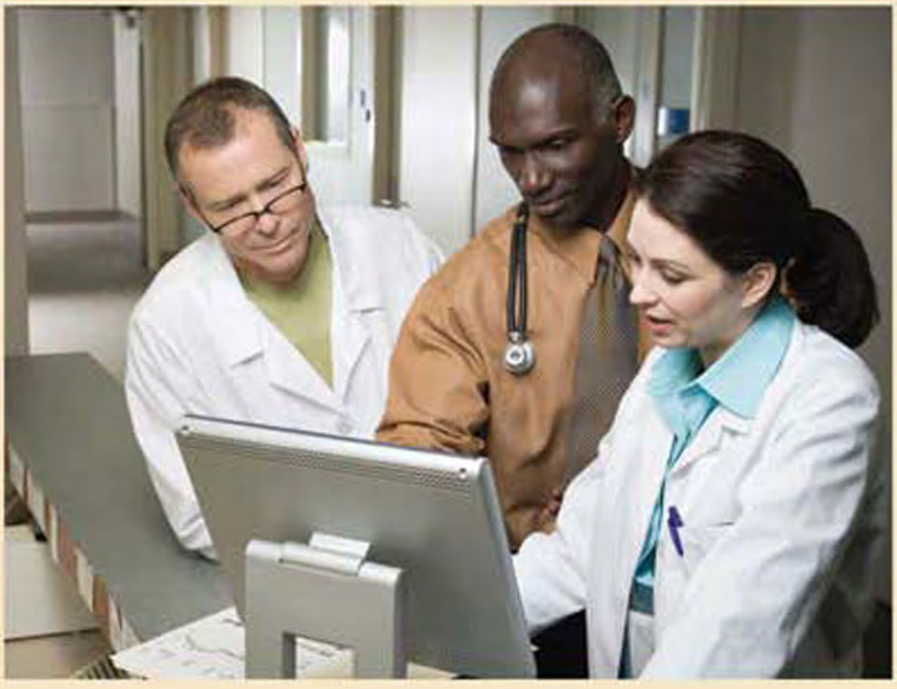
Encyclopedia of

MEDICAL DECISION MAKING



MICHAEL W. KATTAN, EDITOR

Encyclopedia of

MEDICAL DECISION MAKING

Editorial Board

Encyclopedia of MEDICAL DECISION MAKING

MICHAEL W. KATTAN, EDITOR
Cleveland Clinic

MARK E. COWEN, ASSOCIATE EDITOR
St. Joseph Mercy Health System, Michigan

SAGE

Los Angeles | London | New Delhi
Singapore | Washington DC

A SAGE Reference Publication

For information:

SAGE Publications, Inc.
2455 Teller Road
Thousand Oaks, California 91320
E-mail: order@sagepub.com

SAGE Publications Ltd.
1 Oliver's Yard
55 City Road
London, EC1Y 1SP
United Kingdom

SAGE Publications India Pvt. Ltd.
B 1/I 1 Mohan Cooperative Industrial Area
Mathura Road, New Delhi 110 044
India

SAGE Publications Asia-Pacific Pte. Ltd.
33 Pekin Street #02-01
Far East Square
Singapore 048763

Printed in the United States of America.

Library of Congress Cataloging-in-Publication Data

Encyclopedia of medical decision making/editor, Michael W. Kattan ; associate editor, Mark E. Cowen.
p. ; cm.
Includes bibliographical references and index.
ISBN 978-1-4129-5372-6 (cloth : alk. paper)
1. Medicine—Decision making—Encyclopedias. 2. Clinical medicine—Decision making—Encyclopedias. 3. Diagnosis—Decision making—Encyclopedias. I. Kattan, Michael W. II. Cowen, Mark E.
[DNLM: 1. Clinical Medicine—Encyclopedias—English. 2. Decision Making—Encyclopedias—English. 3. Costs and Cost Analysis—Encyclopedias—English. 4. Decision Support Techniques—Encyclopedias—English. 5. Patient Participation—Encyclopedias—English. 6. Quality of Health Care—Encyclopedias—English. WB 13 E56289 2009]

R723.5.E53 2009
610.3—dc22 2009004379

This book is printed on acid-free paper.

09 10 11 12 13 10 9 8 7 6 5 4 3 2 1

Publisher:	Rolf A. Janke
Assistant to the Publisher:	Michele Thompson
Acquisitions Editor:	Jim Brace-Thompson
Developmental Editor:	Carole Maurer
Reference Systems Manager:	Leticia Gutierrez
Reference Systems Coordinator:	Laura Notton
Production Editor:	Tracy Buyan
Copy Editor:	QuADS Prepress (P) Ltd.
Typesetter:	C&M Digitals (P) Ltd.
Proofreaders:	Jenifer Kooiman, Scott Oney
Indexer:	Virgil Diodato
Cover Designer:	Gail Buschman
Marketing Manager:	Amberlyn McKay

Contents

List of Entries

Reader's Guide

The alphabetical organization of an encyclopedia facilitates access to information when the reader can identify the topic of interest. Some readers, on the other hand, may prefer to use the encyclopedia as a source for topical study and to sample key concepts of an academic discipline sequentially. This Reader's Guide is an attempt to catalog essays to mirror the components of the decision-making process. Many essays could be grouped under more than one category, and some titles might have more than one connotation. The following organization is offered as one of many possible ways to guide topical reading:

Basis for Making the Decision. These essays examine criteria by which the optimal choice among available alternatives can be identified. Some methods, particularly those drawn from the field of health economics or used by decision analysts, are quantitative and permit the rank ordering of potential decision strategies. However, there are also other considerations used in making a final decision, sometimes on philosophical or ethical grounds.

Biostatistics and Clinical Epidemiology: The Assessment of the Likelihood of Possible Consequences or Outcomes. These entries present some of the techniques used to determine the probabilities of health outcomes, to determine if the results of a clinical study are due to chance or some alternate explanation, and to assess the accuracy of diagnostic tests and prognostic algorithms. These concepts often underlie the approaches described in essays throughout the Encyclopedia, and provide background for understanding the Methods sections of scientific publications.

Decision Analysis and Related Mathematical Models. These essays present techniques for a rational or prescriptive decision-making process for an individual or population. The choice to be made, the possible consequences, their value or cost, and their likelihood of occurring are combined to identify the optimal decision.

Health Outcomes and Measurement. These essays discuss some of the possible health outcomes that follow a medical or health policy decision and how they can be measured or quantified. Some of these essays provide a foundation for understanding health-related surveys and evaluations of the quality of care provided by health professionals or health systems.

Impact or Weight or Utility of the Possible Outcomes. These essays examine the value or "utility" placed on certain health outcomes to indicate their relative level of desirability or preference in the eyes of patients or the general population. Some essays describe the methods for determining this value, whereas others describe how utilities can be combined and aggregated in decision analyses or economic analyses.

Other Techniques, Theories, and Tools to Understand and to Assist Decision Making. Although the following essays do not fit neatly into the other categories, they represent a valuable array for understanding patients and healthcare delivery systems, and provide potential resources for guiding clinicians and patients in making sound decisions.

Perspective of the Decision Maker. The relevant components of decision making and the considerations

used to determine a course of action may differ according to the point of view or perspective of the person or entity empowered to choose. The following essays include examples of health-related scenarios for decision making by individuals, clinicians, health systems, governments, and other entities.

The Psychology Underlying Decision Making. These essays represent scholarly work to understand how humans appropriate, use, and process information when they make their choices. Influences on, vulnerabilities associated with, and strategies developed to improve decision making are discussed.

Basis for Making the Decision

Acceptability Curves and Confidence Ellipses
Beneficence
Bioethics
Choice Theories
Construction of Values
Cost-Benefit Analysis
Cost-Comparison Analysis
Cost-Consequence Analysis
Cost-Effectiveness Analysis
Cost-Minimization Analysis
Cost-Utility Analysis
Decision Quality
Distributive Justice
Dominance
Equity
Evaluating Consequences
Expected Utility Theory
Expected Value of Perfect Information
Extended Dominance
Health Production Function
League Tables for Incremental Cost-Effectiveness Ratios
Marginal or Incremental Analysis, Cost-Effectiveness Ratio
Monetary Value
Moral Choice and Public Policy
Net Benefit Regression
Net Monetary Benefit
Nonexpected Utility Theories
Pharmacoeconomics
Protected Values
Rank-Dependent Utility Theory
Return on Investment
Risk-Benefit Trade-Off
Subjective Expected Utility Theory
Toss-Ups and Close Calls
Value-Based Insurance Design
Welfare, Welfarism, and Extrawelfarism

Biostatistics and Clinical Epidemiology

Analysis of Covariance (ANCOVA)
Analysis of Variance (ANOVA)
Attributable Risk
Basic Common Statistical Tests: Chi-Square, *t* Test, Nonparametric Test
Bayesian Analysis
Bayesian Evidence Synthesis
Bayesian Networks
Bayes's Theorem
Bias
Bias in Scientific Studies
Brier Scores
Calibration
Case Control
Causal Inference and Diagrams
Causal Inference in Medical Decision Making
Conditional Independence
Conditional Probability
Confidence Intervals
Confounding and Effect Modulation
Cox Proportional Hazards Regression
Decision Rules
Diagnostic Tests
Discrimination
Distributions: Overview
Dynamic Treatment Regimens
Effect Size
Equivalence Testing
Experimental Designs
Factor Analysis and Principal Components Analysis
Fixed Versus Random Effects
Frequentist Approach
Hazard Ratio
Hypothesis Testing
Index Test
Intraclass Correlation Coefficient
Likelihood Ratio
Logic Regression
Logistic Regression
Log-Rank Test

About the Editors

General Editor

Michael W. Kattan is Chairman of the Department of Quantitative Health Sciences at Cleveland Clinic. He is also Professor of Medicine, Epidemiology, and Biostatistics at Cleveland Clinic Lerner College of Medicine of Case Western Reserve University. Prior to joining Cleveland Clinic, he was Associate Attending Outcomes Research Scientist at Memorial Sloan-Kettering Cancer Center and Associate Professor of Biostatistics in Urology at Cornell University in New York City. He began his academic career as an assistant professor of urology and medical informatics at Baylor College of Medicine in Houston, Texas, where he also obtained his postdoctorate in medical informatics.

His primary research interest lies in prescriptive medical decision making—how physicians and patients should make decisions. Specifically, he is most interested in medical prediction: how, why, and when. He has received multiple patents for his work in this area and coauthored about 300 articles in peer-reviewed journals. In 2008, he received the Eugene Saenger Distinguished Service Award from the Society for Medical Decision Making. He serves or has served on the editorial boards for several journals, including *Cancer Investigation, Nature Clinical Practice Urology, Medical Decision Making, Clinical Genitourinary Cancer, Urologic Oncology, Journal of Urology,* and *Urologic Oncology: Seminars and Original Investigation*. His PhD is in management information systems, with a minor in statistics, from the University of Houston. He also has a master's of business administration, with concentration in computer information systems and quantitative analysis, from the University of Arkansas. His undergraduate degree is in food science, also from the University of Arkansas.

Associate Editor

Mark E. Cowen, MD, is Chief of Clinical Decision Services at the St. Joseph Mercy Health System, Ann Arbor, Michigan. He was founder and president of a private group practice in internal medicine and served as a clinical instructor for the University of Michigan Medical School for a number of years. After receiving a master of science degree in epidemiology from the Harvard School of Public Health, he was Vice President, Performance Improvement, of Allegiance LLC, a physician-hospital organization for managed care. He is currently on the editorial board of the *American Journal of Managed Care*. His research has been largely driven by questions arising from daily responsibilities with managed care or hospitalized patient populations. A concurrent interest in prostate cancer screening and treatment decisions led to the development of a Markov model and collaborations with Dr. Kattan to study various components of decision making regarding this disease. He has been a member of the prostate cancer outcomes task force of the American Urological Association. He received his undergraduate and medical degrees from the University of Michigan.

Contributors

A E Ades
University of Bristol

Arpita Aggarwal
Virginia Commonwealth University

Laith Alattar
University of Michigan

Daniel Almirall
Duke University School of Medicine

Allen Andrew Alvarez
University of the Philippines

Andrea Angott
University of Michigan

Noriaki Aoki
University of Texas–Houston

Arlene S. Ash
Boston University School of Medicine

Koula Asimakopoulou
King's College London

Carla Bann
RTI International

Paul G. Barnett
U.S. Department of Veterans Affairs

Barbara A. Bartman
Agency for Healthcare Research and Quality

Daniel Beavers
Baylor University

J. Robert Beck
Fox Chase Cancer Center

James F. Bena
Cleveland Clinic

George Bergus
University of Iowa

Whitney Berta
University of Toronto

Harald Binder
University Medical Center

Jakob B. Bjorner
QualityMetric Inc.

Eugene H. Blackstone
Cleveland Clinic

Gerhard Blasche
Medical University of Vienna

Han Bleichrodt
Erasmus University

Donald Bordley
University of Rochester Medical Center

Emanuele Borgonovo
Bocconi University

Brian H. Bornstein
University of Nebraska–Lincoln

Mari Botti
Deakin University

Dave Bouckenooghe
Vlerick Leuven Gent Management School

Aziz A. Boxwala
Harvard Medical School

Eduard Brandstätter
Johannes Kepler University of Linz

John E. Brazier
University of Sheffield

Karen E. Bremner
University Health Network

Frank Brennan
Calvary Hospital

Andrew H. Briggs
University of Glasgow

Arndt Bröder
University of Bonn

Werner Brouwer
Erasmus MC Rotterdam

Tracey Bucknall
Deakin University

Marc Buelens
Vlerick Leuven Gent Management School

Robert S. Butler
Cleveland Clinic

Scott B. Cantor
University of Texas M. D. Anderson Cancer Center

Linda Jean Carroll
University of Alberta

Lydia L. Chen
University of Michigan

Ying Qing Chen
Fred Hutchinson Cancer Research Center

Karis K. F. Cheng
Chinese University of Hong Kong

V. K. Chetty
Boston University

Ling-Hsiang Chuang
Centre for Health Economics

Felix K.-H. Chun
University of Hamburg

Leslie Citrome
New York University Medical Center

Nancy S. Clark
University of Rochester Medical Center

Karl Claxton
University of York

Phaedra Corso
University of Georgia

Michael Cousins
University of Sydney Pain Management Research Institute

William Dale
University of Chicago

Jarrod E. Dalton
Cleveland Clinic

Laura J. Damschroder
Ann Arbor VA HSR&D Center of Excellence

Raisa Deber
University of Toronto

Richard A. Demme
University of Rochester Medical Center

Francisco J. Díez
UNED (Spain)

Peter H. Ditto
University of California, Irvine

Robert S. Dittus
Vanderbilt University

Jason N. Doctor
University of Southern California

Ray Dolan
University College London

Michael R. Dougherty
University of Maryland

Stephan Dreiseitl
Upper Austria University of Applied Sciences

Marek J. Druzdzel
University of Pittsburgh

Michael Dunn
University of Oxford

Mette Ebbesen
University of Aarhus

Mark H. Eckman
University of Cincinnati

Heather Edelblute
University of North Carolina at Chapel Hill

Eric L. Eisenstein
Duke University Medical Center

A. Christine Emler
Veterans Administration

David Epstein
University of York

Ronald Epstein
University of Rochester Medical Center

Steven Estrada
Cornell University

Margot M. Eves
Cleveland Clinic

Zhaozhi Fan
Memorial University of Newfoundland

Deb Feldman-Stewart
Queen's University

Elisabeth Fenwick
University of Glasgow

Paul J. Ford
Cleveland Clinic

Liana Fraenkel
Yale University

Daniel J. France
Vanderbilt University Medical Center

Jenny V. Freeman
University of Sheffield

Alex Z. Fu
Cleveland Clinic

Amiram Gafni
McMaster University

Wolfgang Gaissmaier
Max Planck Institute for Human Development

Mirta Galesic
Max Planck Institute for Human Development

Rocio Garcia-Retamero
University of Granada

Jason Gatliff
Cleveland Clinic

Constantine Gatsonis
Brown University

R. Brian Giesler
Butler University

Gerd Gigerenzer
Max Planck Institute for Human Development

John Gilmour
Gilmour and Associates Physiotherapy

Alan Girling
University of Birmingham

Julie Goldberg
University of Illinois at Chicago

Morton P. Goldman
Cleveland Clinic

Mithat Gönen
Memorial Sloan-Kettering Cancer Center

Banu Gopalan
Cleveland Clinic

Carolyn C. Gotay
University of British Columbia

Markus Graefen
University of Hamburg

Erika Graf
University Medical Center

Dan Greenberg
Ben Gurion University of the Negev

Kalle Grill
Royal Institute of Technology

Erik J. Groessl
VA San Diego/University of California, San Diego

Scott D. Grosse
Centers for Disease Control

Enzo Grossi
Bracco

Frank M. Guess
University of Tennessee

Alexander Haese
University of Hamburg

C. Gregory Hagerty
Robert Wood Johnson Medical School

Susan Halabi
Duke University

Bruce P. Hallbert
Idaho National Laboratory

Robert M. Hamm
University of Oklahoma Health Sciences Center

Seunghee Han
Carnegie Mellon

Ronald B. Harrist
University of Texas, Austin Regional Campus

Kate Haswell
Auckland University of Technology

Katherine Hauser
Cleveland Clinic

Daniel Hausmann
University of Zurich

Ron D. Hays
University of California, Los Angeles

Christopher Hebert
Cleveland Clinic

Glenn Heller
Memorial Sloan-Kettering Cancer Center

Joshua Hemmerich
University of Chicago

Lidewij Henneman
VU University (Amsterdam) Medical Center

Adrian V. Hernandez
Cleveland Clinic

Jørgen Hilden
University of Copenhagen

Richard A. Hirth
University of Michigan School of Public Health

Jeffrey S. Hoch
St. Michael's Hospital

Eduard Hofer
Retired Mathematician

Søren Holm
Cardiff University

Kirsten Howard
University of Sydney

Xuelin Huang
University of Texas M. D. Anderson Cancer Center

Yunchen Huang
Mississippi State University

M. G. Myriam Hunink
Erasmus University Medical Center

Jordan Hupert
University of Illinois College of Medicine

Don Husereau
Health Technology Assessment Council

Lisa I. Iezzoni
Harvard Medical School/ Institute for Health Policy, Massachusetts General Hospital

Lee H. Igel
New York University

Peter B. Imrey
Cleveland Clinic

Hemant Ishwaran
Cleveland Clinic

John L. Jackson Jr.
University of Pennsylvania

Philip Jacobs
University of Alberta

Eric W. Jamoom
University of Florida

Stephen Jan
George Institute for International Health

Naveed Zafar Janjua
Aga Khan University, Karachi, Pakistan

Ruth Jepson
University of Stirling

Ava John-Baptiste
University of Toronto

Michael L. Johnson
University of Houston

Robert M. Kaplan
University of California, Los Angeles

Matthew Karafa
Cleveland Clinic

Pierre I. Karakiewicz
University of Montreal

Jonathan Karnon
University of Adelaide

Catherine Kastanioti
Technological Educational Institution of Kalamata

David A. Katz
University of Iowa Carver College of Medicine

H. J. Keselman
University of Manitoba

J. Kievit
Leiden University Medical Center

Sunghan Kim
University of Toronto

Sara J. Knight
University of California, San Francisco

Spassena Koleva
University of California, Irvine

Charles Kooperberg
Fred Hutchinson Cancer Research Center

Wendy Kornbluth
Cleveland Clinic

Olga Kostopoulou
University of Birmingham

Murray Krahn
THETA Collaborative

Ronilda Lacson
Harvard Medical School

Elizabeth B. Lamont
Harvard Medical School

Audrey Laporte
University of Toronto

Franklin N. Laufer
New York State Department of Health

France Légaré
Université Laval

Allen J. Lehman
Arthritis Research Centre of Canada

Harold Lehmann
Johns Hopkins Medical Institutions

Jennifer S. Lerner
Harvard University

Scott R. Levin
Johns Hopkins University

Liang Li
Cleveland Clinic

Matthew H. Liang
Harvard University

Richard Lilford
University of Birmingham, UK

Carol L. Link
New England Research Institutes

Joseph Lipscomb
Rollins School of Public Health, Emory University

Benjamin Littenberg
University of Vermont

Lisa M. Lix
University of Saskatchewan

Hilary A. Llewellyn-Thomas
Dartmouth Medical School

Karen E. Lutfey
New England Research Institutes

Sílvia Mamede
Erasmus University Rotterdam

Edward C. Mansley
Merck & Co., Inc.

P. J. Marang-van de Mheen
Leiden University Medical Centre

Lisa D. Marceau
New England Research Institutes

Kathryn Markakis
University of Rochester Medical Center

Ed Mascha
Cleveland Clinic

Josephine Mauskopf
RTI Health Solutions

Madhu Mazumdar
Weill Cornell Medical College

Dennis J. Mazur
Department of Veterans Affairs Medical Center

Christine M. McDonough
Dartmouth Institute for Health

Craig R. M. McKenzie
Rady School of Management and Psychology Department

John B. McKinlay
New England Research Institutes

Michael McMillan
Cleveland Clinic

Katherine Mead
George Washington University

Alan Meisel
University of Pittsburgh

J. Michael Menke
University of Arizona

Lesley-Ann N. Miller
University of Texas M. D. Anderson Cancer Center

Wilhelmine Miller
GWU School of Public Health and Health Services

Britain Mills
Cornell University

Alex J Mitchell
Consultant and Honorary Senior Lecturer

Nandita Mitra
University of Pennsylvania

Liz Moliski
University of Chicago

Barbara Moore
Gilmour and Associates Physiotherapy

Chaya S. Moskowitz
Memorial Sloan-Kettering Cancer Center

Stephanie Müller
University of Granada

Susan A. Murphy
University of Michigan

Jonathan D. Nelson
University of California, San Diego

Peter J. Neumann
Tufts–New England Medical Center

Angela Neumeyer-Gromen
Max Planck Institute for Human Development

J. Tim Newton
King's College London

Jerry Niederman
Rush University College of Medicine

Annette O'Connor
Ottawa Health Research Institute

Lucila Ohno-Machado
Brigham and Women's Hospital, Harvard Medical School

Sachiko Ohta
Center for Health Service, Outcomes Research and Development–Japan (CHORD-J)

Stephen Olejnik
University of Georgia

Christopher Y. Olivola
Princeton University

Obinna Onwujekwe
College of Medicine, University of Nigeria Enugu

Daniel M. Oppenheimer
Princeton University

Monica Ortendahl
Royal Institute of Technology

Katherine S. Panageas
Memorial Sloan-Kettering Cancer Center

Robert Panzer
University of Rochester Medical Center

Robert Patrick
Cleveland Clinic

Katherine Payne
University of Manchester

Niels Peek
Academic Medical Center (Amsterdam)

Alleene M. Ferguson Pingenot
California State University, Stanislaus

Petra Platzer
Cleveland Clinic

Harold A. Pollack
University of Chicago

Maarten J. Postma
University of Groningen

Georges Potworowski
University of Michigan

Robert K. Pretzlaff
University of California, Davis Medical Center

Lisa Prosser
University of Michigan Health System

Timothy E. Quill
University of Rochester Medical Center

Rob Ranyard
University of Bolton

J. Sunil Rao
Case Western Reserve University

Thomas C. Redman
Navesink Consulting Group

Valerie F. Reyna
Cornell University

Remy Rikers
Erasmus University Rotterdam

Stephen D. Roberts
North Carolina State University

Virginie Rondeau
Bordeaux School of Public Health

Aubri S. Rose
Dartmouth College

Geoffrey L. Rosenthal
Childrens Hospital, Pediatric Institute, Cleveland Clinic

Ingo Ruczinski
Johns Hopkins University

Tracey H. Sach
University of East Anglia

Michi Sakai
Center for Health Service, Outcomes Research and Development–Japan (CHORD-J)

Arash Salehi
Mississippi State University

Lacey Schaefer
Mississippi State University

Marilyn M. Schapira
Medical College of Wisconsin

Michael Schlander
University of Heidelberg, Mannheim Medical Faculty

Henk G. Schmidt
Rotterdam

Jesse D. Schold
University of Florida

Alan Schwartz
University of Illinois at Chicago

Mark Sculpher
Centre for Health Economics

John W. Seaman Jr.
Baylor University

Karen R. Sepucha
Massachusetts General Hospital

Anuj K. Shah
Princeton University

James Shanteau
Kansas State University

Ya-Chen Tina Shih
University of Texas M. D. Anderson Cancer Center

Michael Shwartz
Boston University

Uwe Siebert
UMIT–University for Health Sciences (Austria)

Bruce Siegel
George Washington University

Mahender P. Singh
Massachusetts Institute of Technology

Grant H. Skrepnek
University of Arizona

Dean G. Smith
University of Michigan

Kenneth J. Smith
University of Pittsburgh

Martin L. Smith
Cleveland Clinic

Richard D. Smith
London School of Hygiene and Tropical Medicine

Claire F. Snyder
Johns Hopkins School of Medicine

Frank A. Sonnenberg
UMDNJ–Robert Wood Johnson Medical School

Chenni Sriram
Cleveland Clinic

James Stahl
Massachusetts General Hospital

James D. Stamey
Baylor University

Thomas R. Stewart
University at Albany, State University of New York

Ewout W. Steyerberg
Erasmus Medical Center

Anne M. Stiggelbout
Leiden University Medical Center

Theo Stijnen
Leiden University Medical Center

Lesley Strawderman
Mississippi State University

J. Shannon Swan
MGH Institute for Technology Assessment

Carmen Tanner
University of Zurich

Curtis Tatsuoka
Cleveland Clinic

Rick P. Thomas
Oklahoma University

Danielle R. M. Timmermans
EMGO Institute, VU University (Amsterdam) Medical Center

Richard J. Tunney
University of Nottingham

Diane M. Turner-Bowker
QualityMetric Inc.

M. Dolores Ugarte
Universidad Publica de Navarra

Erin Winters Ulloa
VA Boston Healthcare System

Wilbert van den Hout
Leiden University Medical Center

Ben Vandermeer
University of Alberta

Tyler J. VanderWeele
University of Chicago

René (M) van Hulst
University of Groningen

Elisabeth van Rijen
Erasmus University Rotterdam

Marion Verduijn
Leiden University Medical Center

Andrew J. Vickers
Memorial Sloan-Kettering Cancer Center

Ivo Vlaev
University College London

Esteban Walker
Cleveland Clinic

David A. Walsh
University of Southern California

Declan Walsh
Cleveland Clinic

Jochen Walz
Institut Paoli-Calmettes

Bin Wang
University of South Alabama

Xiao-Feng Wang
Cleveland Clinic

Elke U. Weber
Columbia University

Noah J. Webster
Case Western Reserve University

Douglas H. Wedell
University of South Carolina

Saul J. Weiner
VA Center for the Management of Complex Chronic Care/ University of Illinois at Chicago

Kevin Weinfurt
Duke Clinical Research Institute

Brian J. Wells
Cleveland Clinic

Robert L. Winkler
Duke University

Eve Wittenberg
Brandeis University

Sarah E. Worley
Cleveland Clinic

J. Frank Yates
University of Michigan

Jun-Yen Yeh
Cleveland Clinic

Andrew Peng Yu
Analysis Group, Inc.

Changhong Yu
Cleveland Clinic

Marcel Zeelenberg
Tilburg University

Han Zhang
Mississippi State University

Li Zhang
Cleveland Clinic

Sue Ziebland
University of Oxford

Armineh Zohrabian
Centers for Disease Control

Foreword

As the chief academic officer in a cancer research institution, Fox Chase Cancer Center, I meet with each of the new faculty members shortly after their arrival. Recently, a young radiation oncologist came to my office, where we discussed the usual junior faculty issues: adjustment to the center, mentoring, the promotion and tenure process. When I asked him about his research interests, he startled me by expressing a desire to conduct "willingness to pay" studies of new modalities in the radiation therapy of prostate cancer. Then he asked me if I knew anything about this type of research. I admitted I did know a little about it, and offered to refer him to various texts on cost-effectiveness and cost-utility analysis. I could also have sent him to Becker, DeGroot, and Marschak's 1964 paper in *Behavioral Science*, "Measuring Utility by a Single-Response Sequential Method," a foundational article in willingness-to-pay studies, and encouraged him to perform a forward citation search using a tool such as the Web of Science.

Becker et al. (1964) have been cited 255 times since its publication, and the citing articles cover a broad range from econometrics to neuroscience. This might not be the easiest way to learn about a technical topic in valuing health outcomes. Another approach might be a keyword search, focusing on the biomedical literature. PubMed, the Web-based search engine to the comprehensive holdings in the U.S. National Library of Medicine, matches more than 1,100 articles to the text phrase "willingness to pay." Browsing the most recent 50 or so citations turns up a familiar name, Joel Tsevat, a friend who trained with my mentor, Steve Pauker. Downloading a recent paper of Joel's from *Medical Decision Making*, I find in the reference list a few contemporary methods papers on willingness to pay. If these papers suffice for Joel and his team, they are likely good enough for my young colleague.

Such is a typical approach to exploring a specific research topic in biomedical research. I confess to adding Wikipedia to my routine search strategy, as well as Google Scholar. Admittedly, the thrill of the hunt motivates some of my exploration, but the field of medical decision making could use a comprehensive reference. A field that draws from economics, mathematics, medicine, philosophy, psychology, and sociology (and occasionally from many others) is particularly in need of a compendium of ideas and techniques.

This encyclopedia aims to address this need. Didactic articles on more than 300 headwords have been prepared by well over 200 contributors from around the world. Joel Tsevat is on the advisory board for the encyclopedia, along with a number of other leaders who span the disciplines within the field of medical decision making. The article on willingness to pay is written by Obinna Onwujekwe, a health economist and clinician from the London School of Tropical Medicine and Hygiene, based at the University of Nigeria in Enugu, and funded through the Gates Malaria Partnership. I don't know Dr. Onwujekwe, but through this contact, I have discovered more resources on the Internet that can support my research and build my professional network. A well-researched encyclopedia can contribute much to the furthering of knowledge and the application of appropriate techniques to current problems. I look forward to having this reference for our current and future trainees.

At the publication of this encyclopedia, the field is 50 years old if one dates from the publication of Ledley and Lusted's seminal "Reasoning Foundations of Medical Diagnosis" (*Science*, 1959). Table 1 lists frequently cited articles in the field from that point forward, using title words and search terms from MEDLINE and Web of Science. A number of important technical manuscripts are included in this list, as well as "first papers" in several disciplines. Of course, this list is subject to the vagaries of article indexing; papers that focus on "risk" are relatively underrepresented in this set.

Table 1 Highly cited articles in the field of medical decision making, beginning with Ledley and Lusted's 1959 *Science* paper

First Author	*Short Title*	*Journal*	*Year*	*Cited*
Ledley, R. S.	Reasoning foundations of medical diagnosis	*Science*	1959	312
Wennberg, J. E.	Small area variations in health care delivery	*Science*	1973	751
Schwartz, W. B.	Decision analysis and clinical judgment	*American Journal of Medicine*	1973	193
McNeil, B. J.	Primer on certain elements of medical decision making	*New England Journal of Medicine*	1975	974
Pauker, S. G.	Therapeutic decision-making—cost-benefit analysis	*New England Journal of Medicine*	1975	223
Pauker, S. G.	Coronary artery surgery—use of decision analysis	*Annals of Internal Medicine*	1976	107
Kassirer, J. P.	Principles of clinical decision-making—introduction to decision analysis	*Yale Journal of Biology and Medicine*	1976	102
Weinstein, M. C.	Foundations of cost-effectiveness analysis . . .	*New England Journal of Medicine*	1977	1,043
Eisenberg, J. M.	Sociologic influences on decision-making by clinicians	*Annals of Internal Medicine*	1979	225
Shortliffe, E. H.	Knowledge engineering for medical decision making	*Proceedings of the IEEE*	1979	145
Pauker, S. G.	The threshold approach to clinical decision making	*New England Journal of Medicine*	1980	480
Griner, P. F.	Selection and interpretation of diagnostic tests . . .	*Annals of Internal Medicine*	1981	600
McNeil, B. J.	On the elicitation of preferences for alternative therapies	*New England Journal of Medicine*	1982	604
Beck, J. R.	A convenient approximation of life expectancy (the DEALE) 2 . . .	*American Journal of Medicine*	1982	256
Beck, J. R.	A convenient approximation of life expectancy (the DEALE) 1 . . .	*American Journal of Medicine*	1982	253
Beck, J. R.	Markov process in medical prognosis	*Medical Decision Making*	1983	473
Spiegelhalter, D. J.	Statistical and knowledge-based approaches to CDSS	*Journal of the Royal Statistical Society: Series A (General)*	1984	156
Greenfield, S.	Expanding patient involvement in care: Effect on health outcomes	*Annals of Internal Medicine*	1985	635

First Author	*Short Title*	*Journal*	*Year*	*Cited*
Sox, H. C.	Probability-theory in the use of diagnostic tests	*Annals of Internal Medicine*	1986	274
Pauker, S. G.	Decision-analysis	*New England Journal of Medicine*	1987	426
Kassirer, J. P.	Decision-analysis—a progress report	*Annals of Internal Medicine*	1987	190
Shortliffe, E. H.	Computer programs to support clinical decision making	*Journal of the American Medical Association*	1987	125
Swets, J. A.	Measuring the accuracy of diagnostic systems	*Science*	1988	1,339
Detsky, A. S.	Clinician guide to cost-effectiveness analysis	*Annals of Internal Medicine*	1990	435
Boyd, N. F.	Whose utilities for decision analysis	*Medical Decision Making*	1990	191
Fryback, D. G.	The efficacy of diagnostic-imaging	*Medical Decision Making*	1991	303
Hillner, B. E.	Efficacy and cost-effectiveness of adjuvant chemotherapy . . .	*New England Journal of Medicine*	1991	152
Sonnenberg, F. A.	Markov models in medical decision making . . .	*Medical Decision Making*	1993	773
Fleming, C.	A decision analysis . . . clinically localized prostate cancer	*Journal of the American Medical Association*	1993	448
Wu, Y. Z.	Artificial neural networks in mammography	*Radiology*	1993	250
Smith, T. J.	Efficacy and cost-effectiveness of cancer treatment	*Journal of the National Cancer Institute*	1993	157
Jaeschke, R.	Users' guides to the med. lit. 3. How to use an article about a diagnostic test B. What are the results?	*Journal of the American Medical Association*	1994	936
Krahn, M. D.	Screening for prostate cancer—a decision analytic view	*Journal of the American Medical Association*	1994	256
Omeara, J. J.	A decision analysis . . . for deep vein thrombosis	*New England Journal of Medicine*	1994	105
Davis, D. A.	Changing physician performance—A review of CME strategies	*Journal of the American Medical Association*	1995	1,298

(Continued)

Table 1 Continued

First Author	*Short Title*	*Journal*	*Year*	*Cited*
Wilson, I. B.	Linking clinical variables with HRQOL	*Journal of the American Medical Association*	1995	763
Weinstein, M. C.	Recommendations of the panel on CE in health and medicine	*Journal of the American Medical Association*	1996	829
Russell, L. B.	The role of cost-effectiveness analysis in health and medicine	*Journal of the American Medical Association*	1996	502
Pestotnik, S. L.	Implementing antibiotic practice guidelines through CDSS	*Annals of Internal Medicine*	1996	325
Gambhir, S. S.	Decision tree sensitivity analysis for cost-effectiveness of FDG-PET	*Journal of Nuclear Medicine*	1996	170
Partin, A. W.	Combination of PSA, clinical stage, and Gleason score . . .	*Journal of the American Medical Association*	1997	895
Schrag, D.	Decision analysis—effects of prophylactic mastectomy and oophorectomy	*New England Journal of Medicine*	1997	239
Fine, M. J.	A prediction rule . . . low-risk patients with community-acquired pneumonia	*New England Journal of Medicine*	1997	1,141
Bates, D. W.	Effect of CPOE . . . on prevention of serious medication errors	*Journal of the American Medical Association*	1998	631
Hunt, D. L.	Effects of CDSS on physician performance . . .	*Journal of the American Medical Association*	1998	506
Briggs, A.	An introduction to Markov modelling for economic evaluation	*PharmacoEconomics*	1998	143
Gambhir, S. S.	Analytical decision model for . . . solitary pulmonary nodules	*Journal of Clinical Oncology*	1998	116
Grann, V. R.	DA of prophylactic mastectomy/oophorectomy in BRCA-1 positive . . .	*Journal of Clinical Oncology*	1998	115
Schulman, K. A.	The effect of race and sex on physicians' recommendations for cardiac catheterization	*New England Journal of Medicine*	1999	665
Braddock, C. H.	Informed decision making in outpatient practice	*Journal of the American Medical Association*	1999	285
Claxton, K.	A rational framework for decision making by the NICE	*Lancet*	2002	116
Weinstein, M. C.	Principles of good practice for DA modeling in health care evaluation	*Value Health Care*	2003	170

A recently proposed measure of scientific prominence, the h-index, ranks articles in order of times cited since publication. A scholar's or institution's h-index is represented by that article whose rank in terms of times cited is nearest to the actual number of citations. For example, the editor of this encyclopedia, Michael Kattan, has an h-index of 56: His 56th most cited paper has been cited 56 times through this writing. Taking this idea to a search, the topic and title phrase "medical decision making" in the Web of Science (which, unsurprisingly, incorporates more types of articles than covered in this encyclopedia) has an h-index of 131. By comparison, "bioinformatics" has an h-index of 122 and "medical informatics," 40, whereas the comprehensive biomedical topic, "chemotherapy," has an h-index of 239. This suggests to me that the field has attained a level of maturity where comprehensive reference works such as this encyclopedia will add value to teachers and learners. I look forward to browsing this reference and to following the scholarly output of its distinguished board of editors and contributors.

J. Robert Beck, MD
Fox Chase Cancer Center

Introduction

Healthcare decisions affect all of us, whether on a personal, professional, or societal level. We are human, as are all decision makers, and so are blessed and bound by the resources and limitations of the human mind. We cannot predict the future perfectly; we cannot arrange all positive and negative events to our liking; and we may not always understand the available choices. Moreover, even if effective treatments are recognized, financial constraints may force the selection of one option to the exclusion of others. This encyclopedia provides an introduction to some of the pitfalls and potential solutions our species has developed in the quest for achieving better decisions with less regret.

The audience for this encyclopedia is broad, and the need for a compilation of short essays over an equally expansive range of topics is great. There are a number of examples. Patients may wish to understand their vulnerability in interpreting the level of risks and benefits of treatment options, how their decisions are shaped by culture and emotions, or how physicians assess evidence and make diagnoses. Policy makers may seek firsthand knowledge on the basics of economic analyses, health measurement, and bioethics. Clinicians may desire deeper insights into the influences on their own processes of making diagnoses or choosing treatments or understanding the steps by which decision algorithms found in the literature are constructed and evaluated. Ironically, many participating in medical decision making have not accessed the impressive and exciting body of study and scholarship available. The anticipated time commitment and availability of formal courses may have prevented some from exploring the contributions of cognitive psychology, decision analysis, ethics, health economics, health outcomes, biostatistics, and clinical epidemiology. Others may have tried some self-study but become frustrated when confronted with new vocabulary or advanced mathematics. Some potential readers may have had lectures or training in the past but struggle now to retrieve particular points quickly and apply them to their current practices. And many in the target audience may be experts in one aspect of medical decision making but wish to enhance or energize their work by understanding a different perspective.

Satisfying the interests, needs, and time constraints of a diverse audience is challenging. We have attempted to address these, first, by making the encyclopedia instructional, not simply informational. The authors have written clearly, explained carefully the general sense of the mathematical formulae when presented, and provided generously the many and varied examples so that a wide range of readers can understand and appreciate the material. Certainly no encyclopedia can substitute for a textbook or formal course on a particular topic; this encyclopedia provides a quick and comprehensible introduction. Next, we wanted each essay to be understandable on its own account, not critically dependent on previous readings or coursework. Nevertheless, the authors have also suggested related topics and further readings at the ends of the articles. A third consideration guiding the development of this work was that it should reflect international scholarship. The authors represent nearly every continent. Many have contributed to the primary foundations of medical decision making; to have their work represented here together may eventually be viewed as historical.

Given the universality of medical decision making, the list of potential topics to include can quickly grow to an unmanageable length. A conceptual framework is needed to identify the key

ideas and organize their elaboration. One approach is to classify studies of medical decision making as either prescriptive (also called normative) or descriptive. Work in the prescriptive area investigates the processes and technology by which optimal medical decisions should be determined. In contrast, descriptive studies examine how decisions actually are made. Perhaps not surprisingly, these two strategies are often in disagreement. Prescriptive decision making often employs formal analyses and algorithms, often via a computer, to calculate the best choice under the circumstances, for instance, maximizing benefit relative to the costs. These algorithms may be complicated, inaccessible, not trusted, and thus not used. In their absence, patients, physicians, or policy makers might use less formal methods to make decisions yet may do as well or better than the more sophisticated approach.

Our encyclopedia addresses both categories—prescriptive and descriptive—through a conceptual structure consisting of six components of classical decision analysis. The first component concerns identification of the decision maker—in other words, who must choose. In general, there are three levels of decision makers, each with a particular perspective: the individual patient or surrogate, the clinician, and society. The second component is the identification of the decision to be made, for instance, the selection of the most likely diagnosis or the therapy with the best chance of cure. The essays pertaining to these first two components generally fall into the descriptive category: how decisions are influenced, finalized, and reviewed afterwards. Generally speaking, these draw heavily from the field of cognitive psychology. The third component concerns the consequences or outcomes of decisions and how these are defined and measured. The corresponding entries generally concern health econometrics and health-related quality-of-life measurement. The fourth category is related, and examines the value of the potential outcomes, often expressed as a monetary sum or a level of desirability termed *utility*. The fifth component in the conceptual framework involves the likelihood or probability of the possible consequences through essays on statistical concepts and clinical epidemiology. The sixth category concerns the mechanism by which individuals, clinicians, and society determine the best decision. This involves ethics, cultural considerations informed by sociology and anthropology, and prescriptive approaches such as utility maximization as well as descriptive approaches related to cognitive psychology. We have also included a seventh category for the encyclopedia, broadly characterized as pertaining to methods and techniques used to predict outcomes and analyze decisions, whether at the individual patient, cohort, or societal level. The pertinent essays cover mathematical models of disease progression, diagnosis, and prognosis as well as economic evaluations.

The encyclopedia was developed in five basic steps.

Step 1: Leading medical decision-making experts around the world were invited to serve on the editorial board.

Step 2: The senior editorial board editor and the associate editor created a master list of topics corresponding to the conceptual framework presented.

Step 3: The editorial board was asked to nominate individuals to author the list of entries. We also searched PubMed and the Web sites of universities to find people publishing on certain topics, and we consulted with our colleagues for additional suggestions.

Step 4: Contributors were given basic guidelines and instructions regarding the writing of their entries. As previously mentioned, we encouraged them to be thorough in describing the entire topic area and to write in nontechnical, accessible language.

Step 5: The editor and associate editor then reviewed all the entries and asked authors for revisions as necessary.

As with the subject matter, this encyclopedia has its own limitations and imperfections. The first concerns the selection of topics. We anticipate surprise with the selection of some included and chagrin with those not found. Our generic response to those questioning an inclusion is that many of the techniques and tools used in medical decision making are based on methodology developed in a related discipline such as statistics or psychometrics. We wanted to provide interested readers with the opportunity to obtain background in these supporting topics to enhance their

enjoyment of the other essays. For those disappointed in the lack of a particular topic, we beg your understanding and trust that your curiosity will lead you to the appropriate reference. The second limitation concerns our attempt to make each essay understandable as a single entity. An inevitable consequence of this editorial approach is some redundancy of content among related essays. We do not foresee this being too problematic with the encyclopedia format. The third limitation concerns the difference between empirical evidence and hypothetical examples for teaching purposes. The field of medical decision making continually develops and includes concepts supported with various levels of evidence. Many of the examples within the essays summarize formal studies referenced at the end of the article. However, other examples are provided as relevant illustrations of the underlying concepts. These should not be taken as firm evidence of the decision practices of a particular culture, profession, or specialty, much less a judgment on the decisions or actions of a given individual.

We conclude with some practical advice for those still reading this Introduction. Start wherever your curiosity is most urgent. If your reading halts due to unfamiliar mathematical notation, consult the essay on "Statistical Notation." If you finish an essay without understanding its major points, read a few entries on the related topics, then return to the original essay. This encyclopedia is a treasure. In the course of its compilation, we have reexperienced the initial joy of discovering concepts and techniques that ultimately changed the directions of our careers.

Michael W. Kattan and Mark E. Cowen

Acknowledgments

I am now severely further indebted to my great and longtime friend, Associate Editor Mark Cowen. He has put an enormous amount of effort into this encyclopedia. Its positive attributes largely belong to him, while the shortcomings rest with me. I picked Mark immediately in the development process because I knew how thoughtful he was. It clearly shows in this encyclopedia.

The SAGE team has been great. Carole Maurer and Laura Notton, most prominently, have really held my hand during this entire project. The SAGE group has put a lot of thought into the encyclopedia process, and it is very good. I thank Neil Salkind at Studio B for identifying me to spearhead this project.

I also thank my star-studded advisory board. I was very pleasantly surprised with how willing these very accomplished folks were to volunteer for this project, given their many time commitments on their time. Their diversity was of tremendous value in identifying the broad listing of topics necessary for a comprehensive text like this. Furthermore, they helped me identify excellent authors for many of the entries.

On a personal level, many people have taken it on themselves to position me to be able to edit this encyclopedia. The individual who has done the most is, without question, Peter T. Scardino, Chief of Surgery at Memorial Sloan-Kettering Cancer Center. For over a decade, Peter single-handedly sponsored and promoted my career. Whatever success I have had can easily be traced to him, an unmatched combination of brilliance, kindness, and humility. More thanks go to Bob Beck and Scott Cantor, who initiated and propelled, respectively, my medical-decision-making exposure. Along those lines, I thank my many colleagues in the Society for Medical Decision Making for their relentless thirst for new knowledge. I must also thank the leadership at Cleveland Clinic for protecting my time so that I might devote it to this project. And finally, I thank my family (Grace, Madeleine, and Lily), who put up with a lot of unattractive behavior on my part.

Michael W. Kattan

A

Acceptability Curves and Confidence Ellipses

Acceptability curves and confidence ellipses are both methods for graphically presenting the uncertainty surrounding the estimate of cost-effectiveness. A confidence ellipse provides a visual representation of the region containing x% (where x is usually 95) of the uncertainty. An acceptability curve provides a graphical representation of the probability that an intervention is cost-effective compared with the alternative(s), given the data. Confidence ellipses can only be used for comparisons between two interventions, whereas acceptability curves can be produced for decisions involving multiple interventions. Confidence ellipses are determined parametrically from information about the distribution of costs and effects (mean, variance, and covariance). The acceptability curve can be determined from the confidence ellipse or direct from the data following an assessment of uncertainty through bootstrapping (for trial data) or probabilistic sensitivity analysis (of modeling analyses). Both are specified as appropriate methods for presenting uncertainty in cost-effectiveness in the *Guide to the Methods of Technology Appraisal* produced by the National Institute for Clinical Excellence (NICE) in the United Kingdom. This entry reviews the concepts of confidence ellipses and cost-effectiveness acceptability curves (CEACs) for the presentation of uncertainty surrounding the cost-effectiveness, detailing their construction, use, and interpretation. The concept of the cost-effectiveness acceptability frontier (CEAF) is also introduced.

Confidence Ellipse

A confidence ellipse provides a visual representation of the uncertainty surrounding costs and effects (or indeed any two variables). The ellipse provides a region on the cost-effectiveness plane that should contain x% (e.g., 95%) of the uncertainty. By varying x, a series of contour lines can be plotted on the cost-effectiveness plane, each containing the relevant proportion of the cost and effect pairs. Figure 1 illustrates 95%, 50%, and 5% confidence ellipses.

Construction of the confidence ellipse requires the assumption that the costs and effects follow a bivariate normal distribution, that is, for each value of cost, the corresponding values of effect are normally distributed (and vice versa).

The drawback with the confidence ellipse is that while it presents the uncertainty around the costs and effects, it does not deal with the uncertainty surrounding the incremental cost-effectiveness ratio (ICER). One solution to this is to use the boundaries of the relevant confidence ellipse to approximate confidence intervals (e.g., 95%) for the ICER. This interval is given by the slopes of the rays from the origin, which are just tangential to the relevant ellipse (identified in Figure 2). Note that these will be overestimates of the confidence interval.

The particular shape and orientation of the confidence ellipse will be determined by the covariance

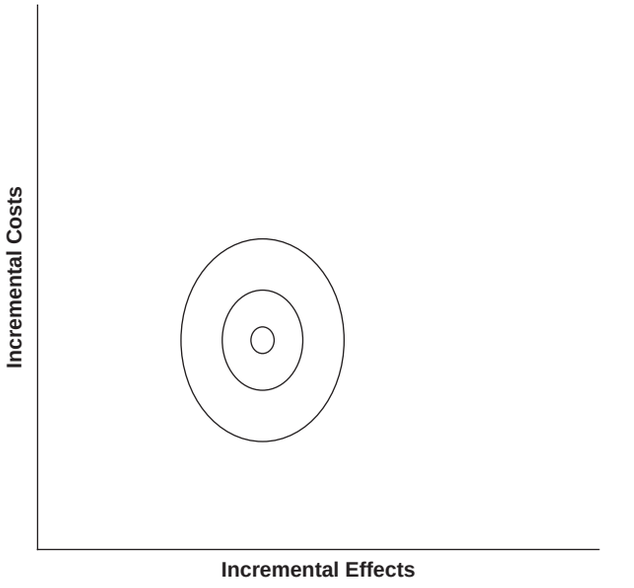


Figure 1 Confidence ellipses on the cost-effectiveness plane

of the costs and effects. This will in turn affect the confidence intervals estimated from the ellipse. Figure 3 illustrates the influence of the covariance on the confidence ellipse and the confidence limits.

Cost-Effectiveness Acceptability Curves

In contrast, the acceptability curve (or cost-effectiveness acceptability curve [CEAC]) focuses on the uncertainty surrounding the cost-effectiveness. An acceptability curve provides a graphical presentation of the probability that the intervention is cost-effective (has an ICER below the cost-effectiveness threshold) compared with the alternative intervention(s), given the data, for a range of values for the cost-effectiveness threshold. It should be noted that this is essentially a Bayesian view of probability (probability that the hypothesis is true given the data) rather than a frequentist/classical view of probability (probability of getting the data, or data more extreme, given that the hypothesis is true). It has been argued that this is more appropriate to the decision maker, who is concerned with the probability that the intervention is cost-effective (hypothesis is correct) given the cost-effectiveness results. However, a frequentist interpretation of the acceptability curve has been suggested, as the $1 - p$ value of a one-sided test of significance.

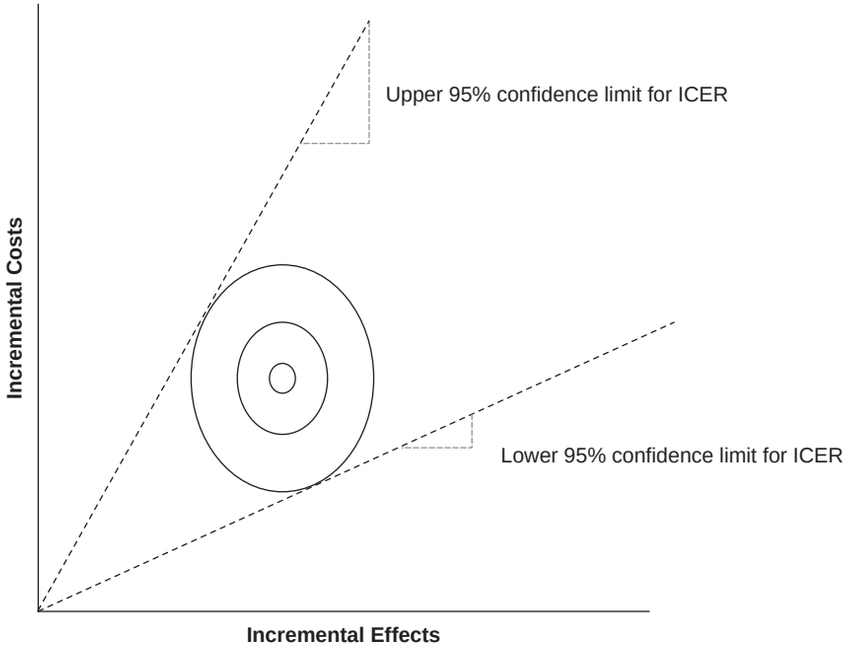


Figure 2 Estimation of the confidence interval from the confidence ellipse

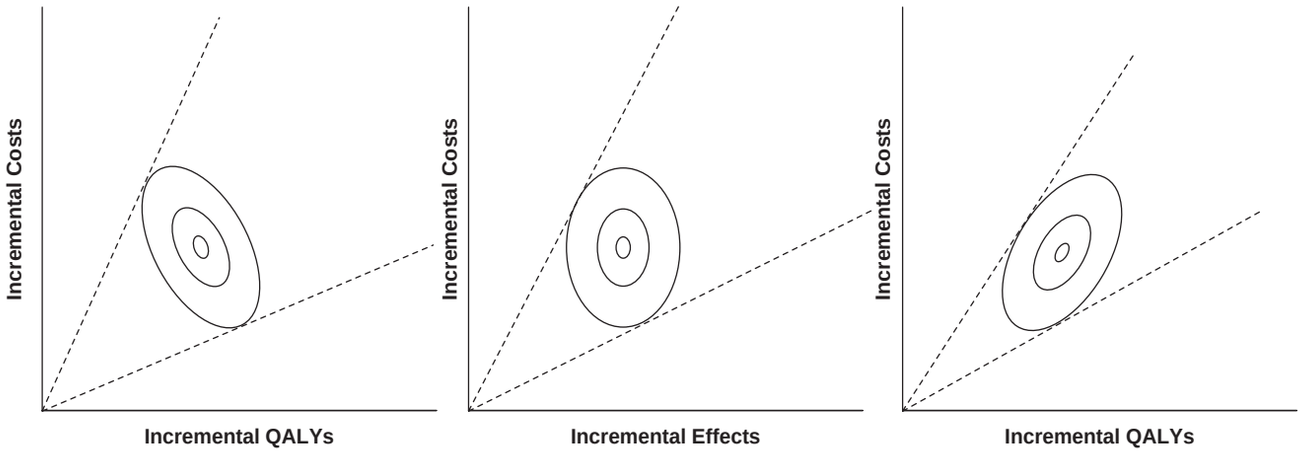


Figure 3 Covariance and the confidence ellipse: (a) negative covariance between cost and effect, (b) independent cost and effect (0 covariance), and (c) positive covariance between cost and effect

Acceptability curves were originally introduced as an alternative to presentation of confidence intervals around the ICER, given the methodological difficulties involved with determining confidence intervals for ratio statistics, including the nonnegligible probability of a small or nonexistent effect difference that would cause the ICER to be undefined and make the variance intractable. Figure 4 presents a CEAC for an intervention.

Constructing a CEAC

The CEAC is derived from the joint distribution of incremental costs and incremental effects. When cost and effect data originate from a clinical trial, the joint distribution is generally determined through nonparametric bootstrapping. When a model has been used, probabilistic sensitivity analysis (Monte Carlo simulation) can be used to translate the uncertainty surrounding the model parameters into uncertainty in costs and effects. As such, the construction of the acceptability curve has no requirement for parametric assumptions regarding the joint distribution of costs and effects.

For any specified cost-effectiveness threshold, the probability that the intervention is cost-effective is calculated simply as the proportion of the cost and effect pairs (plotted on the cost-effectiveness plane) lying below a ray with slope equal to the specific threshold. Since the cost-effectiveness threshold is generally not explicitly defined, this calculation is repeated for different values of the cost-effectiveness threshold. The process usually starts with the threshold = 0 (indicating that society cares only for reduced costs) and ends with the threshold = ∞ (indicating that society cares only for increased effects). The acceptability curve is constructed by plotting probabilities (y-axis) against the cost-effectiveness threshold (x-axis). Figure 5 illustrates the process of constructing the acceptability curve illustrated in Figure 4.

Rules for the CEAC

1. The value at which the acceptability curve cuts the y-axis (i.e., when cost-effectiveness threshold = 0) is determined by the extent of the joint distribution that falls below the x-axis on the cost-effectiveness plane (i.e., involves cost savings). If any of the joint distribution involves cost savings, the curve will not start at 0.

2. The value to which the acceptability curve asymptotes (as the cost-effectiveness threshold approaches infinity) is determined by the extent of the joint distribution falling to the right of the y-axis (i.e., involving increased effects). If any of the joint distribution involves negative effects, the curve will not asymptote to 1.

3. The shape of the acceptability curve will depend solely on the location of the joint distribution within the incremental cost-effectiveness plane.

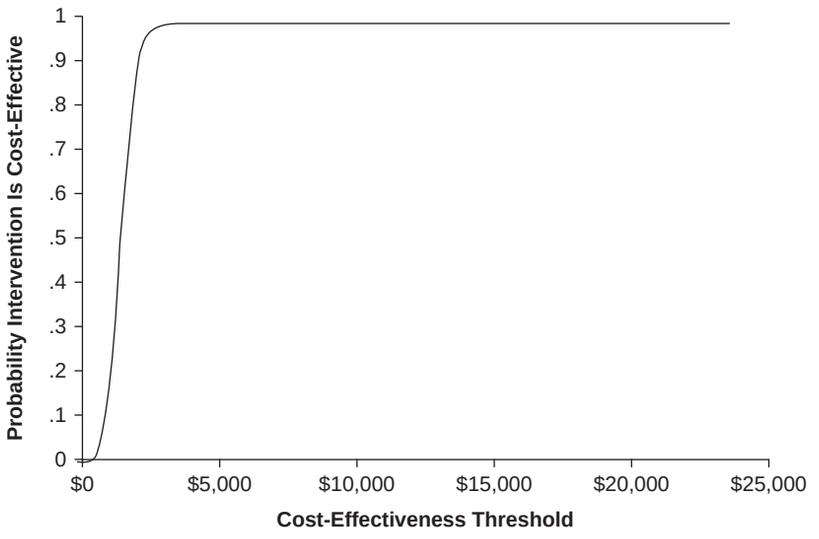


Figure 4 Cost-effectiveness acceptability curve

Incremental cost-effect pairs that fall in the northwest quadrant are never considered cost-effective and, therefore, are never counted in the numerator of the estimate. Incremental cost-effect pairs that fall in the southeast quadrant are always considered cost-effective and, therefore, are always counted in the numerator of the estimate. As the threshold increases from zero to infinity, incremental cost-effect pairs in the northeast and southwest quadrants may or may not be considered cost-effective (and therefore included in the numerator) depending on the value of the threshold. As such, the acceptability curve is not necessarily monotonically increasing with the cost-effectiveness threshold, and therefore, it does not represent a cumulative distribution function.

Interpreting and Misinterpreting the CEAC

For a specific cost-effectiveness threshold (x-axis), the acceptability curve presents the probability (read off on the y-axis) that the data are consistent with a true cost-effectiveness ratio falling below that value. It presents a summary measure of the joint uncertainty in the estimate of incremental cost-effectiveness, thus providing the decision maker with a measure of the uncertainty associated with the selection of a particular intervention as cost-effective.

Note that the acceptability curve *should not* be read in the opposite direction (i.e., from the y-axis to the x-axis) as this would imply that the cost-effectiveness threshold is flexible and determined by the required probability level (confidence) rather than externally set and based on society's willingness to pay for health effects. For example, the curve should not be read to determine the cost-effectiveness threshold (x-axis) required to provide at least a .95 probability that the intervention is cost-effective ($p < .05$).

Statements concerning the acceptability curve should be restricted to those regarding the uncertainty of the estimate of cost-effectiveness. An acceptability curve *should not*, in general, be used to make statements about whether the intervention is actually cost-effective compared with the alternative(s).

Presenting Multiple Acceptability Curves

There are two situations in which it may be useful and/or necessary to present multiple acceptability curves: (1) where there are different patient subgroups and (2) where there are multiple interventions to be compared. The methods for handling and displaying these two situations are very different.

Multiple Patient Subgroups

With analyses involving different patient subgroups, the cost-effectiveness of the intervention for each subgroup is entirely independent from that for other subgroups. Each acceptability curve presents the probability that the intervention is cost-effective compared with the comparator(s), given the data, for a particular subgroup. As such,

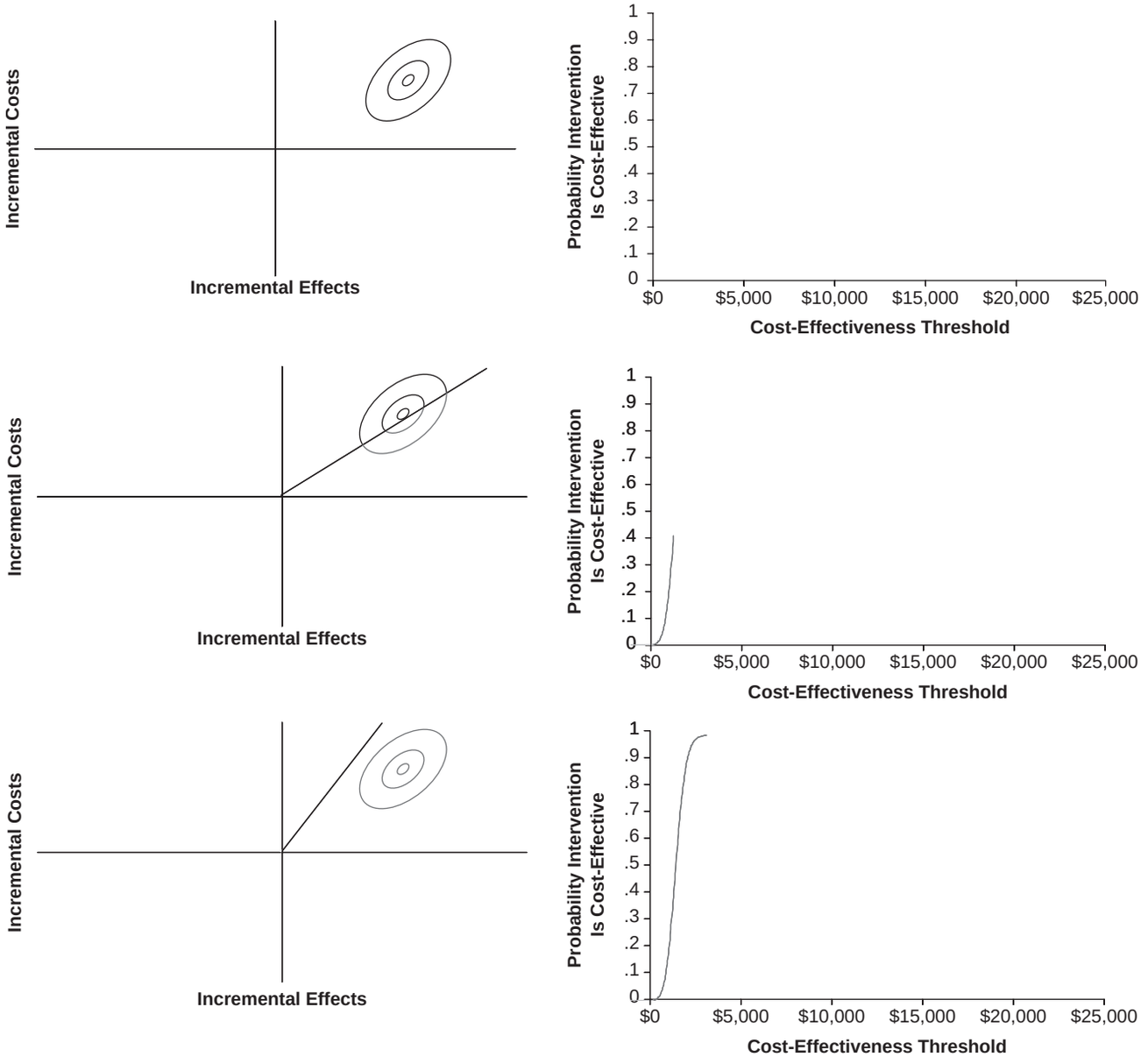


Figure 5 Creating the cost-effectiveness acceptability curve

each acceptability curve should be read and interpreted independently. Such curves can be plotted separately or, to save space, together, but the interpretation remains the same.

Multiple Interventions

With analyses involving multiple (mutually exclusive) interventions, the cost-effectiveness of each intervention must be compared with the available alternatives and assessed simultaneously. The same is true of the probability that each intervention is cost-effective compared with the available alternatives, given the data. With mutually exclusive, collectively exhaustive interventions, the vertical sum of the probabilities must equal 1 for every value of the cost-effectiveness threshold (i.e., one of the interventions must be cost-effective). Therefore, in contrast to the multiple subgroup case, when presenting acceptability curves for multiple (mutually exclusive) interventions, the curves should be read and interpreted together. However, this presentation of multiple acceptability curves can cause confusion with interpretation and lead to a temptation to

identify the cost-effective intervention from the acceptability curves, as that with the highest probability for each cost-effectiveness threshold. As stated above, the acceptability curves present only the probability that the intervention is cost-effective compared with the alternative(s), given the data. They do not identify whether the intervention, or which intervention, is cost-effective. This is identified through comparison of the ICER with the cost-effectiveness threshold, with the cost-effective intervention identified as that with the largest ICER falling below the cost-effectiveness threshold.

Acceptability Frontier

One method suggested to avoid the problem of misinterpretation associated with multiple acceptability curves is the presentation of a CEAF. The CEAF is created by graphing the probability that the intervention is cost-effective only over the range at which it is identified as such on the basis of the ICER. As the name suggests, this provides a frontier produced from the relevant sections of the individual acceptability curves. It should be noted that the appropriate construction of the CEAF requires that the cost-effective intervention is identified for each value of the threshold and then the probability is plotted for this intervention for this threshold. Breaks in the acceptability frontier may occur at the point where the cost-effective intervention changes (i.e., where the cost-effectiveness threshold equals the ICER between the two interventions). Note that the acceptability frontier, created in this way, is not necessarily the same as that created from the outermost boundary of the individual acceptability curves. Figure 6 presents multiple acceptability curves and the associated CEAF.

Elisabeth Fenwick

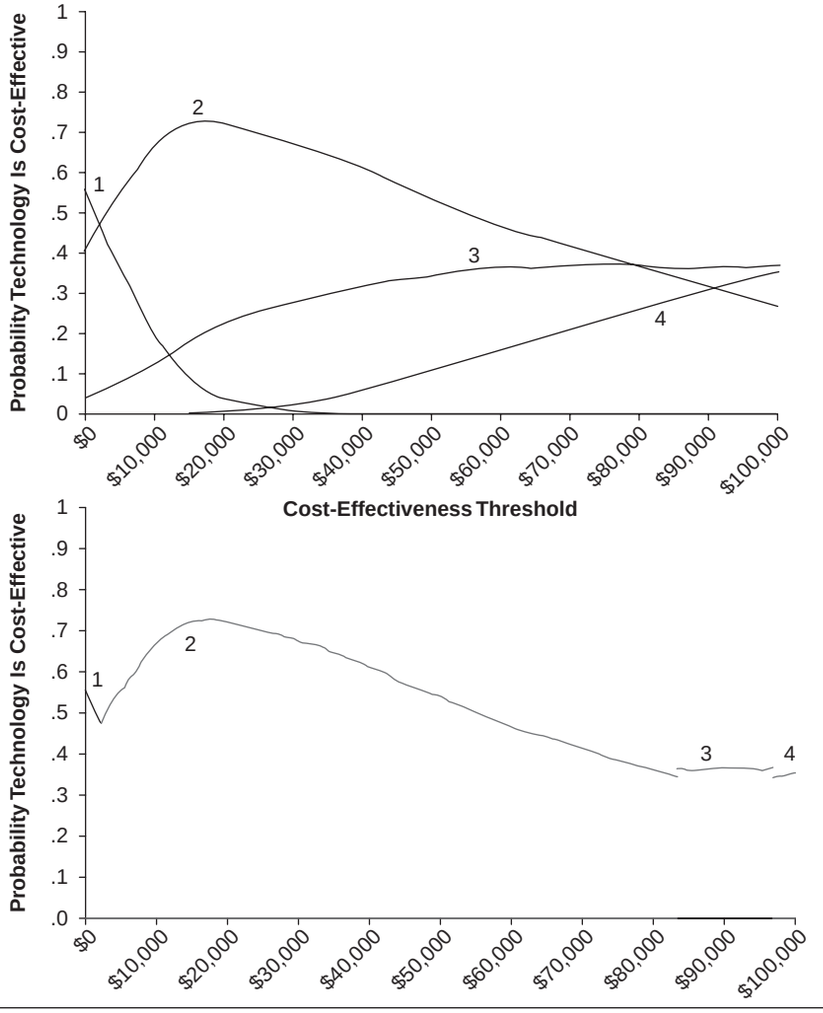


Figure 6 Multiple acceptability curves and associated acceptability frontier

See also Confidence Intervals; Cost-Effectiveness Analysis; Decision Trees: Sensitivity Analysis, Basic and Probabilistic; Managing Variability and Uncertainty; Marginal or Incremental Analysis, Cost-Effectiveness Ratio

Further Readings

Briggs, A. H., & Fenn, P. (1998). Confidence intervals or surfaces? Uncertainty on the cost-effectiveness plane. *Health Economics, 7,* 723–740.

Fenwick, E., Claxton, K., & Sculpher, M. (2001). Representing uncertainty: The role of cost-effectiveness acceptability curves. *Health Economics, 10,* 779–787.

Fenwick, E., O'Brien, B., & Briggs, A. (2004). Cost-effectiveness acceptability curves: Facts, fallacies and frequently asked questions. *Health Economics, 13,* 405–415.

Van Hout, B. A., Al, M. J., Gordon, G. S., & Rutten, F. F. H. (1994). Costs, effects and c/e-ratios alongside a clinical trial. *Health Economics, 3,* 309–319.

Accountability

Accountability refers to the implicit or explicit expectation that one may be called on to justify one's beliefs, feelings, and actions to others. Although most theories of decision making have conveniently assumed that decision makers act as isolated individuals, decision makers, including those in the field of medicine, seldom think and act free from social influences.

Decision making in the field of medicine is fraught with complex, conflicting pressures from various parties, including patients, physicians, hospitals, health policy makers, and insurers, that promote distinct and often competing objectives, such as maximizing life expectancy versus optimizing quality of life, or weighing quality of treatment against economic constraints. Therefore, to best structure accountability relationships and ultimately to improve the quality of decisions in the medical setting, careful analysis of accountability is warranted.

This entry reviews findings from empirical research that addresses the impact of many types of accountability on decision making and attempts to identify the conditions under which accountability will improve decision making.

Many Kinds of Accountability

It is intuitive to think that accountability will breed hard thinking and that thinking harder will translate to thinking better. But according to reviews of the accountability literature, accountability promotes self-critical and effortful thinking only under certain conditions.

Different types of accountability can be distinguished based on the specific nature of justification an individual is expected to provide for his or her decisions: To whom is he or she accountable, for what, and according to what ground rules must he or she justify his or her decisions? For example, a decision maker may be accountable to an audience with known versus unknown views, to authority figures whom the decision maker may perceive as legitimate or illegitimate, and for either the outcome or the process of the decision.

Based on their review of the accountability literature, Jennifer Lerner and Phillip Tetlock reported that decision makers engage in more careful thinking only when they learn prior to forming any opinions about the decision that they will be accountable to an audience (a) whose views are unknown, (b) who is interested in accuracy, (c) who is more interested in processes rather than outcomes, (d) who is reasonably well informed, and (e) who has a legitimate reason for probing the reasons behind decisions. Therefore, simply leading decision makers to expect to justify their decisions to others is insufficient to promote thorough decision making. Instead, organizations and authorities must methodically tailor accountability structures to promote more careful thought processes.

Will Accountability Improve Decision Making?

Although making a decision maker accountable to an unknown audience before the decision is made promotes more careful thought processes, employing this specific kind of accountability by no means ensures improved decision making. Rather, the effects of accountability depend on the types of decisions and the cognitive processes involved, resulting

in some improved decisions, some unchanged decisions, and some degraded decisions.

When Accountability Improves Decision Making

Predecisional accountability to an unknown audience improves decision making to the extent that suboptimal decisions would—under default conditions—result from lack of effort and self-critical attention to the decision process. In other words, as long as improvements in decision making require only greater attention to the information provided, and not acquisition of special skills or training in formal decision rules, the concentrated thinking motivated by accountability pressure will result in thinking better. For example, research has shown that accountable decision makers with a heightened awareness of decision processes made better decisions, specifically, by reducing the tendency for happiness from an unrelated event to elicit heuristic, stereotypic judgments; by reducing blind commitment to a prior course of action in an effort to recoup sunk costs; and by decreasing the likelihood of mindlessly rating a conjunctive event (e.g., shy librarian) as more likely than a simple event (e.g., librarian).

When Accountability Has No Effect on Decision Making

Predecisional accountability to an unknown audience has no effect on decision making if knowledge of formal decision rules (e.g., Bayes's theorem, expected utility theory) that cannot be acquired through increased attention to the decision process is critical for improvements on decision tasks. For instance, accountability had no effect on insensitivity to base rate information; even with increased awareness of their decision process, decision makers often failed to adjust their probability estimates for the frequency of a specific event in some relevant population. As an example, when asked to estimate the probability of a woman having breast cancer given a positive mammogram with 90% sensitivity and 93% specificity, most participants failed to take the base rate of breast cancer in the woman's age group (.8%) into account even when it was clearly provided to them, no matter how hard they were pressured to think.

When Accountability Degrades Decision Making

Predecisional accountability to an unknown audience can actually degrade decision making when certain decision-making biases result from using normatively proscribed information or when the option that appears easiest to justify also happens to be a biased option. For example, increased effort in accountable decision makers led them to increase integration of nondiagnostic information into predictions and resulted in dilution of critical diagnostic information.

Decomposing Accountability

To fully understand how accountability influences a given decision context, it is worth recognizing that even the simplest form of accountability necessarily implicates several empirically distinguishable subphenomena: (a) *the mere presence of another person* (decision makers expect that another person will observe their performance), (b) *identifiability* (decision makers expect that what they say or do will be linked to them personally), (c) *evaluation* (decision makers expect that their performance will be assessed by another person according to some normative ground rules and with some implied consequences), and (d) *reason giving* (decision makers expect that they must give reasons for what they say or do). More research is needed to clarify how these phenomena might affect the impact of accountability.

Accountability and Medical Decision Making

Assuming that accountability is a social panacea, people propose accountability as a solution to all sorts of problems. However, research has documented that accountability is not a singular phenomenon that solves every problem. Only highly specialized forms of accountability will elicit increased cognitive effort in decision makers. More cognitive effort is not always beneficial and sometimes makes matters even worse. Moreover, accountability inherently implicates empirically distinguishable subphenomena, which may or may not influence decision makers in a consistent direction. Accountability as a whole is a complex construct that interacts with individual characteristics of the decision maker and properties of the

decision-making environment to produce an array of effects. Decision makers and their superiors should carefully research the decision environment and decision task to use accountability pressure to advantage in medical decision making.

Seunghee Han and Jennifer S. Lerner

See also Bias; Cognitive Psychology and Processes; Decision Quality; Judgment; Social Factors

Further Readings

Lerner, J. S., & Tetlock, P. E. (1999). Accounting for the effects of accountability. *Psychological Bulletin, 125,* 255–275.

Lerner, J. S., & Tetlock, P. E. (2003). Bridging individual, interpersonal, and institutional approaches to judgment and choice: The impact of accountability on cognitive bias. In S. Schneider & J. Shanteau (Eds.), *Emerging perspectives on judgment and decision making* (pp. 431–457). Cambridge, UK: Cambridge University Press.

Tetlock, P. E. (1999). Accountability theory: Mixing properties of human agents with properties of social systems. In J. Levine, L. Thompson, & D. Messick (Eds.), *Shared cognition in organizations: The management of knowledge* (pp. 117–137). Hillsdale, NJ: Erlbaum.

Advance Directives and End-of-Life Decision Making

Advance directives are oral or written statements given by competent individuals regarding the medical treatment they would like to receive should an incapacitating injury or illness preclude their ability to make or express their own decisions. They are most often used to make decisions when a person is near the end of life, and difficult choices must be made about the use or withdrawal of life-sustaining medical treatment.

Rapid advances in medical technology over the past several decades have made end-of-life decision making an increasingly important and complex challenge for patients, their families, and healthcare professionals. Advance directives play a role in many end-of-life decisions, and their use is encouraged by medical professionals and supported by state and federal law. This entry describes the main types of advance directives, their social and legal history, some of their limitations as aids to effective end-of-life decision making, and some strategies suggested for addressing these limitations.

Types of Advance Directives

There are two primary types of advance directives. *Instructional advance directives*, also known as living wills, contain instructions about the type of life-sustaining treatment an individual would like to receive should he or she become incapacitated. Such instructions can range from legal documents prepared with the help of an attorney to verbal statements made to a family member or a physician. They can be general and express values and goals that the individual feels should guide medical care (e.g., emphasize quality over quantity of life) or relevant religious values. Or they can be specific and carefully delineate particular medical treatments to be used or withheld in particular medical conditions. Most often, instructional directives express a desire to withhold aggressive life-sustaining treatments, but they can also be used to request such treatments. In addition, they can specify preferences regarding pain management, organ donation, or dying at home as opposed to in a hospital.

Proxy advance directives designate another person as a surrogate decision maker, or a proxy, for the patient should he or she become incapacitated. Proxy directives are also known as *durable powers of attorney for healthcare* and *surrogate appointments*. The surrogate decision maker is usually a spouse or another close family member. Proxy directives convey the legal right to make treatment decisions but do not necessarily contain explicit guidance regarding what those treatments should be.

Advance directives can be created without using any preprepared forms, but the majority of U.S. states provide standard forms that follow specific state statutes. Verbal statements are also

considered legal advance directives, especially if recorded by a medical professional in a patient's chart.

Another common kind of instructional advance directive is a *Do Not Resuscitate* (DNR) order, which is recorded in a medical chart and indicates a desire to not receive cardiopulmonary resuscitation (CPR). Because resuscitating treatments often fail, such orders are also sometimes called *Do Not Attempt Resuscitation* (DNAR) orders. Also, because decisions besides those involving resuscitation must often be made, a more comprehensive type of medical order form called *Physician Orders for Life-Sustaining Treatment* (POLST) has recently been developed and adopted for use in several states. POLST forms record a patient's wishes for a number of different life-sustaining treatments and require both patients and physicians to sign, indicating that they have discussed these preferences.

Advance Directives Versus Physician-Assisted Suicide

Advance directives should not be confused with the more controversial issue of physician-assisted suicide. Advance directives involve choices about whether to accept or refuse particular kinds of life-sustaining medical treatment in the event of incapacitation. Physician-assisted suicide involves a competent, terminally ill person asking a physician to knowingly and intentionally provide the means to end his or her life. The use of advance directives to refuse unwanted medical treatment near the end of life is endorsed widely by medical associations and supported by U.S. state and federal law. Advance directives have achieved similar levels of acceptance in a number of European countries. In contrast, physician-assisted suicide is much more controversial and, at this time, is legal only in the state of Oregon and a few European countries (e.g., the Netherlands) under very narrow sets of conditions.

The Social and Legal History of Advance Directives

The concept of advance directives emerged in the late 1960s as medical technology made it increasingly possible to prolong the lives of seriously ill individuals, especially individuals with minimal cognitive functioning or severe and chronic pain, who have little or no hope for ultimate recovery. Many people view the use of life-sustaining medical treatment in such situations as not so much extending life as extending the process of dying. This created a challenge to the "technological imperative" that physicians should use all means at their disposal to prolong life. The concept of advance directives was thus created to allow people to exert some control over the medical treatment they receive at the end of their lives.

Advance directives were a response to a practical problem. At the time difficult medical decisions must be made about the use of life-sustaining treatments, many patients are already too sick to decide for themselves. In 1969, attorney Luis Kutner suggested that individuals too ill to make decisions for themselves could maintain their ability to influence the use of life-sustaining medical treatments by documenting treatment wishes prior to incapacitation in what he termed a "living will."

The issues of advance directives and end-of-life decision making did not enter public consciousness, however, until the controversial 1976 court case of *In re Quinlan*. In that case, the New Jersey Supreme Court considered the dilemma of Karen Ann Quinlan, a young woman who suffered severe brain damage after mixing alcohol and tranquilizers at a party and was left in a persistent vegetative state. Her parents sought to remove her from the respirator that was maintaining her life, but hospital administrators asked for a court ruling on the matter because of concerns about legal liabilities. The court granted her parents' request for removal of the respirator, finding that it infringed on Quinlan's right to privacy protected under the Constitution. The decision was important because it concluded that not only did a competent person have a constitutionally protected right to refuse life-sustaining treatment but that this right was not diminished by Quinlan's incapacitation. The court went on to say that while Quinlan could obviously not exercise this right herself, her parents could on her behalf, using their "best judgment" on how she would decide for herself.

An even more crucial legal decision supporting the use of instructional advance directives was *Cruzan v. Director, Missouri Department of Health*, decided by the U.S. Supreme Court in 1990. The case involved 24-year-old Nancy Cruzan, who suffered a car accident that left her in a persistent vegetative state with no hope for recovery. Cruzan's parents sought legal action to remove her from life support but were opposed by Missouri state officials. The U.S. Supreme Court confirmed not only Cruzan's constitutionally protected right to refuse medical treatment but also a state's right to set its own standard for determining sufficient evidence of an incompetent person's wishes. In this case, Missouri's standard required "clear and convincing evidence" of an incompetent patient's prior wishes, and an instructional advance directive is often seen as the best method of meeting this strict evidentiary standard.

The controversy surrounding the Cruzan case helped spur important legislation, and in 1990, the U.S. Congress passed the Patient Self-Determination Act. The act stipulates that all hospitals receiving Medicaid or Medicare reimbursement must inform patients of (a) their right to accept or refuse treatment, (b) their rights under existing state laws regarding advance directives, and (c) any policies the institution has regarding the withholding or withdrawing of life-sustaining treatments. Institutions are also required to engage in ongoing educational activities for both their employees and the general public regarding the right to accept or refuse treatment and the opportunity for drafting or signing advance directives. Moreover, state legislation has been passed over the past two decades making some form of advance directives (instructional, proxy, or both) legal in all 50 states and the District of Columbia.

More recently, the case of Theresa Marie (Terri) Schiavo brought intense worldwide media attention to the issue of end-of-life decision making. Schiavo was a 26-year-old Florida housewife when her heart unexpectedly stopped in 1990, leaving her immobile and uncommunicative for the next 15 years. Schiavo left no advance directive and members of her immediate family disagreed vehemently about whether or not she should be removed from the machines that were supplying her with food and fluids. Although a series of court decisions had sided with the arguments of Schiavo's husband, Michael, that she should be removed from life support, her parents and siblings continued to battle, both in legal court and in the court of public opinion, arguing that she would want to be kept alive in her current condition, and even that she was currently responsive to external stimulation. Schiavo died on March 31, 2005, 13 days after her feeding and fluid tubes were ordered disconnected by a Florida trial judge. The case raised public awareness of advance directives and the complex and emotionally charged nature of end-of-life decision making.

Limitations of Advance Directives

A number of researchers and ethicists now express skepticism regarding the effectiveness of advance directives to improve end-of-life medical decision making. The challenges of making decisions for incapacitated individuals are complex and multifaceted. End-of-life decisions involve multiple individuals, including the patient, his or her loved ones, and physicians. Information must be passed from one individual to another, and each individual has motivations that may conflict and decision-making limitations that must be overcome. Of particular concern are low completion rates of advance directives (particularly among some ethnic groups), the stability of preferences for life-sustaining treatment across changes in an individual's psychological and medical condition, and the effectiveness and accuracy of surrogate decision making.

The first challenge facing the use of advance directives is that most people do not have one. Estimates suggest that fewer than 25% of U.S. adults have an advance directive. Completion rates are not substantially higher for individuals with serious chronic diseases, and interventions designed to increase the rate of advance directive completion have shown limited effectiveness. Completion rates are particularly low for some ethnic groups, including African Americans, Latinos, and Native Americans. One source of cultural differences may be differential value placed on autonomy. In Western philosophy, family members are generally viewed as a source of emotional support, not active participants in the decision-making process. In

many East Asian and other cultures, however, the importance of filial duty or protecting the elderly may lead a family to make decisions for a fully competent adult and withhold information about prognosis. In addition, in traditional Hawaiian, Chinese, and Japanese cultures, it is commonly believed that talking about death may bring on death or spiritual pollution. Planning ahead via advance directives is often resisted by individuals with these cultural backgrounds because it is seen as interfering with deeply held cultural traditions and the natural course of life and death.

A second problem with instructional advance directives in particular concerns the appropriateness of projecting treatment wishes of competent individuals onto future states of incompetence. Preferences for life-sustaining treatment have been found to be highly context dependent and can be altered by an individual's current psychological and physical state, as well as the way questions soliciting treatment preferences are framed. People may have difficulty imagining what life would be like in severely impaired health states. Research suggests that almost one third of individuals change their preferences about any given life-sustaining medical treatment over a period of 1 to 2 years. Moreover, the majority of individuals whose life-sustaining treatment preferences change over time are unaware of these changes and, thus, are unlikely to revise their advance directives. These issues raise concerns about whether an instructional directive completed years before an incapacitating illness can be taken as an accurate representation of a patient's current treatment wishes.

A parallel concern exists for the usefulness of proxy directives. Researchers have examined the ability of potential surrogate decision makers to predict a close relative's life-sustaining treatment wishes. In these studies, an individual records his or her treatment preferences for various end-of-life scenarios (e.g., irreversible coma, end-stage cancer, debilitating stroke), and a surrogate decision maker (e.g., a loved one or physician) is asked to predict those preferences. Research has consistently shown that surrogate accuracy in predicting a patient's life-sustaining treatment wishes rarely exceed chance levels. Surrogate decision makers have been found to show at least two types of prediction biases. The first is an overtreatment bias, that is, predicting that family members will want life-sustaining treatment more often than they really do, thus choosing to "err on the side of life." This bias is weaker in predictions made by physicians, who have sometimes been found to show an undertreatment bias. The second is a projection bias in which surrogates (both family members and physicians) have been found to err by assuming that individuals will have wishes for life-sustaining treatment that are similar to their own.

Last, it should be noted that decisions about treatment for a loved are not purely rational ones. Individuals who are placed in the position of being directly responsible for taking the action that ends the life of a loved one may experience strong emotional conflict. Thus, even if a surrogate knows full well that a loved one does not want to receive life-sustaining treatment, the surrogate may find it difficult to honor that wish. Another point of conflict may occur if the patient's known wishes conflict with religious or other deeply held values of the surrogate, as well as if different family members disagree about what the patient would have wanted.

Improving End-of-Life Decision Making

Although research has uncovered a number of important limitations of advance directives, several strategies have been advocated that may improve their effectiveness.

Studies show that when asked about their personal wishes, most individuals express generally positive attitudes about planning for the end-of-life, but many express ambivalence toward completing specific instructional directives and, instead, seem more positively inclined toward informal discussion that focuses on general values and goals. Many individuals are comfortable leaving end-of-life medical decisions to their families and indicate that in the event of a disagreement between their own documented preferences and the opinions of their loved ones, their family's rather than their own directions should be followed. As noted above, such attitudes are particularly pronounced in some cultural groups. Therefore, broad-based attempts to encourage healthy people to document increasingly specific instructional advance directives may be misguided. Instead, some scholars have argued that it is better to focus on encouraging the completion of proxy advance directives, and virtually all agree that people should be encouraged to

view completion of an advance directive document as only one part of a broader strategy of advance care planning that includes maintaining an ongoing discussion about end-of-life treatment wishes with loved ones and physicians.

Another approach that attempts to overcome the hypothetical nature of general advance directives is the use of disease-specific advance directives. These are directives developed for patients with a particular medical condition (e.g., AIDS) and allow them to document their wishes for the specific decisions that individuals with their condition are most likely to face. Proponents of this approach argue that because the patient already has some experience with the illness, treatment choices are less hypothetical and, thus, more durable and authentic.

Finally, some shortcomings of standard advance directives may be overcome by the use of medical orders for life-sustaining treatment. Like disease-specific advance directives, medical orders can be written based on the individual's current medical condition and, thus, may be more accurate and up-to-date expressions of end-of-life wishes than generic directives completed months or years prior to hospitalization. Advocates of the POLST program argue that in contrast to standard instructional advance directives that are typically more philosophical reflections of an individual's preferences about an unknown future, the POLST is immediately actionable and can be followed by licensed medical staff such as nursing facility nurses and emergency medical technicians. Some recent research supports the effectiveness of the POLST program in ensuring that patients' treatment preferences are honored.

Peter H. Ditto and Spassena Koleva

See also Biases in Human Prediction; Bioethics; Context Effects; Cultural Issues; Decision Making in Advanced Disease; Surrogate Decision Making

Further Readings

Brett, A. S. (1991). Limitations of listing specific medical interventions in advance directives. *Journal of the American Medical Association, 266,* 825–828.

Buchanan, A. E., & Brock, D. W. (1990). *Deciding for others: The ethics of surrogate decision making*. Cambridge, UK: Cambridge University Press.

Cicirelli, V. G. (1997). Relationship of psychosocial and background variables to older adults' end-of-life decisions. *Psychology and Aging, 12,* 72–83.

Ditto, P. H., Danks, J. H., Smucker, W. D., Bookwala, J., Coppola, K. M., Dresser, R., et al. (2001). Advance directives as acts of communication: A randomized controlled trial. *Archives of Internal Medicine, 161,* 421–430.

Ditto, P. H., Hawkins, N. A., & Pizarro, D. A. (2005). Imagining the end of life: On the psychology of advance medical decision making. *Motivation and Emotion, 29,* 475–496.

Emanuel, L. L., Danis, M., Pearlman, R. A., & Singer, P. A. (1995). Advance care planning as a process: Structuring the discussions in practice. *Journal of the American Geriatrics Society, 43,* 440–446.

Fagerlin, A., Ditto, P. H., Danks, J. H., Houts, R., & Smucker, W. D. (2001). Projection in surrogate decisions about life-sustaining medical treatment. *Health Psychology, 20,* 166–175.

Hickman, S. E., Hammes, B. J, Moss, A. H., & Tolle, S. W. (2005). Hope for the future: Achieving the original intent of advance directives. *Hastings Center Report Special Report, 35*(6), S26–S30.

Kwak, J., & Haley, W. E. (2005). Current research findings on end-of-life decision making among racially or ethnically diverse groups. *The Gerontologist, 45,* 634–641.

The President's Council on Bioethics. (2007). *Taking care: Ethical caregiving in our aging society*. Washington, DC: Government Printing Office.

Allais Paradox

The *independence axiom* of expected utility theory offers a compelling reason for making a decision. According to this axiom, a choice between two alternatives should depend only on features in which alternatives differ but not on features in which the alternatives are equal. Any feature that is the same for both alternatives, therefore, should not influence the choice a rational person makes. For instance, when choosing between two therapies with exactly the same side effects, a rational doctor would ignore these side effects. That is, rational choice is *independent* of the alternatives' shared features.

This axiom seems very intuitive; if two therapies have the same side effects, it does not matter

whether they are small or severe. Hence, rational decision makers base their choices on the distinctive rather than the shared features of the choice alternatives. In the early 1950s, however, French economist Maurice Allais proposed choice problems that challenged the independence axiom as a descriptive principle for risky choice. To illustrate this paradox, known as the Allais paradox, consider the following Allais-type choice problems presented by Adam Oliver: Which of the following would you prefer?

A: Living for 12 years in full health then death, with a chance of 100%

B: Living for 18 years in full health then death, with a chance of 10%

Living for 12 years in full health then death, with a chance of 89%

Immediate death, with a chance of 1%

The majority of people selected Alternative *A* over *B*.

C: Living for 12 years in full health then death, with a chance of 11%

Immediate death, with a chance of 89%

D: Living for 18 years in full health then death, with a chance of 10%

Immediate death, with a chance of 90%

In the second problem, most people chose Alternative *D*, which constitutes a violation of the independence axiom. Table 1 shows why.

Alternatives *A* and *B* share an 89% chance of living for 12 years. Because this shared feature should not influence the choice, it can be cancelled out. Similarly, Alternatives *C* and *D* share an 89% chance of immediate death, which can be cancelled out again. Importantly, after the shared features in each problem (i.e., the bold column in Table 1) have been cancelled out, both problems become identical. A rational decision maker, thus, should choose *A* and *C* or *B* and *D*, but not *A* and *D*.

Table 1 Illustration of the Allais paradox

Alternative	*10 Blue*	*89 Red*	*1 Green*
A	12	**12**	12
B	18	**12**	0
C	12	**0**	12
D	18	**0**	0

Note: The chances in the Allais paradox are symbolized by an urn containing 10 blue balls, 89 red balls, and 1 green ball. Cell entries represent numbers in years living in full health for each alternative in the Allais paradox.

Explaining the Allais Paradox

To account for the Allais paradox, two prominent explanations have surfaced: prospect theory and the priority heuristic. Prospect theory by Daniel Kahneman and Amos Tversky explains the Allais paradox by adding complex nonlinear transformations of utilities and probabilities on top of the expected utility framework. The priority heuristic by Eduard Brandstätter, Gerd Gigerenzer, and Ralph Hertwig is motivated by first principles, so as to avoid ending up with the worst of two minimum consequences. The heuristic consists of three steps (assuming nonnegative consequences). In the first step, people compare the alternatives' *minimum consequences*. They select the alternative with the higher minimum consequence, if this difference is large (i.e., equal to or larger than 10% of the problem's best consequence). Otherwise, they compare the *chances* of the minimum consequences. They select the alternative with the smaller chance of the minimum consequence, if this difference is large (i.e., equal or larger than 10%). Otherwise, they compare the *maximum consequences* and select the alternative with the higher maximum consequence.

In the choice between *A* and *B*, 12 and 0 years represent the minimum consequences. Because this difference is large (i.e., 12 years exceeds 10% of 18 years), people are predicted to select the alternative with the higher minimum consequence, which is *A*. That is, the heuristic predicts the majority choice correctly.

In the second choice problem, the minimum consequences (0 and 0) do not differ. In the second step, the chances of the minimum consequences, 89% and 90%, are compared, and this difference is small (i.e., less than 10 percentage points). The higher maximum consequence, 18 versus 12 years, thus, decides choice, and people are predicted to

select Alternative *D*, which is the majority choice. Together, the pair of predictions makes the Allais paradox.

Oliver asked participants to think aloud while making both decisions. In the first problem, living for 12 years with certainty was often a decisive reason for choosing Alternative *A*. In the second problem, participants most often stated that the difference between a chance of 10% and 11% (i.e., the logical complements to 90% and 89%) was negligible and that the maximum consequence determined their choice. The latter protocol conforms with the priority heuristic, which assumes comparisons across alternatives, but not with prospect theory, which assumes utility calculations within alternatives.

Adhering to the independence axiom, as implied by expected utility theory, is one criterion for rational choice. Avoiding the worst consequence, as implied by the priority heuristic, is another compelling reason. In conclusion, the Allais paradox makes clear that people do not always follow one principle only.

Eduard Brandstätter

See also Bounded Rationality and Emotions; Certainty Effect; Expected Utility Theory; Prospect Theory

Further Readings

Allais, M. (1979). Criticism of the neo-Bernoullian formulation as a behavioural rule for rational man. In M. Allais & O. Hagen (Eds.), *Expected utility hypotheses and the Allais paradox* (pp. 74–106). Dordrecht, the Netherlands: Reidel.

Brandstätter, E., Gigerenzer, G., & Hertwig, R. (2006). The priority heuristic: Making choices without trade-offs. *Psychological Review, 113,* 409–432.

Oliver, A. J. (2003). A quantitative and qualitative test of the Allais paradox using health outcomes. *Journal of Economic Psychology, 24,* 35–48.

Analysis of Covariance (ANCOVA)

Analysis of covariance (ANCOVA) is a statistical model introduced by Sir Ronald Fisher that combines features of analysis of variance (ANOVA) with those of regression analysis. The purpose of ANCOVA is to examine differences between levels of one or more grouping variables on an outcome measure after controlling for variation or differences between populations on one or more nuisance variables. The grouping variable often represents different treatments, the outcome measure is the consequence of those treatments, and the nuisance variable either obscures true treatment differences or is a confounding variable that offers an alternative explanation for differences on the outcome other than the treatments.

The ANCOVA model is often underused in experimental research and misinterpreted in quasi-experimental studies. Researchers may not recognize the benefit of using a covariate to reduce unexplained variation among units to increase statistical power in experimental studies. The inclusion of the covariate can substantially increase the sensitivity of group comparisons or reduce the necessary sample size to detect meaningful population differences. In quasi-experiments, researchers may fail to recognize the limitations of the ANCOVA model and overinterpret the results of the analyses. Because of specification and measurement errors, the statistical model cannot totally compensate for a lack of random assignment and equate the populations being compared. However, when used properly, the ANCOVA model can be an essential statistical tool to identify differences among populations on outcomes of interest.

Research Design

The simplest application of this model involves one grouping variable (*G*) having two levels (e.g., a herbal supplement treatment vs. a placebo), a single outcome variable (*Y*) (e.g., blood pressure) and a single nuisance variable, referred to as a covariate (*X*) (e.g., body mass index [BMI]) measured before the formation of the groups or before the start of the treatments. While the application of the model is identical when groups are formed using a random or nonrandom process, the primary purpose and the interpretation of the results are substantially different. When the formation of the groups is based on a random process (e.g., use of random numbers matched with participant identification numbers to assign individuals to

treatment levels), the research design is referred to as an *experiment* and is often represented as follows:

$$R\ X\ G_1\ Y,$$
$$R\ X\ G_2\ Y,$$

where R represents the random assignment of units to the treatment groups; X is a covariate; G_1 and G_2 represent intervention and placebo groups, respectively; and Y is the outcome of interest. When group formation is based on a nonrandom process (e.g., self-selection) the research design is referred to as a *quasi-experiment* and is often represented as follows:

$$X\ G_1\ Y,$$
$$X\ G_2\ Y,$$

where terms are defined as above.

Data Example

Suppose a sample of 12 overweight patients having high systolic blood pressure volunteered to investigate the usefulness of a herbal supplement over a 2-month trial period. Half of the volunteers are randomly assigned to receive the herbal supplement, while the other half are given a placebo. Before beginning the investigation, each individual's BMI is computed. When the treatment period ends, systolic blood pressure is assessed. Table 1 presents hypothetical data along with means and standard deviations (*SD*s). These data will be used to demonstrate the use and interpretation of the ANCOVA model.

Structural Model

The ANCOVA model that can represent data from both designs can be written as follows:

$$Y_{ij} = \mu + \alpha_j + \beta_{Y|X}(X_{ij} - X) + \varepsilon_{ij},$$

where Y_{ij} is the outcome score for individual i in Group j ($i = 1, \ldots, n_j$; $j = 1, \ldots, J$), μ the grand mean on the outcome measure, α_j the deviation of the mean of population j on the outcome measure from the grand mean, $\beta_{y|x}$ the common regression slope of the outcome on the covariate, X_{ij} the covariate (e.g., pretest) score for individual i in Group j, X the observed grand mean on the covariate measure, and ε_{ij} the model error, a measure of individual differences.

Table 1 Body mass index and systolic blood pressure scores (post) for volunteers receiving an herbal diet supplement or placebo

	Herbal Supplement		*Placebo*	
	BMI	*Post*	*BMI*	*Post*
	50	150	45	147
	40	142	26	135
	7	120	40	152
	32	129	30	128
	45	132	52	165
	36	138	37	140
Mean	38.3	135.2	38.3	144.5
SD	8.45	10.51	9.56	13.16

Before discussing the hypotheses that can be tested with this model, it is very important to note that a common regression slope, β, of *Y* on *X* is assumed for this model. That is, the regression slope of *Y* on *X* is assumed to be identical for all populations being compared. This assumption is important for two reasons. First, if the slopes are not equal, the statistical model is incorrect and the subsequent ANCOVA hypothesis tests may be statistically invalid. Second, unequal regression slopes indicate that there is an interaction between the grouping variable and the covariate. That is, differences between the populations vary depending on the value of the covariate. For example, the difference in blood pressure between a population receiving an herbal supplement and the placebo may only occur for individuals having high BMI scores. In this context, testing for average differences between populations can be inappropriate or misleading. When an interaction is present, alternative analyses (e.g., Johnson-Neyman procedure) may be recommended.

Hypotheses

To determine whether the assumption of a common regression slope is tenable, a statistical test for the equality of the separate regression slopes should be conducted (i.e., H_0: $\beta_{y|x1} = \beta_{y|x2}$) with the criterion for statistical significance set at a slightly elevated level (e.g., $\alpha = .10$ or .15) to reduce the risk of concluding equal slopes when in fact they differ.

For the data in Table 1, the regression slopes of post on BMI are 1.03 and 1.25, respectively. The observed difference between sample estimates is not statistically significant ($F(1, 8) = .230$, $p = .644$). The ANCOVA model is therefore judged appropriate for these data.

If the ANCOVA model is appropriate, two hypotheses can be tested. One hypothesis examines the relationship between the covariate and the outcome measure: H_0: $\beta_{y|x} = 0$. From a substantive perspective, this hypothesis is generally of little interest. Often the covariate and the outcome measures are obtained from the same test administered twice, so a relationship is to be expected. If there is no relationship between the covariate and the outcome measure, then *X* and *Y* are independent and knowledge of *X* is of little statistical value. For the current data set, the pooled or average regression slope is 1.15. The relationship between BMI and postsystolic blood pressure is statistically significant at $\alpha = .05$ ($F(1, 9) = 28.64$, $p = .000$).

A second hypothesis, and the primary hypothesis of interest, that can be tested with the ANCOVA model can be written as: H_0: $\alpha_j = 0$ for all *j*, or equivalently as H_0: ${}_{\text{adj}}\mu_1 = {}_{\text{adj}}\mu_2 = \cdots = {}_{\text{adj}}\mu_j$. The exact meaning of this hypothesis depends on whether the research design is experimental or quasi-experimental. An adjusted mean for population *j* is defined as

$$ {}_{\text{adj}}\mu_j = \mu_{Y_j} - \beta_{Y|X}(\mu_{X_j} - \mu_X), $$

where μ_{Y_j} and μ_{X_j} are the means for population *j* on the outcome and covariate measures, respectively, and $\mu_{X_{..}}$ is the grand mean across all populations on the covariate.

If two populations are compared, the hypothesis may be written as

$$ H_0 : \text{adj}\mu_1 - {}_{\text{adj}}\mu_2 = 0, $$

or

$$ H_0 : (\mu_{Y_1} - \mu_{Y_2}) - \beta_{Y|X}(\mu_{X_1} - \mu_{X_2}) = 0. $$

The hypothesis on difference between the adjusted population means can be seen as a hypothesis on the difference between the population means on the outcome measure minus the product of the difference between the population covariate means and $\beta_{Y/X}$. Where $\beta_{Y/X}$ is a measure of the degree to which the covariate can predict the outcome measure. An estimate of the difference between adjusted population means is provided by substituting sample estimates for the parameters in the hypothesis:

$$ (Y_1 - Y_2) - b_{Y|X}(X_1 - X_2). $$

Experimental Design

When units are randomly assigned to the groups, there would be no difference between the populations on the covariate measure, $\mu_{x1} - \mu_{x2} = 0$, and no true adjustment is made nor is one necessary. In an experiment, the hypothesis on the adjusted population means is identical to the hypothesis

tested in a posttest-only design using analysis of variance. The equality of means on the covariate measure refers to only the populations, not the sample means. Sample means typically differ slightly and small differences between adjusted and unadjusted sample outcome means are generally observed. But hypotheses are statements regarding populations, not samples, so the small differences in sample means can be safely ignored. In the present example, sample BMI means are identical (i.e., $X_1 = X_2 = 38.3$).

Quasi-Experimental Design

In quasi-experimental studies, populations being compared typically differ on the covariate measure $\mu_{x1} - \mu_{x2} \neq 0$. For example, individuals who choose to take herbal supplements may also exercise more than individuals who do not take the supplements. The difference in blood pressure between the two populations may be related to the amount of exercise rather than the herbal supplement. With the ANCOVA model, differences on the outcome measure can to some extent be adjusted for the difference on the covariate. The question, however, is whether this adjustment is sufficient. The answer is generally no. There are two problems when the populations being compared are not equivalent on all relevant variables that could explain differences on the outcome variable other than the treatments. First, if populations differ on one variable, X, they are likely to differ on other variables as well, and these additional variables might also provide an alternative explanation for population differences on the outcome. It is possible to extend the ANCOVA model to include multiple covariates, but it is impossible to know and to specify all the other relevant confounding variables. This is known as the specification error problem. Second, even if the populations differed on only one variable, X, the adequacy of the adjustment would depend on the estimation of the population slope $\beta_{y|x}$. The reliability (i.e., consistency) with which the covariate is measured affects the estimate of $\beta_{y|x}$. The relationship between $\beta_{Y|X}$ and the sample estimate $b_{Y|X}$ is $b_{Y|X} = \beta_{Y|X}\,\rho_{XX}$, where ρ_{xx} is the reliability of the covariate measure (e.g., BMI). Because the covariate is never perfectly reliable, measurement error leads to an underestimation of the relationship between X and Y, and the pooled regression slope, $b_{Y|X}$, is too small and the difference in outcome means is underadjusted. This is known as the measurement error problem. In our example, the pooled slope was computed as $b_{Y|X} = 1.15$. If the BMI is measured with .70 reliability, the correct adjustment should have been 1.64. Consequently, the adjustment is insufficient, and it is not possible to attribute differences in the outcome variable solely to the treatment. In the current example, the mean BMI score for both groups was identical, so the underestimation of the relationship is irrelevant. No adjustment to postsystolic blood pressure is needed.

The hypothesis regarding the grouping variable tested with the ANCOVA model is therefore different when the research design is experimental or quasi-experimental. In an experimental design, the hypothesis tested is unambiguous. Differences in the outcome variable can be attributed to differences in the grouping variable. But in a quasi-experimental study, because of measurement error with the covariate and the inability to specify and measure all relevant confounding variables, differences between populations on the outcome measure cannot be attributed solely to differences in the grouping variable. The ANCOVA model cannot be used to completely compensate for a lack of random assignment, and the results of the analysis must be interpreted cautiously.

Statistical Power

As discussed above in an experimental study, the ANCOVA and ANOVA models test the hypothesis that the population means on the outcome variable are identical. It might then be asked, why go to the trouble and expense of collecting additional data prior to the formation of the groups? The answer is greater sensitivity (i.e., statistical power) to detect a difference between populations. Both ANOVA and ANCOVA models compute a test statistic, F, by taking the ratio of the variation among group means multiplied by n, the common group size (e.g., $n = 6$) to the unexplained variation of units within the groups. Because in an experiment adjusted and unadjusted means are, within sampling error, equivalent, the two statistics differ only in terms of the unexplained variation among the units. The unexplained variation is individual differences attributable to multiple causes (e.g.,

initial blood pressure, BMI, activity levels). With the ANOVA model, the unexplained variation of the units in the populations being compared on the outcome measure can be represented as $\sigma^2_{Y|G}$. If a covariate is available and is used, it can explain some of the unexplained variation in the outcome measure, and the remaining variation for the ANCOVA model can be written as $\sigma^2_{Y|GX} = \sigma^2_{Y|G}(1 - \rho^2)$, where ρ^2 is the population correlation between the covariate and the outcome measure. The greater the correlation between the two measures, the smaller the unexplained variation in the ANCOVA model relative to the ANOVA model, $\sigma^2_{Y|GX} < \sigma^2_{Y|G}$. The smaller the unexplained variation, the more sensitive the analysis to a true population difference between the intervention and the placebo. This sensitivity is manifested in a larger computed *F* statistic.

In the current data set, if the BMI is ignored, $\sigma^2_{Y|G}$ is estimated using the average within-group variance on postsystolic blood pressure,

$$141.8 = \frac{(10.51)^2 + (13.16)^2}{2}.$$

Including BMI scores as a covariate, $\sigma^2_{Y|GX}$ is estimated as 37.7. Ignoring the BMI scores, the observed difference between posttreatment means (135.2 vs. 144.5) is not statistically significant ($F(1, 10) = 1.843$, $p = .204$). But after considering individual differences in the BMI scores, the difference between means on the postsystolic blood pressure is statistically significant ($F(1, 9) = 6.936$, $p = .027$).

Effect Size

The statistical evaluation of α_j in the ANCOVA model is useful in determining whether observed difference in the adjusted sample means represent a true difference in population means or is an artifact of sampling error (i.e., chance differences between units in the samples studied). But this analysis provides no information on the magnitude of the true difference. Two useful indices of effect size are the standardized mean difference and η^2.

The standardized mean difference (δ) is useful when comparing two populations, and it defines the difference in population means in terms of the population standard deviation on the outcome measure:

$$\delta = \frac{{}_{\text{adj}}\mu_{G_1} - {}_{\text{adj}}\mu_{G_2}}{\sigma_{Y|G}}.$$

A sample estimate of δ is provided by using sample estimates of the parameters:

$$d = \frac{{}_{\text{adj}}Y_{G_1} - {}_{\text{adj}}Y_{G_2}}{S_{Y|G}},$$

where $S_{Y|G}$ equals the pooled within-group standard deviation on the outcome measure. Note that when computing the standardized-mean difference, the denominator includes the variation associated with the covariate. For the current data, *d* is computed to equal –.78 [(135.2–144.5)/√141.8]. The herbal supplement reduced systolic blood pressure .78 standard deviation units compared with the placebo.

Eta-square is useful when it is desirable to define the effect as the proportion of the total variation that is associated with the grouping variable:

$$\eta^2 = \frac{\sigma^2_G}{\sigma^2_G + \sigma^2_{\text{ID}}},$$

where σ^2_G is the variation associated with the grouping variable and σ^2_{ID} the unexplained variation due to individual differences.

A sample estimate of η^2 is provided using sample estimates of the parameters:

$$\hat{\eta}^2 = \frac{SS_G}{SS_G + SS_{\text{ID}}},$$

where SS_G is the sum of squares for the grouping variable and SS_{ID} the sum of squares for individual differences.

Individual differences include unexplained variation and variation associated with the covariate, that is, $SS_{\text{ID}} = SS_{\text{X}} + SS_{Y|GX}$. Both the results of the statistical test for population mean differences and effect size should be reported when summarizing the results of the ANCOVA model. For the current data,

$$\hat{\eta}^2 = 0.156\left(= \frac{261.333}{261.333 + 1079.237 + 339.096}\right).$$

Contrast Analysis

If more than two populations are compared simultaneously (e.g., herbal supplement vs. yoga vs. placebo) the omnibus hypothesis test H_0: $\alpha_j = 0$ for

all j does not identify which populations differ. To identify specific differences between and among populations, contrasts must be examined and tested. A contrast is a linear composite of means: $\psi = \Sigma c_j \mu_j$, with $\Sigma c_j = 0$, where c_j is the contrast coefficient for population j. The hypothesis tested is H_0: $\psi = 0$, (e.g., $H_{0(1)}: \psi = \mu_1 - \mu_2 = 0$, or $H_{0(2)}: \psi = .5\mu_1 + .5\mu_2 - \mu_3, = 0$). A sample estimate, $\hat{\psi}$, is provided using sample estimates of the parameters, for example, $\hat{\psi} = {}_{\text{adj}}Y_1 - {}_{\text{adj}}Y_2$. To test the hypothesis, a t test statistic is formed by taking the ratio of the sample estimate of the contrast to the standard error of the contrast,

$$t = \frac{\hat{\psi}}{S_{\hat{\psi}}}.$$

Because multiple contrasts are generally tested in a single study, several strategies have been suggested for evaluating the t statistic depending on what is judged to be an acceptable risk of a Type I error and statistical power.

Data Assumptions

The statistical validity of the hypotheses tested using the ANCOVA model depends on whether several assumptions regarding the units in the populations being compared are met. In addition to the assumption that the separate regression slopes of the outcome on the covariate are the same for all populations, which was discussed earlier, the ANCOVA model also assumes that the relationship between covariate and the outcome is linear and that the model errors, ε_{ij}, are (a) independent of each other, (b) normally distributed at each level of the covariate, and (c) have equal variance at each level of the covariate both within each population and between the populations being compared.

The assumption of linearity can be examined by testing within each group the statistical significance of the Pearson correlation between the covariate and the outcome. For the current data, the separate correlations between BMI and postsystolic blood pressure are .827 and .906 for the herbal and placebo groups, respectively. Both correlations are statistically significant at the .05 level. Further examining a scatter plot of the data shows a consistent increase in postsystolic blood pressure with increasing BMI scores for each group. A linear relationship is reasonable to assume.

Model errors refer to the difference between actual postsystolic pressure and predicted postsystolic blood pressure from BMI, $Y_{ig} - \hat{Y}_{ig}$. These errors are sometimes referred to as residuals. The independence assumption implies that individuals do not influence each other with respect to the outcome under investigation. Determination of whether this assumption is tenable is best judged based on how data were collected. If there is little interaction among the units between and within each group, the assumption is likely met.

The assumptions regarding the distributions of model errors (normality and equal variance) are best examined by plotting the errors (residuals) for each group around the separate regression lines using the common slope. The homogeneity of error variance assumption can also be examined by comparing mean square error estimates, $S^2_{Y|X_j}$ from the regression lines in each group. For the current study, $S^2_{Y|X_1} = 43.6$ and $S^2_{Y|X_2} = 38.8$, so the variance of errors between the groups appear similar.

The ANCOVA model is generally robust to moderate violations of these data assumptions, particularly when the number of units per group is equal.

Stephen Olejnik and H. J. Keselman

See also Analysis of Variance (ANOVA); Hypothesis Testing

Further Readings

Harwell, M. (2003). Summarizing Monte Carlo results in methodological research: The single-factor fixed-effect ANCOVA case. *Journal of Educational and Behavioral Statistics, 28,* 45–70.

Huitema, B. E. (1980). *The analysis of covariance and alternatives*. New York: Wiley.

Kirk, R. E. (1995). *Experimental design procedures for the behavioral sciences* (3rd ed.). Pacific Grove, CA: Brooks/Cole.

Olejnik, S., & Algina, J. (2000). Measures of effect size for comparative studies: Applications, interpretations, and limitations. *Contemporary Educational Psychology, 25,* 242–286.

Olejnik, S., & Algina, J. (2003). Generalized eta and omega squared statistics: Measures of effect size for some common research designs. *Psychological Methods, 8,* 434–447.

Pedhazur, E. J. (1997). *Multiple regression in behavioral research: Explanation and prediction* (3rd ed.). Fort Worth, TX: Harcourt Brace.

Pedhazur, E.J., & Schmelkin, L. P. (1991). *Measurement, design, and analysis: An integrated approach*. Hillsdale, NJ: Lawrence Erlbaum.

Porter, A. C., & Raudenbush, S. W. (1987). Analysis of covariance: Its model and use in psychological research. *Journal of Counseling Psychology, 34,* 383–392.

Shaffer, J. P. (1995). Multiple hypothesis testing. *Annual Review of Psychology, 46,* 561–584.

Toothaker, L. E. (1991). *Multiple comparisons for researchers*. Newbury Park, CA: Sage.

Analysis of Variance (ANOVA)

Consider a study in which a randomized trial is undertaken to compare a control group, an intervention group receiving a standard treatment, and an intervention group receiving a new treatment on a single continuous outcome measure, such as health status. How can it be determined whether there is a statistically significant difference in the mean outcome score among the three groups? The conventional method of analysis for these data is analysis of variance (ANOVA). ANOVA encompasses a broad collection of statistical procedures used to partition variation in a data set into components due to one or more categorical explanatory variables (i.e., factors). The topics covered in this entry are (a) a description of the applications of ANOVA in medical research, (b) a review of the computations for the ANOVA test statistic, and (c) criteria to assess the reporting of ANOVA results in medical literature.

Applications

Data arising from many different types of studies can be analyzed using ANOVA, including the following:

One-way independent groups design, in which two or more groups of study participants are to be compared on a single outcome measure. This is the simplest type of design in which ANOVA is applied.

One-sample repeated measures design, in which a single group of study participants is observed on two or more measurement occasions. The measurements for each participant are typically correlated (i.e., related).

Factorial independent groups design, in which two or more factors are crossed so that each combination of categories, or cell of the design, comprises an independent group of study participants. Interaction and main effects will usually be tested in factorial designs. A statistically significant two-way interaction implies that the effect of one factor is not constant at each level of the second factor.

Mixed designs, which contain both independent groups and repeated measures factors. Within-subjects interaction and main effects, as well as the between-subjects main effect, may be tested in a mixed design. A significant within-subjects two-way interaction effect indicates that the repeated measures effect is not constant across groups of study participants.

Computing an ANOVA Test Statistic

The sidebar outlines the goal of ANOVA in a one-way independent groups design, the required computations, and the decision rule for the test statistic. The method is described for the simplest situation, in which all the group sizes are equal. A numeric example is also provided.

In an independent groups design, the assumptions that underlie validity of inference for the ANOVA *F* test are as follows:

1. The outcome variable follows a normal distribution in each population from which data are sampled.
2. Variances are equal (i.e., homogeneous) across the populations.
3. The observations that comprise each sample are independent (i.e., unrelated).

In one-sample repeated measures designs or mixed designs, measurements taken from the same study participant are correlated, but measurements from different study participants are assumed to be unrelated. In these designs, the data are assumed to follow a multivariate normal distribution and conform to the assumption of multisample sphericity. Multivariate normality means that the marginal

Computing the ANOVA *F* Test When Group Sizes Are Equal

Goal of ANOVA: To test the plausibility of the null hypothesis, H_0: $\mu_1 = \mu_2 = \ldots = \mu_J$, the hypothesis of equal population means for J groups, against the alternative hypothesis, H_A, at least one of the means is different from the others.

Computations: Compute the sample means, $\bar{y}_1, \bar{y}_2, \ldots, \bar{y}_J$; the grand (i.e., overall) mean, $\bar{y}$; and the sample variances $s_1^2, s_2^2, \ldots, s_J^2$. When the same number of study participants are in each group, n, the total number of study participants is $n \times J = N$. The numerator of the test statistic, the variability between groups, is

$$\text{MSBG} = \frac{n}{J-1}\left[(y_1 - y)^2 + (y_2 - y)^2 + \ldots + (y_J - y)^2\right].$$

The farther apart the means of the groups, the larger this quantity will be.

The denominator of the test statistic, the variability within groups, is

$$\text{MSWG} = \frac{1}{J}\left[s_1^2 + s_2^2 + \ldots + s_J^2\right].$$

This quantity will be larger when there is more variability within groups.

Test statistic:

$$F = \frac{\text{MSBG}}{\text{MSWG}}.$$

Decision rule: Reject H_0 if F exceeds a critical value from an F distribution with numerator degrees of freedom $df_1 = J - 1$ and denominator degrees of freedom $df_2 = N - J$ for a prespecified level of significance (e.g., $\alpha = .05$). The F statistic will be large when H_0 is not true.

Example: Suppose that a researcher collects data for three groups of study participants, with 10 participants in each group. Let the group means be $\bar{y}_1 = 12.0$, $\bar{y}_2 = 8.0$, and $\bar{y}_3 = 11.5$. Then the grand mean, $\bar{y} = 10.5$. Let the group variances be $s_1^2 = 15.0$, $s_2^2 = 10.5$, and $s_3^2 = 18.0$. Then the numerator of the test statistic is MSBW = 47.5 and the denominator is MSWG = 14.5. The test statistic, $F = 3.28$, is compared with a critical value from the F distribution with $df_1 = 2$ and $df_2 = 27$, which is equal to $F_{crit} = 2.96$ when $\alpha = .05$. The p value is .0362. The null hypothesis, H_0: $\mu_1 = \mu_2 = \mu_3$, is rejected.

distribution for each measurement occasion, that is, the distribution of scores for each measurement occasion, ignoring all other occasions, is normal and the joint distribution of the measurement occasions (i.e., the distribution of all occasions together) is normal. Multisample sphericity means that the difference scores for all pairs of repeated measurements have a common variance and also that this common variance is the same for all groups of study participants.

The F test is not robust to assumption violations; this means that it is sensitive to changes in those factors that are extraneous to the hypothesis being tested. In fact, the F test may become seriously biased when assumptions are not satisfied, resulting in spurious decisions about the null hypothesis.

The assumptions that underlie the ANOVA F test are unlikely to be satisfied in many studies. Outliers or extreme observations are often a significant concern and can result in a substantial loss of statistical power to detect study effects. Furthermore, study participants who are exposed to a particular healthcare treatment or intervention may exhibit greater (or lesser) variability on the outcome measure than study participants who are not exposed to it. Inequality of variances can have serious consequences for control of the Type I error rate, the probability of erroneously rejecting a true null hypothesis.

Researchers who rely on ANOVA to test hypotheses about equality of means may, therefore, unwittingly fill the literature with nonreplicable results or at other times may fail to detect effects when they are present. This is of concern because the results of statistical tests are routinely used to make decisions about the effectiveness of clinical interventions and to plan healthcare delivery. In this era of evidence-informed decision making, it is crucial that the statistical procedures applied to a set of data will produce valid results.

Researchers often regard nonparametric procedures based on rank scores, such as the Kruskal-Wallis test or Friedman's test, as appealing alternatives to the ANOVA F test when the assumption of normality is suspect. However, nonparametric

procedures test hypotheses about equality of distributions rather than equality of means. They are therefore sensitive to heterogeneous variances; distributions with unequal variances will necessarily result in rejection of the null hypothesis. Rank-transform test procedures are also appealing because they can be implemented using existing statistical software packages. A rank-transform ANOVA *F* test is obtained by converting the original scores to ranks prior to computing the conventional *F* statistic. One limitation of rank-transform procedures is that they cannot be applied to tests of interaction effects in factorial designs. The ranks are not a linear function of the original observations; therefore, ranking the data may introduce additional effects into the statistical model. Furthermore, ranking may alter the pattern of the correlations among the measurement occasions in repeated measurement designs. Rank-transform tests, while insensitive to departures from normality, must therefore be used with caution.

Transformations of the data, to stabilize the variance or reduce the influence of extreme observations, are another popular choice. Logarithmic, square root, and reciprocal transformations are common. The primary problem with applying a transformation to one's data is that it may become difficult to interpret the null hypothesis when the data are no longer in the original scale of measurement. Also, a transformation may not accomplish the goal of getting rid of outliers.

When variance equality cannot be assumed, robust procedures such as the Welch test for the one-way independent groups design are recommended alternatives to the ANOVA *F* test. Welch's test does not pool the group variances in the computation of the test statistic denominator and modifies the degrees of freedom with a function of the sample sizes and the variances. Welch's test does, however, assume that the data are normally distributed. If normality is not tenable, then a modification of the Welch test should be considered. One alternative involves substituting robust means and variances for the usual means and variances in the computation of the test statistic. Robust means and variances are less affected by the presence of outlying scores or skewed distributions than the usual mean. There are a number of robust statistics that have been proposed in the literature; among these, the trimmed mean has received substantial attention because of its good theoretical properties, ease of computation, and ease of interpretation. The trimmed mean is obtained by removing, or censoring, the most extreme scores in the distribution, which have the tendency to shift the mean in their direction. Current recommendations are to remove between 10% and 20% of the observations in *each* tail of the distribution. A consistent robust estimator of variability for the trimmed mean is the Winsorized variance, which is computed by replacing the most extreme scores in the distribution with the next most extreme observations. While robust measures are insensitive to nonnormality, they test a null hypothesis different from traditional estimators. The null hypothesis is about equality of trimmed population means. In other words, one is testing a hypothesis that focuses on the majority (i.e., central part) of the population rather than the entire population.

Finally, computationally intensive methods, such as the bootstrap method, have also been used to develop alternatives to the ANOVA *F* test. The bootstrap method can be described as follows: The usual ANOVA *F* test is computed on the original observations, but statistical significance is assessed using a critical value from the empirical distribution of the test statistic rather than a critical value from the *F* distribution. The empirical distribution is obtained by generating a large number (e.g., 1,000) of data sets; each data set is a random sample (sampling with replacement) from the original observations. Sampling with replacement means that any observation can potentially be sampled multiple times. The *F* test is computed for each bootstrap data set. The bootstrapped test statistics are ranked in ascending order; the critical value for assessing statistical significance corresponds to a preselected percentile of the empirical distribution, such as the 95th percentile. Bootstrap test procedures have good properties in the presence of assumption violations. For example, the bootstrapped ANOVA *F* test for repeated measures designs will control the rate of Type I errors to α, the nominal level of significance, under departures from both normality and sphericity.

Assessing ANOVA Results

For decision makers to have confidence in ANOVA results reported in the medical literature, it is

important that the choice of test procedures is justified and the analytic strategy is accurately and completely described. The reader should be provided with a clear picture of the characteristics of the data under investigation. This can be accomplished by reporting exploratory descriptive analysis results, including standard deviations or variances, sample sizes, skewness (a measure of symmetry of the distribution) and kurtosis (a measure of peakedness of the distribution), and normal probability plots. As a general rule of thumb, skewness and kurtosis measures should be within the range from +1 to −1 to assume that the data follow a normal distribution. The normal probability plot is a graphic technique in which the observations are plotted against a theoretical normal distribution; if all the points fall on an approximate diagonal line, then normality is likely to be a tenable assumption.

While preliminary tests of variance equality, such as Levene's test, or tests of sphericity, such as Mauchly's test, are available in statistical software packages, their use is not recommended in practice. Many tests about variances are sensitive to departures from a normal distribution, and those that are insensitive to nonnormality may lack statistical power to detect departures from the null hypothesis of equal variances, which can result in erroneous decisions about the choice of follow-up tests.

For factorial designs, unless there is theoretical evidence that clearly supports the testing of main effects only, the analysis should begin with tests of interactions among the study factors. Graphic presentations of the cell means are often useful to characterize the nature of the interaction.

Each test of a main or interaction effect should be completely described. This includes reporting the numeric value of the test statistic, degrees of freedom, and p value or critical value.

A statistically significant ANOVA F test is routinely followed by multiple comparisons to identify the localized source of an effect. The choice of a multiple comparison test statistic and procedure for controlling the familywise error rate, the probability of making at least one Type I error for the entire set of comparisons, should be explicitly identified in the reporting of results. A simple Bonferroni approach may suffice, in which each of m comparisons is tested at the α/m level of significance. However, this multiple comparison procedure is often less powerful than modified Bonferroni procedures, such as Hochberg's procedure.

Conclusion

ANOVA is one of the most popular test procedures for analyzing medical data because it can be used in a wide variety of research applications. Researchers may be reluctant to bypass the conventional ANOVA F test in favor of an alternative approach. This reluctance may stem, in part, from the belief that the F test is robust to departures from derivational assumptions. While Type I error rates may be relatively robust to the presence of nonnormal distributions, power rates can be substantially affected. This is a critical issue, particularly for small-sample designs, which are common in clinical trials. Departures from variance homogeneity and sphericity can result in seriously biased tests of between-subjects and within-subjects effects, respectively. Statistical procedures that are robust to assumption violations have been developed for both simple and complex factorial designs and are now routinely available in many statistical software packages.

Lisa M. Lix and H. J. Keselman

See also Analysis of Covariance (ANCOVA); Measures of Central Tendency; Multivariate Analysis of Variance (MANOVA); Variance and Covariance

Further Readings

Conover, W. J., & Iman, R. L. (1981). Rank transformation as a bridge between parametric and nonparametric statistics. *The American Statistician, 35,* 124–129.

Hill, M. A., & Dixon, W. J. (1982). Robustness in real life: A study of clinical laboratory data. *Biometrics, 38,* 377–396.

Hochberg, Y. (1988). A sharper Bonferroni procedure for multiple tests of significance. *Biometrika, 75,* 800–802.

Keselman, H. J. (2005). Multivariate normality tests. In B. S. Everitt & D. C. Howell (Eds.), *Encyclopedia of statistics in behavioural science* (Vol. 3, pp. 1373–1379). Chichester, UK: Wiley.

Keselman, H. J., Wilcox, R. R., & Lix, L. M. (2003). A generally robust approach to hypothesis testing in

independent and correlated groups designs. *Psychophysiology, 40,* 586–596.
Lix, L. M., Keselman, J. C., & Keselman, H. J. (1996). Consequences of assumption violations revisited: A quantitative review of alternatives to the one-way analysis of variance *F* test. *Review of Educational Research, 66,* 579–619.
Scariano, S. M., & Davenport, J. M. (1987). The effects of violations of independence assumptions in the one-way ANOVA. *The American Statistician, 41,* 123–129.
Toothaker, L. E. (1991). *Multiple comparisons for researchers*. Newbury Park, CA: Sage.
Vickers, A. J. (2005). Parametric versus non-parametric statistics in the analysis of randomized trials with non-normally distributed data. *BMC Medical Research Methodology, 5,* 35.
Wasserman, S., & Bockenholt, U. (1989). Bootstrapping: Applications to psychophysiology. *Psychophysiology, 26,* 208–221.
Wilcox, R. R. (1995). ANOVA: A paradigm for low power and misleading measures of effect size? *Review of Educational Research, 65,* 51–77.
Wilcox, R. R., & Keselman, H. J. (2003). Modern robust data analysis methods: Measures of central tendency. *Psychological Methods, 8,* 254–274.
Zimmerman, D. W. (2004). A note on preliminary tests of equality of variances. *British Journal of Mathematical and Statistical Psychology, 57,* 173–181.

Applied Decision Analysis

Decision analysis (DA) is a methodology by which the various aspects of a decision are represented in an explicit and quantitative model to support or improve the procedure and/or outcome of decisions under uncertainty. The term *decision analysis* is used both for the domain and for the actual single exercise of construction and quantifying a model for a particular problem. DA can be used purely for the sake of knowledge itself (such as to increase one's own understanding or that of others in a teaching setting) and also with the purpose to apply that knowledge to real-life medical dilemma, where a choice has to be made. This is called *applied decision analysis*, although the more correct term might be *applicable decision analysis*, as, however one may be with the model, it remains to be seen whether it will convince doctors and patients sufficiently to be used in clinical practice. In this entry, an overview is given of applied DA, its history, the why and how, and its present and future potential and limitations.

History

The history of applied DA starts with Stephen Pauker's famous "Clinical Decision Making Rounds at the New England Medical Center," which started in 1981 and filled the first 10 issues of the journal *Medical Decision Making*. Typically, these papers dealt with problems in individual patients and elaborated from that individual to more general issues. Since that pioneering time, the number of publications on clinically applied DA has increased strongly over time and keeps doing so. Papers on applied DA increase more than twice as fast as those on other clinical issues (with a doubling time of 4.2 years as compared with 9.9 years).

The "Why" of Applied Decision Analysis

Why one should perform a DA for a real-life clinical problem may not be clear to everyone in the first place. Normally, medical choices in healthcare are dealt with in a more or less implicit way, where doctors or other healthcare professionals rely on their knowledge or experience to estimate which choice alternative would (probably) provide the best outcome. Dissatisfaction with the subjectivity and the lack of transparency of this process and its outcomes has over several decennia led to the development of more explicit and quantitative methods of dealing with clinical issues that include not only DA but also evidence-based medicine (EBM). DA differs in its ambition from EBM.

EBM builds on the assumption that doctors only have an information problem that can be solved by providing them with the right data. EBM, and in particular the Cochrane collaboration, has therefore put an enormous effort in tracking that information and making the data available to doctors.

As has been argued by Arthur Elstein and others, DA goes one step further and assumes that in addition to the information problem, doctors also have a judgment problem. That judgment problem has to do with the complexity of integrating all the available information elements and their relations, and condensing it into the right choice, and with

the fact that making choices entails taking into account not only probabilities but likewise valuing potential outcomes. That may be a reason why doctors may have a better feel about EBM (here is the info, so you can decide) than about DA (here is the info and the advice, because you cannot be trusted to make the right judgment). The added value that DA brings is in structuring and summarizing available knowledge and in supporting or steering actual decisions.

There are several advantages of using applied DA to tackle clinical problems. First, a global problem is dissected into parts, so that the intricacies and complexities of the decision problem are made clearer. Second, using data synthesis methods (an abundance of), available data will be combined and restructured into a limited number of variables that are essential to the (solution of the) problem, so that available knowledge is summarized in a clear and concise way. Third, far from being mechanistic or generalizing, the DA approach allows for individualization of choices if variables are used that characterize individuals and their (relevant) characteristics. Fourth, the root causes of clinical disagreement (if present) will become more clear by the process of dissecting the problem, building the model, and combining it with the available information. This will pinpoint why clinicians (if they do) differ in opinion and will allow for the testing of the arguments of each camp against available evidence. Finally, and perhaps most important for clinical purposes, the (ir)relevance of various elements of the problem may be tested by using sensitivity and threshold analysis. Thus, the understanding of the intricacies of the clinical issues may improve considerably, thanks to quantitative answers on "what if" questions (albeit about different patients, settings, and/or any key variables).

Self-evidently, there are also potential disadvantages to using DA to solve clinical problems. Real-life problems are inevitably simplified when they are translated into DA models, thus providing solutions that fit the model but do not necessarily solve the real-life problem. This means that other issues (such as feasibility, experience and expertise, safety, and acceptability), that go beyond the purely medical information represented in the model, but may be highly relevant, may not be taken into account in the model's advice. Then, the information necessary to quantify all the variables in the model may be deficient, and the pragmatic use of low-quality data entails the risk of "garbage in, garbage out." One should not forget that building a decision model, and in particular collecting and summarizing available evidence into the necessary variables, requires decision analytic expertise and experience, and it may take a considerable amount of time. Finally, however careful a decision model is constructed, there is always the risk that errors within the model may go unobserved, which will jeopardize the validity of the answers supplied (example).

Thus, before one embarks on doing a DA, one should take heed. Building a decision tree and "playing with it" to better understand the various issues may be relatively easy. But upscaling it to something that is sufficiently convincing for incorporation into guidelines, or for publication in a peer-reviewed journal, is an endeavor of a different scale.

The "How" of Applied Decision Analysis

The technical execution of an applied DA can be subdivided into several stages or aspects:

1. Identifying the precise nature of the real-life medical problem, including
   a. the type of question (diagnosis, therapy, diagnostic-therapeutic management), the patient category, the setting (primary care, center of excellence);
   b. the relevant outcomes such as mortality, (quality-adjusted) life expectancy, disease-free survival, cost, and so on;
   c. the available policy alternatives, both the realistic ones, and the extremes of doing nothing and treating everyone always; and
   d. the various arguments in favor of, or against, each policy alternative.
2. Structuring the problem first in a flowchart or an algorithm (using "if-then-else" sequences only, and no variables yet), and only then in a full-fledged decision tree.
3. Obtaining the necessary data on underlying diseases and their natural histories on prior probabilities and relations with determinants in subgroups, on characteristics of available diagnostic tests and of relevant therapies, and on (the value of) relevant outcomes.

4. Performing calculations on expected outcomes and costs, and including sensitivity analysis (what matters?), threshold analysis (when should one choose differently?), and resulting in overall conclusions.

For the first stage, close cooperation with experts in the (clinical) field is of the utmost importance to make sure that the decision analyst understands why there is a problem and what the real-life options and outcomes may be. It is particularly important to identify the relevant stakeholders and to know their optimization criteria (what do they consider important, either as something to achieve or as something to avoid). To convince the stakeholders, and in particular the decision makers, one should know which issues and aspects of the problem and its potential solutions they hold important. Stage 2 requires adequate logical understanding and sufficient technical expertise to create a model in whatever technical context, while Stage 3 requires considerable experience with literature searching and data synthesis. Before calculations are performed and any conclusions may be drawn from them, all aspects of the model should be checked and double checked, ideally by two or more experts in the field. Most experienced decision analysts have had personal experience with the impact that mistakes in structure and formulas may have on model outcomes. In general, errors are weeded out before publication by rigid testing of the model through sensitivity analyses and other testing procedures and by having others check and double check the model. However, discussions about the correctness of model assumptions and details, in relation to the intricacies of the clinical problem, may be heated, and may continue even after publication.

Suitability for Real-Life Problems

The choice to perform an applied DA should not be taken lightly. One should have a clear idea of what a DA may add to the usual clinical (more implicit and less quantitative) way of dealing with a problem, used by experienced clinicians, and whether DA is suitable for the problem at hand. Not all clinical problems are ideally suitable for this approach, and the overall balance of advantages and disadvantages may strongly differ for different clinical problems. In general, DA can be used more advantageously for clinical problems

- that concern risk or uncertainty,
- that are structurally complex,
- about which sufficient quantitative data are available,
- in which solutions differ for different patients or patient categories,
- in which different and conflicting interests have to be considered and weighed, and
- of which the frequency of occurrence or the magnitude of the problem justifies the effort of performing a DA.

Impact on Clinical Performance

The impact of applied DA, in particular to what extent papers on applied DA are actually changing medical practice, is not easy to assess. Likewise, it has long been unclear which factors contribute to success or failure in real-life clinical practice. In medication prescription, A. Holbrook found that factors such as system speed, convenience of use, quality, relevance to the task at hand, and integration with workflow are important determinants of success. Recent reviews on the effectiveness of clinical decision support systems in improving clinical performance have increased our insight in these matters. However, many of these decision support systems that are assessed in these systematic reviews differ from "classical decision analyses" with their decision trees and tables of variables. Four factors have been found to independently predict success: (1) automatic provision of decision support as part of clinician workflow, (2) provision of recommendations rather than just assessments, (3) provision of decision support at the time and location of decision making, and (4) computer-based decision support. These findings confirm what decision analysts have experienced over many years—that the results of formal decision tree calculations should be transformed into clinically more acceptable formats. In addition, automatic prompting to use the system is a success factor as it reduces the burden and threshold of use. Not surprisingly, most studies where the authors were the creators of the system are more positive about a system's ability to improve clinical performance.

Other Issues

Real-life clinical problems are not solved by the completion of a DA alone. In practice, actions have to be taken by the decision maker. In this process, several other issues may come into play, which have been described by decision psychologist Frank Yates. For any decision, he identifies 10 cardinal checks. They range (apart from the first question on whether there is the need to decide at all) from the "who and how" of the clinical decision, to more practical issues such as acceptability and implementation. All these checks emphasize the fact that there is a real-life world out there, beyond the model, and that for decisions to be successful, one should look beyond the model to real-life situations and to both their potential and their limitations.

Health Technology Assessment

Health technology assessment and cost-effectiveness analysis are next-generation family members of DA. By the fact that they take not only medical outcomes into account, but likewise costs and equity issues, they are becoming more and more relevant to healthcare systems that suffer from the strain that expanding medical technologies and increasing public demands put on the limited available healthcare resources in many countries. One of the most striking examples of the use of such methods at the macrolevel is the use of health technology assessment to steer policy making in the U.K. National Health Service by National Institute for Clinical Excellence (NICE). Whether new medications or other interventions are allowed and are paid for by the National Health Service is based on a rigid analysis of both the avail able evidence on their effectiveness and of the cost burden they would put on the National Health Service (and thereby on the British taxpayer). NICE's approach has set quite an example of methodological rigor and is a success story of practical potential of applied DA and health technology assessment methods. It is therefore all the more striking that NICE is quite regularly depicted as bureaucratically denying patients the access to "wonderful new drugs" on the basis of cost containment only. This contrast is an illustration of the fact that however right one may be from an intellectual point of view, the value of being able to explain clearly and simply through the right channels to all stakeholders (and maybe of using the right communication/PR methods while doing so) cannot be overestimated, and it confirms the relevance of Yates's warnings.

Success and Effectiveness

Applied DA and its descendants such as health technology assessment and clinical decision support systems have come a long way since their start in the early 1980s. Research as well as experience suggests that the success of applying DA methodologies to real-life problems depends on many factors, not least of all an intense exchange of ideas between analysts and potential users right from the start and the continuing realization that there is a reality beyond the model and its calculations and that feasibility, acceptability, and other implementation issues will codetermine the effectiveness of applied DA.

J. Kievit

See also Cost-Effectiveness Analysis; Evidence-Based Medicine

Further Readings

Beck, J. R., Plante, D. A., & Pauker, S. G. (1981). A 65-year-old Chinese woman with lymphadenopathy and progressive pulmonary infiltrates. One million Chinese women can't all have tuberculosis. *Medical Decision Making, 1,* 391–414.

Claxton, K., Ginnelly, L., Sculpher, M., Philips, Z., & Palmer, S. (2004). A pilot study on the use of decision theory and value of information analysis as part of the NHS Health Technology Assessment programme. *Health Technology Assessment, 8,* 1–103, iii.

Computerization of medical practice for the enhancement of therapeutic effectiveness. COMPETE Publications. Retrieved January 16, 2009, from http://www.compete-study.com/publications.htm

Elstein, A. S. (2004). On the origins and development of evidence-based medicine and medical decision making. *Inflammation Research, 53*(Suppl. 2), S184–S189.

Garg, A. X., Adhikari, N. K., McDonald, H., Rosas-Arellano, M. P., Devereaux, P. J., Beyene, J., et al. (2005). Effects of computerized clinical decision support systems on practitioner performance and

patient outcomes: A systematic review. *Journal of the American Medical Association, 293,* 1223–1238.
Hanney, S., Buxton, M., Green, C., Coulson, D., & Raftery, J. (2007). An assessment of the impact of the NHS Health Technology Assessment Programme. *Health Technology Assessment, 11,* iii–xi, 1.
Kawamoto, K., Houlihan, C. A., Balas, E. A., & Lobach, D. F. (2005). Improving clinical practice using clinical decision support systems: A systematic review of trials to identify features critical to success. *British Medical Journal, 330, 765.*
National Institute for Clinical Excellence (NICE). http://www.nice.org.uk
Yates, J. F. (2003). *Decision management.* San Francisco: Jossey-Bass.

Artificial Neural Networks

Research on artificial neural network modeling started in the early 1940s when the first scientific paper by Warren McCulloch and Walter Pitts was published. The motivation came from the fields of artificial intelligence and neuroscience when initial investigators attempted to model the workings of neurons in the human brain. One of the hypothesized reasons for the brain's superiority compared with common computers lies in the fact that neurons function in parallel. There are approximately 10^{12} neurons in the human brain, all interconnected and receiving input from many other neurons, as well as stimulating many others in a conglomeration of complex interconnections. Thus, neural networks are able to perform highly complex computing tasks in an efficient and powerful manner. In addition, they are able to integrate newly acquired data, or experiences, into existing ones, thus allowing for efficient learning and inference. Figure 1 illustrates a basic representation of the neuron and how it gets activated to fire (stimulate) other connected neurons.

As shown in Figure 1, the neuron collects and processes input from structures referred to as dendrites. It then sends out electrical activity through a long strand called an axon. This axon splits into multiple branches, and at the end of a branch, a synapse converts the electrical activity from the axon and sends stimuli to the neighboring neuron. This activity is either excitatory or inhibitory. Prior information affects signal transfer functions and influences how neurons respond to any future stimuli; synaptic processing mimics learning in this sense. Information transmission and processing across multiple neurons influence the development of artificial neural network models.

The simplest representation of a single-layer artificial neural network is shown in Figure 2. Similar to neuronal processing, information is passed between nodes (neurons) interconnected by links (synapses) with modifiable weights. In the case of a single-layer neural network, input into the node is often represented as a vector of features $X = (x_1, x_2, \ldots, x_n)$. (A single-layer network is also referred to as a two-layer network corresponding to the number of layers of input and output units. Often, it is referred to as a single-layer because there is only one layer of modifiable weights.) Each of these feature values, x_i, is multiplied by a corresponding weight, w_i. Thus, the effective input at the output unit would be the sum of all the products $\sum w_i x_i$. Adding a constant bias term ($x_0 = 1$) with a corresponding weight w_0 produces the formula for a single-unit perceptron.

$$\mathrm{f(x, w)} = \mathrm{w0} + \sum_{\mathrm{i}=1}^{\mathrm{n}} \mathrm{xiwi}. \tag{1}$$

This could then be represented as

$$\mathrm{f(x, w)} = \sum_{\mathrm{i}=0}^{\mathrm{n}} \mathrm{xiwi}. \tag{2}$$

A clinical scenario where an artificial neural network would be useful would be in predicting mortality after a procedure (e.g., angioplasty) in patients with chronic renal failure. The input might comprise several clinical features, such as age, gender, hypertension, diabetes, heart failure, and coronary artery involvement. The output would be mortality after 6 months, which is binary in this example although not necessarily so for artificial neural networks. The artificial neural network is useful for achieving increased accuracy of prediction when the features might have nonlinear interactions.

The input into the node is then processed to generate an optimal output. This is determined by a function $y = g(f(x, w))$. Typically, $f(x, w)$ is linear as in Equation 1. The function g, on the other

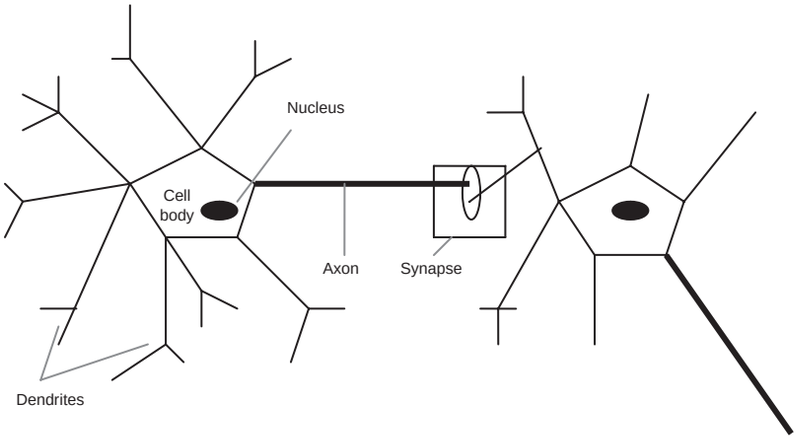


Figure 1 Structure of a typical neuron and a synaptic junction

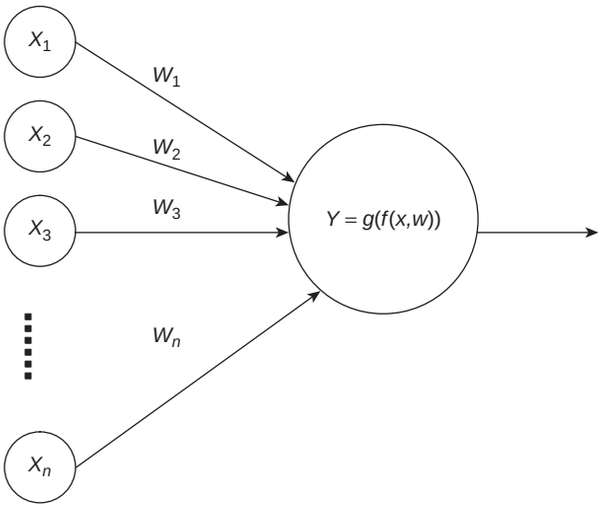


Figure 2 Single-layer artificial neural network

hand, is commonly referred to as the activation function. It is chosen from a selection of functions, including the following:

$$g(x) = x, \text{ a linear function.} \tag{3}$$

$$g(x) = x_+, \text{ producing a nonnegative value.} \tag{4}$$

$$g(x) = \tanh(x), \text{ producing output between } -1 \text{ and } 1. \tag{5}$$

$$g(x) = \sin(x), \text{ where the output is 1 if } x \geq 0 \text{ and } -1 \text{ if } x < 0. \tag{6}$$

$$g(x) = [\sin(x) + 1]/2, \text{ where the output is 0 or 1.} \tag{7}$$

$$g(x) = [1 + e^{-x}]^{-1}, \text{ with a sigmoidal output between 0 and 1.} \tag{8}$$

The specifications for an artificial neural network are determined by two mechanisms, the architecture of the network and optimization of the network parameters generally based on performance in a given data set.

Architecture

Single Layer

A typical single-layer network is shown in Figure 2. As shown in this simplified example,

artificial neural networks consist of layers with input nodes and corresponding modifiable weights. In a single-layer network, all nodes connect to the output node(s) where the activation function generates an output.

The learning process or network optimization involves recursive modification of weights as more training data get processed. The recursive algorithm is described as follows.

Taking the angioplasty example in the previous section, the network learns by adding examples from the training data set. Suppose a new observation datum is to be added into the model (x_m, z_m), where x_m corresponds to the feature vector for one patient (e.g., 60 years of age, male gender, nonhypertensive, diabetic, with heart failure and left main coronary artery involvement), and z_m corresponds to the actual output (e.g., death after 6 months). The weights for each of the current nodes would be modified as follows:

1. Calculate the error derived from the predicted output for each output unit, compared with the desired (actual) output, z_m. This could be represented as the mean squared error:

$$E(w) = \frac{1}{2}\sum_{r=1}^{N}(z_r - p_r)^2. \tag{9}$$

Following matrix transposition of Equation 2, the predicted output for the training data, p_r, is obtained.

$$p_r = \sum_{\mathrm{i}=0}^{\mathrm{n}} \mathrm{x_i} w_i = x_r^T w_r. \tag{10}$$

2. Given the weights for the nodes in the single layer, $w_i = w_1, w_2, \ldots, w_n$, the weights change according to the following rule: $w_i = w_i + \Delta w_i$, where

$$\delta \mathrm{w_i} = \eta E(w) x_i. \tag{11}$$

η (greater than 0) is the learning rate. To minimize the error, E, using the gradient descent method for optimization, the steps include

$$\delta w_\mathrm{i} = -\eta \frac{\partial E(w)}{\partial w_i}. \tag{12}$$

To substitute E, where p_{mr} is the predicted output for the training data,

$$D = \{(x_r, z_r,), r = 1, \ldots, N\}$$

$$E = \frac{1}{2}\sum_{r=1}^{N}(z_m r - p_m r)^2. \tag{13}$$

Further substituting p_{mr} with Equation 10,

$$E = \frac{1}{2}\sum_{r=1}^{N}(z_m r - x_m r^T w_m r)^2. \tag{14}$$

For some learning rate η (greater than 0), a recursive version of the steepest descent is obtained. Further substituting Equation 14 into Equation 12 results in $\Delta w_i = \eta(z_{mr} - x_{mr}{}^{\mathrm{T}} w_{mr})\ x_{mi}$., which is similar to Equation 11.

The single-unit perceptron convergence theorem states that if two classes in a training set can be separated by a hyperplane in **R**, then the delta rule (Equation 11) converges to result in a single hyperplane in a finite number of steps. This has been further developed by investigators who worked on cases where the classes are not linearly separable and where there are greater than two classes. In 1969, M. Minsky and S. Papert released a research publication that basically stated that single-layer perceptrons were not able to solve simple problems, most notably the exclusive-OR (XOR) problem. This was addressed subsequently using multiple-layer perceptrons.

Multiple Layers

Neural networks typically have more than a single layer. The most common form has two layers, the second corresponding to a hidden layer. The architecture is arbitrarily designed, with the main components including the number of layers and the number of units in each layer. Figure 3 illustrates a two-layer neural network.

Feedforward Operation

A neural network that has more than a single layer typically proceeds forward to process input from one layer to the next. The only limitation is that each layer only sends signals to the next layer after it. In Figure 3, there is an input layer that processes the external stimuli. There is a second layer, also referred to as a hidden layer. The third column of nodes corresponds to the two output nodes. Weights are specified and modified for each

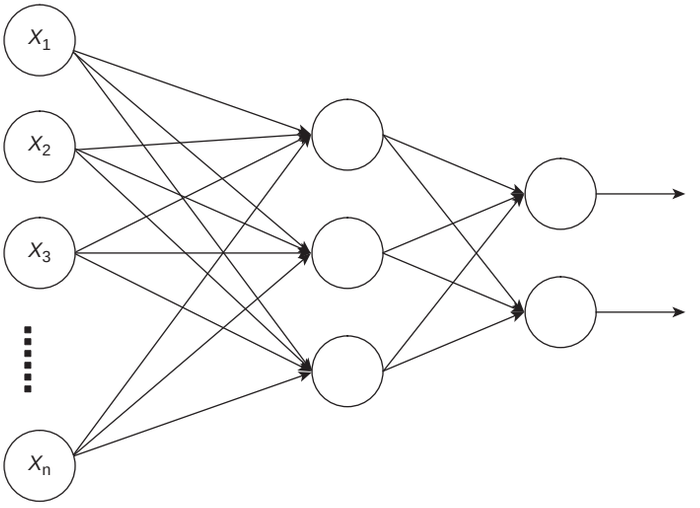


Figure 3 Two-layer artificial neural network with two output nodes

interconnection between nodes. In addition, the hidden layer(s) and output node(s) have activation functions and biases assigned, as described for single-layer artificial neural network. In the multiple-layer artificial neural network, each hidden node computes the weighted sum of all its input from the preceding layers. An activation function, in turn, computes the signal that it then sends to the next group of node(s), which would be another hidden layer or the output node(s). A single hidden layer is able to solve the XOR problem. In fact, A. N. Kolmogorov proved that any continuous function from input to output can be implemented in a two-layer network. However, practical considerations limit the applicability of this theorem. The activation functions would have to be very complex, and there is no principled way that has been suggested to find nonlinear functions based on training data. In addition, some functions are not smooth, which is important for gradient descent learning.

Backpropagation

Backpropagation is one of the most commonly used and simplest methods for training multilayer networks. The simplest way to describe the training method is to follow what happens when a new training datum is added into the network. Suppose the entire training data are represented as follows: $D = \{(x_r, z_r,), r = 1, \ldots, N\}$, where x is the feature vector and z is the actual expected output. When a new datum is added, (x_m, z_m), training of the network ensues. Similar to the method described in the single-layer network, the backpropagation proceeds in the following manner:

1. The output of the network is computed using the feedforward operation for each of the output nodes.
2. The training error between the predicted and actual output is calculated. Typically, it is based on the sum over the output units of the squared difference between the predicted and actual output (Equation 9), which will be referred to as *net*.
3. The weights are initialized with random values and are modified in a direction that will reduce the error, similar to that of Equation 12. The weight update or learning rule is calculated based on the first derivative of $f(net)$, the unit's nonlinear activation function.
4. As was previously done for the hidden-to-output weights, the input-to-hidden weights also get updated.
5. The steps are repeated until the error reaches a specified low threshold.

The description for the two-layer network can readily be generalized into more layers. The activation functions in each node can also vary apart from the bias units and the learning rates.

Special Considerations

Some techniques have been identified to optimize backpropagation and to guide the users in building neural networks. These techniques are briefly described below.

Activation Function

As noted previously, backpropagation should work with any activation function, given that there is continuity of the function and its derivative. However, in selecting an activation function, some guidelines include selecting functions that are nonlinear, that saturate (functions with a minimum and maximum output value), and that have continuity and smoothness. A sigmoid is one such activation function.

Criterion Function

The use of the squared error is described in Equation 10. There are, however, other alternatives that may be used, including cross-entropy error for comparing the separation between probability distributions and Minkowski error for distributions that have long tails.

Number of Hidden Layers

Any number of hidden layers is possible as long as the activation function in each unit is differentiable. However, since the two-layer network can implement any arbitrary function, the addition of an extra layer adds complexity and makes the network more prone to getting caught in local minima. A special condition for using an extra layer includes data transformations, such as rotation or lateral shifts in data.

Number of Hidden Units

The number of hidden units primarily influences the expressivity of the network and how complex the decision boundaries are. Thus, well-separated data will require fewer hidden units. The number of hidden units dictates the number of weights in the network (in addition to the dimensionality of the input vector). Thus, it should not be more than the total number of the training data, n. A rule of thumb is to use $n/10$ hidden units. This can then be adjusted up or down during training.

Initializing Weights

Weights have to be nonzero. The recommended range for the hidden-to-output weights is $-1\sqrt{h}$ to $+1\sqrt{h}$, where h is the number of hidden nodes connected to the output. Similarly, the range for the input-to-hidden weights is $-1\sqrt{d}$ to $+1\sqrt{d}$, where d is the number of input variables connected to the hidden unit.

Learning Rate

The learning rate influences the quality of the network in most instances where training does not reach the training error minimum. In practice, the learning rate is set at .1. It is lowered if the criterion function diverges during learning and is increased if learning is very slow.

Stop Training

Excessive training can lead to poor generalization, also called *overfitting* or *overtraining*. In practice, the goal is to stop training when the error in a separate validation set reaches a minimum.

Applications

Artificial neural network has been used in multiple domains and applications, including image processing, speech recognition, and prediction of financial indices. The use of artificial neural networks in medical decision making ranges from recognition of chromosomal abnormalities, detection of ventricular fibrillation, protein structure prediction, pharmacovigilance applications, and identifying clinical outcomes. Multiple publications review various networks that have been trained and validated in various clinical domains. In addition, many more studies publish the predictive performance of artificial neural network in comparison with other predictive modeling techniques. Artificial neural network has shown comparable performance to several predictive modeling techniques, including logistic regression, decision tree, and support vector machine.

In all, the use of artificial neural network should be tempered with the known constraints of the method. These include the ability to correctly specify the architecture and parameters of the network and, more important, the ability to measure

the contribution of each of the components of the input vector in determining the output of the network. In many clinical domains, the "black box" is not ideally suited for understanding what factors influence specific clinical outcomes. This, in turn, is a deterrent to deciding what interventions to modify clinical factors might need to be recommended for clinical care. On the other hand, artificial neural network has been successfully used for clinical domains when nonlinear interactions need to be modeled in a complex manner. This was illustrated in the successful use of artificial neural network for computerized image analysis of Papanicolaou smears, used for rescreening for cervical abnormalities not previously identified by manual screening. In such clinical settings, more accurately predicting an outcome is of paramount importance in clinical decision making.

Ronilda Lacson and Lucila Ohno-Machado

See also Ordinary Least Squares Regression; Prediction Rules and Modeling

Further Readings

Cheng, B., & Titterington, D. (1994). Neural networks: A review from a statistical perspective. *Statistical Science, 9*(1), 2–54.

Duda, R., Hart, P., & Stork, D. (2001). *Pattern classification* (2nd ed.). New York: Wiley.

Kurkova, V. (1992). Kolmogorov's theorem and multilayer neural networks. *Neural Computation, 5*(3), 501–506.

Lacson, R. C., & Ohno-Machado, L. (2000). Major complications after angioplasty in patients with chronic renal failure: A comparison of predictive models. *Proceedings of the AMIA Symposium,* 457–461.

McCulloch, W. P. W. (1943). A logical calculus of ideas imminent in nervous activity. *Bulletin of Mathematical Biophysics, 5,* 115–133.

Minsky, M., & Papert, S. (1969). *Perceptrons.* Cambridge: MIT Press.

Penny, W., & Frost, D. (1996). Neural networks in clinical medicine. *Medical Decision Making, 16*(4), 386–398.

Ripley, B. D. (1996). *Pattern recognition and neural networks.* Cambridge, UK: Cambridge University Press.

Rosenblatt, F. (1962). *Principles of neurodynamics: Perceptrons and the theory of brain mechanisms.* Washington, DC: Spartan Books.

Ruján, P. (1993). A fast method for calculating the perceptron with maximal stability. *Journal de Physique, 3,* 277–290.

Rumelhart, D., Hinton, G., & Williams, R. (1986). Learning internal representation by back-propagating errors. *Nature, 323,* 533–536.

Williams, R. W., & Herrup, K. (1988). The control of neuron number. *Annual Review of Neuroscience, 11,* 423–453.

Associative Thinking

Associative thinking is used to describe memory-based judgment processes that require the decision maker to infer a diagnosis or other category on the basis of the presence or absence of related features through the activation of associations—memories in which features and categories co-occur. Broadly speaking, the mind automatically associates in memory those experiences or concepts that co-occur. The decision maker later retrieves these associations (again, automatically, and typically unconsciously) in the performance of judgment and decision tasks. For example, when a pediatrician repeatedly sees children who present with persistent sore throat and fever and observes that they are often positive for strep throat, she or he may come to associate the symptoms and the diagnosis, and on the next presentation of a child with sore throat and fever, strep throat is likely to be high on her differential diagnosis. In essence, judgments are evoked by considering the similarity or representativeness of new stimuli to associations previously learned. More frequent and salient co-occurrences result in more memorable associations.

The study of association in thinking has a long history, dating back at least as far as the work of English empiricists in the 17th century. In modern dual-process theories of cognition, associative thinking is often considered to be characteristic of System 1 (intuitive) thinking. It is contrasted with the more effortful and rule-oriented System 2 (deliberative) thinking.

Determinants

According to dual-process theories, associative thinking is automatically performed, but associations may

be suppressed or modified by later deliberation. Associative judgments are more likely to be expressed when deliberation is limited or infeasible. For example, time pressure or cognitive load may increase the likelihood of relying on associative thinking. In other cases, lack of appropriate information or information format may prevent deliberation. For example, Windschitl and Wells showed that eliciting judgments using verbal measures of uncertainty (e.g., "unlikely") evoked associative thinking more frequently than when numerical measures were used.

Associations vary in their strength. Hogarth notes that associations can be reinforced positively or negatively and offers three factors that lead to reinforcement. First, human beings may be genetically predisposed to create particular associations very quickly through operations similar to classical conditioning. Experiences of pain and fear, for example, often rapidly produce or reinforce strong associations with co-occurring events. Second, people can be motivated to increase the strength of an association. Motivation can take the form of either internal motivation to better understand the environment or external motivation (e.g., operant conditioning) from rewards or punishments provided by the environment. For example, associations that lead to decisions that result in approbation are likely to be reinforced. Third, associations are strengthened as the frequency of the association being observed increases. For example, a physician examining a patient within his or her specialty is likely to have developed strong associations between symptoms and diagnoses as a result of the frequency with which the physician examines such patients; a physician examining a patient with a novel diagnosis outside his or her specialty may have fewer and weaker relevant associations.

Advantages and Disadvantages

Because associative thinking allows for rapid categorization and judgment, it can be ecologically adaptive. This is particularly the case when the decision maker has considerable opportunity to develop valid associations and must make decisions in limited time or without other resources necessary to support a more deliberative process. For example, medical decision making in emergent conditions is often greatly facilitated by the ability of the physician to make correct associations rapidly.

On the other hand, when associative knowledge is developed that does not accurately match the actual state of the world, associative thinking can lead to systematic biases in judgment. In addition to such common heuristics for likelihood judgments as availability, representativeness, and value-induced bias, it is also possible to simply make incorrect associations. For example, medical students exposed to dermatological diagnoses and later tested on diagnostic skill have been shown to establish (irrelevant, and therefore incorrect) associations between diagnoses and the body part on which they first learned the diagnosis.

Improving Associative Thinking

Although faulty associative thinking can sometimes be overridden by analytic thinking, associative thinking itself can be improved by developing more veridical and useful associations. This requires either selecting or creating learning environments that provide sufficient exposure to an appropriate set of co-occurring events, information on whether correct associations have been learned (feedback), and suitable rewards for correct associations and adverse consequences for erroneous associations. Hogarth broadly divides learning environments into those that are kind and those that are wicked. Kind environments provide relevant feedback and have exacting consequences for errors; the former allows the learner to adjust associations through observation of their outcomes, and the latter ensures that the learner is well motivated to seek ongoing improvement in, and refinement of, the associations learned. Wicked environments, in contrast, provide either no feedback or distorted feedback, limiting the learner's ability to correct errors, and are lenient in their tolerance of error, reducing motivation to correct errors.

In medical education, learning environments can often be manipulated to provide better control over exposure to co-occurring events. For example, presentation of multiple teaching and practice cases for the differential diagnosis of heart failure in descending order of typicality has been shown to facilitate the development of better associations and improved diagnostic performance in medical students. Similarly, Ericsson has argued that the

development and maintenance of expert-level performance in medicine relies on deliberate practice designed to ensure that the expert continues to seek, acquire, and assess appropriate associations on an ongoing basis.

Alan Schwartz

See also Biases in Human Prediction; Dual-Process Theory; Heuristics

Further Readings

Allen, S. W., Brooks, L. R., Norman, G. R., & Rosenthal, D. (1988). Effect of prior examples on rule-based diagnostic performance. *Proceedings of the Annual Conference on Research in Medical Education, 27,* 9–14.

Ericsson, K. A. (2004). Deliberate practice and the acquisition and maintenance of expert performance in medicine and related domains. *Academic Medicine Research in Medical Education Proceedings of the forty-third annual conference November 7–10, 79*(10), S70–S81.

Hastie, R., & Dawes, R. M. (2001). *Rational choice in an uncertain world: The psychology of judgment and decision making*. Thousand Oaks, CA: Sage.

Hogarth, R. M. (2001). *Educating intuition*. Chicago: University of Chicago Press.

Kahneman, D. (2003). Maps of bounded rationality: A perspective on intuitive judgment and choice. In T. Frangsmyr (Ed.), *Les Prix Nobel. The Nobel Prizes 2002*. Stockholm: Almqvist & Wiksell.

Papa, F. J., Stone, R. C., & Aldrich, D. G. (1996). Further evidence of the relationship between case typicality and diagnostic performance: Implications for medical education. *Academic Medicine, 71*(Suppl. 1), S10–S12.

Papa, F. J., Oglesby, M. W., Aldrich, D. G., Schaller, F., & Cipher, D. J. (2007). Improving diagnostic capabilities of medical students via application of cognitive sciences-derived learning principles. *Medical Education, 41,* 419–425.

Stanovich, K. E., & West, R. F. (2000). Individual differences in reasoning: Implications for the rationality debate? *Behavioral and Brain Sciences, 23,* 645–726.

Windschitl, P. D., & Wells, G. L. (1996). Measuring psychological uncertainty: Verbal versus numeric methods. *Journal of Experimental Psychology: Applied*, *2*(4), 343–364.

Attention Limits

Broadly defined, attention is the focus of cognitive resources on processing information. Research on attention addresses the following questions: (a) What initiates the focus of cognitive resources on objects of psychological concern? (b) What causes the focus of cognitive resources to shift from one object to another? (c) How many objects, or how much information, can be kept in cognitive focus at one moment in time?

The psychological study of attention has investigated the three questions of initiation, change, and capacity of cognitive focus at many levels of information processing. At the lowest level are studies of how cognitive resources are focused when processing sensory information in the visual, auditory, olfactory, gustatory, and tactile domains. At the highest level are studies of cognitive focus on the rich, meaningful content of human thought that underlies making complex, real-world decisions such as those involved in medical diagnosis and treatment.

Attention has relevance to medical decision making at many levels of information processing. At the lowest level of information processing, attention supports a physician's detection of the physical characteristics of a patient that lead to a medical diagnosis of the patient's condition. This might include visual information about the patient's coloration; auditory information from their heartbeat, breathing, and gastrointestinal processes; and tactile and olfactory information that are unique to the patient's condition.

The important factors that initiate attention and limit the capacity of a medical decision maker's attention to sensory input are different from those for simple sensory events in abstract laboratory studies. Whereas the physical characteristics of a stimulus (such as its intensity and duration) have been shown to influence attention in simple laboratory tasks, a medical decision maker's expertise (as defined by his or her background, beliefs, and understanding) creates a mental model (or a schema) that plays a central role in determining what information, and how much information, the decision maker attends to and how that information is interpreted.

A medical decision maker's expertise also plays a central role in determining what information he

or she pays attention to when using executive, cognitive processes, in the absence of sensory input, to reason through a patient's medical conditions either to arrive at a diagnosis or to select a treatment program. The interplay between a decision maker's mental model of a medical problem and the effect of that mental model on directing the decision maker's attention is very important. The importance of the interplay is shaped by (a) a limit on how much information a decision maker can hold in mind at one moment (also known as span of apprehension) and (b) the need of the decision maker to incorporate the most relevant and important information within the span of his or her limited attention if he or she has to make a wise decision.

The practical importance of attention limits on medical decision making is great. Since 1956, psychologists have recognized that the capacity of human attention, or the span of human apprehension, is limited to between five and nine items. Thus, a decision maker presented with a complex medical problem is unlikely to be able to incorporate all the available information about that problem into his or her cognitive focus. Because of these limits, it is exceedingly important that the decision maker has sufficient expertise to prepare him or her to attend to the most important information. Otherwise, the quality of a medical decision is likely to be compromised by being based on a small set of less relevant information. The long medical education, internship, and residency that most doctors go through help develop and hone their mental models for making good medical decisions.

The practical importance of attention limits is especially great for patients involved in their own treatment decision making. Patients must acquire a reasonable understanding of the mental model that medical experts hold of their condition. In the absence of this mental model, patients are unlikely to discover what the most important issues are for understanding, diagnosing, and treating their own condition. Thus, patients interested in playing an active role in their medical treatment need to acquire a mental model that incorporates the set of variables medical experts agree are the most important for making a wise decision for their cases. Otherwise, patients are unlikely to appreciate the medical recommendations made to them and may perhaps insist on following a course of treatment that is unwise.

David A. Walsh

See also Cognitive Psychology and Processes

Further Readings

Deutsch, J. A., & Deutsch, D. (1963). Attention: Some theoretical considerations. *Psychological Review, 70,* 80–90.

Hershey, D. A., & Walsh, D. A. (2000). Knowledge versus experience in financial problem solving performance. *Current Psychology, 19,* 261–291.

Kahneman, D. (1973). *Attention and effort*. Englewood Cliffs, NJ: Prentice Hall.

Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review, 63,* 81–97.

Treisman, A., & Gelade, G. (1980). A feature-integration theory of attention. *Cognitive Psychology, 12,* 97–136.

Wright, R. D., & Ward, L. M. (2008). *Orienting of attention*. Oxford, UK: Oxford University Press.

Attraction Effect

The attraction effect (also known as the decoy effect or the asymmetric dominance effect) refers to a phenomenon in which adding an inferior alternative into an existing choice set increases the probability of choosing an alternative from the original set. The term *attraction effect* comes from the fact that an inferior alternative attracts attention or the choice share to one of the alternatives in the choice set. Because the attraction effect is caused by the addition of an inferior alternative, which is called a *decoy*, to a core choice set, it is also called the *decoy effect* (the decoy effect is a broader term than the attraction effect). Finally, the *asymmetric dominance effect* refers to a specific case of the attraction effect in which the decoy is asymmetrically dominated by one of the alternatives in the set.

The attraction effect has important theoretical implications because it violates some fundamental assumptions of many rational choice models. One such assumption is the *principle of regularity*, by

which the probability of choosing one alternative from an initial choice set cannot be increased by adding a new alternative. The attraction effect also violates an assumption that choices are independent of irrelevant alternatives.

Experimental Paradigm

For the attraction effect to occur, several conditions must be met. In a typical experimental setting (decision environment) in which the attraction effect is demonstrated, alternatives are defined on a few (usually two) attributes (or dimensions) in a decision space (see Figure 1 for a two-attribute decision space). In this decision space, two alternatives (A and B) form a core choice set. These alternatives are selected so that they are nondominating or competitive to each other. In Figure 1, A is weaker on Dimension 1 (e.g., the quality dimension) and stronger on Dimension 2 (e.g., the price dimension), while the reverse is the case for B. Then an alternative that is inferior to only one alternative in the core set, which is called a decoy, is added to the set. The alternative that is directly superior to the decoy is called the *target* (B) and the other alternative that does not have a dominance relation with the decoy is called the *competitor* (A). The attraction effect is demonstrated when the proportion of people choosing the target significantly increases when the decoy is present compared with when the decoy is absent. The decoy is rarely chosen in most cases. The attraction effect has been demonstrated using both the between- and within-subjects designs.

Decoy Types

There are six types of decoys studied in the literature. These decoys can be broadly divided into two categories depending on whether there is an asymmetric dominance relation between the target and the decoy: asymmetrically dominated decoys and nonasymmetrically dominated decoys. There are three decoy types in each category. The asymmetrically dominated decoys have been studied more extensively in the literature because they produce a greater attraction effect.

Asymmetrically Dominated Decoys

The asymmetrically dominated decoys include the range (R), frequency (F), and range-frequency (RF) decoys (see Figure 1). The range (R) decoy extends the range of the target on the dimension on which the target is weaker than the competitor. The frequency (F) decoy increases the frequency of alternatives along the dimension on which the target is stronger than the competitor. The range-frequency (RF) decoy combines the effect of the range decoy with the effect of the frequency decoy. All these three types of decoys are directly dominated by the target but not by the competitor.

Nonasymmetrically Dominated Decoys

The nonasymmetrically dominated decoys include the compromise (C), inferior (I), and range with symmetric dominance (RS) decoys (see Figure 1). The range with symmetric dominance (RS) decoy increases the range downward on the dimension on which the target is weaker than the competitor but is symmetrically dominated by both the target and the competitor. The inferior (I) decoy is similar to the range decoy in that it increases the range on the dimension on which the target is weaker than the competitor. But its value on the dimension on which the target is stronger than the competitor is also raised so that there is no longer direct dominance relation between the target and the decoy. Although the inferior decoy is not directly dominated by the target, it is clearly inferior to the target.

The compromise (C) decoy is produced by raising the value of the inferior decoy further along the dimension on which the target is stronger than the competitor. The compromise decoy appears more attractive than the inferior decoy and helps increase the probability of choosing the target by making the target appear to be a good compromise between the two extreme alternatives (i.e., the decoy and the competitor). The effect produced by the compromise decoy is called the *compromise effect*. Although this effect is similar to the attraction effect in that the decoy increases the choice probability of the target (so categorized as the decoy effect), it is distinguished from the attraction effect because there is no dominance relation between the target and the decoy.

Phantom Decoy

A phantom alternative refers to a choice option that appears real but is unavailable at the time of

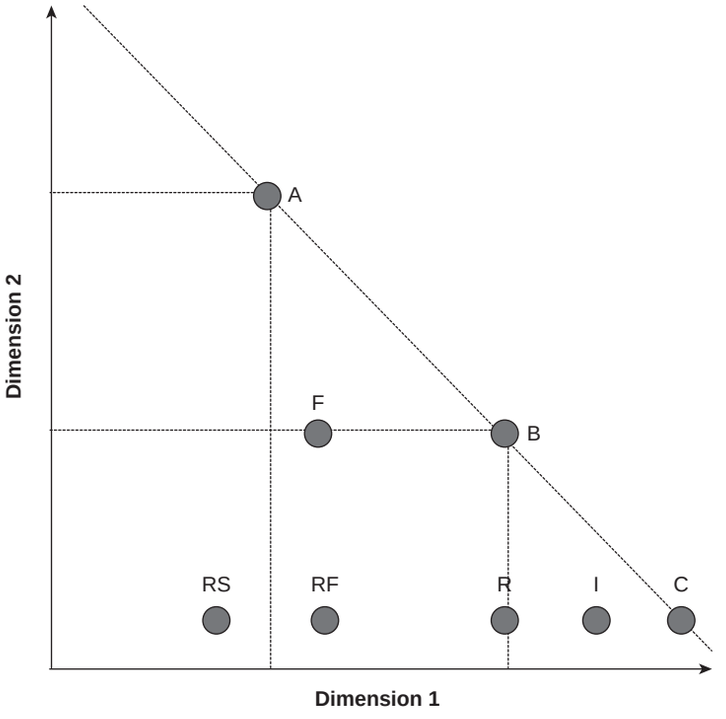


Figure 1 Graphical representation of the attraction effect with competitor (A), target (B), and six decoy types: (1) range, R; (2) frequency, F; (3) range-frequency, RF; (4) compromise, C; (5) inferior, I; and (6) range with symmetric dominance, RS

decision making. Examples include consumer goods that are out of stock, show tickets that are sold out, a job candidate who accepted another job, and so on. The decoys described above (i.e., asymmetrically dominated and nonasymmetrically dominated decoys) are available decoys; they have a potential to increase the choice share of the target by being present along with the target and the competitor. However, it has been demonstrated that the phantom decoy, which is not available when making a choice, also has a similar influence as the physically available decoys. Although some phantom decoys can also increase the choice probability for dominated alternatives, the phantom decoy effect appears to be procedurally similar to the attraction effect.

Decision Domains and Populations

The attraction effect has been demonstrated in various domains with diverse groups of people. The effect has been mostly studied in the consumer choice domain using product categories such as apartment, battery, beer, bicycle, boat, calculator, car, CD player, computer, film, gas barbecue grill, house, light bulb, lottery, microwave oven, mouthwash, orange juice, parking space, plane ticket, printer, restaurant, running shoes, video camera, sunscreen, toothpaste, TV sets, and wine. The effect has also been shown in the in-store and online purchases. In addition to numerous demonstrations of the effect in the consumer domain, it has also been shown in many other domains, such as the choices of partners, job candidates, political candidates in elections, investment options, and medicines. Furthermore, the effect has been shown with a wide range of people, including young adults ranging in age from the late teens to 30s and older adults in their 60s and 70s, undergraduate, graduate, and professional school students, grocery store customers, and internal medicine residents.

Factors Affecting the Attraction Effect

The size of the attraction effect is influenced by several factors, including perceived information relevance or meaningfulness of alternatives (mainly attribute values), product class knowledge, task involvement, perceived similarity between the decoy and the target, relative brand preference, choice share captured by the decoy, and perceived decoy popularity. More specifically, the attraction effect decreases with an increase in the perceived information relevance, product class knowledge (especially when attribute values are presented numerically), task involvement, and preference strength. Meanwhile, the effect increases with an increase in the perceived decoy-target similarity, choice share captured by the decoy, and perceived popularity of the decoy. Also, assuming that the two attributes on which alternatives are defined are quality and price, the attraction effect is stronger when the target is stronger on the quality dimension than the competitor compared with when the target is stronger on the price dimension than the competitor. Finally, the attraction effect (along with the compromise effect) is also influenced by motivational factors, such as prevention and promotion motivations. Specifically, prevention-focused people are more likely to show the compromise effect and less likely to show the attraction effect than promotion-focused people.

Theories and Explanations

Several theories and explanations have been proposed. They assume that the attraction effect occurs because people with limited knowledge do not have strong preformed preferences for attributes (because they do not know which attributes are important to them), and as a result, they are likely to focus on different attributes in different situations as the local decision context changes (as occurs when a decoy is present vs. not along with the core choice set).

Loss Aversion: Decoy as a Reference Point

Based on Tversky and Kahneman's reference-dependent theory of riskless choice that losses (or disadvantages) have greater impact on decision making than gains (or advantages), one explanation for the attraction effect is that a decoy may play a role of a reference point against which other alternatives are compared in terms of expected loss. According to the reference-dependent theory, an alternative with a moderate improvement on one attribute and no loss on the other is more attractive than another with a large improvement on one attribute and a small loss on the other. For example, if the range decoy is viewed as a reference point for both the target and the competitor, the target represents a small improvement on Dimension 2 and no loss on Dimension 1, whereas the competitor represents a large improvement on Dimension 2 and a small loss on Dimension 1 (see Figure 1 for the range decoy), and as a result, the target appears more attractive than the competitor in the presence of the decoy.

Weight Change: Context-Dependent Weighting

The weight-change model argues that adding a decoy changes the relative weights assigned to different attributes. That is, according to the weight-change model, the attraction effect occurs because the decoy causes decision makers to increase the relative weight they assign to the strong attribute of the target or decrease the relative weight they assign to the weak attribute of the target. For example, the relative weight given to a dimension decreases when the range of the value is extended because the attribute value differences become relatively smaller (see Figure 1 for the range decoy) or increases when the number of different attribute values on that dimension increases because the attribute value differences become relatively larger (see Figure 1 for the frequency decoy).

Value Shift: Perceptual Biases

The value-shift model argues that the subjective values assigned to each attribute value are shifted by the presence of the decoy, while weights on attributes remain constant. According to the value-shift model, a change in subjective evaluation of each attribute value leads to an increase in the overall value of the target relative to the competitor. This explanation is based on an idea that the decoy in the attraction effect operates in the same way as Parducci's range-frequency theory. For example, the addition of a decoy that has an

extremely low value on the dimension on which the target is weaker than the competitor should reduce the difference between the target and the competitor in their subjective values on the dimension (see Figure 1 for the range decoy). In another example, the addition of a decoy that has an intermediate value on the dimension on which the target is stronger than the competitor should increase the difference between the target and the competitor in their subjective values on the dimension (see Figure 1 for the frequency decoy). As a result, in the above examples, these decoys increase the relative attractiveness of the target by making it appear less weak than the competitor on the previously weak dimension and stronger than the competitor on the previously strong dimension.

Value Addition: Dominance Heuristic

The value-added model argues that relations among alternatives, such as presence of dominance, add value to the target and, as a result, cause the attraction effect. More specifically, the addition of a decoy to the core choice set creates a dominance relation between the target and the decoy, and this dominance relation adds justifiability value to the target because choosing the target becomes easier to justify with the presence of dominance.

Sunghan Kim

See also Accountability; Choice Theories; Context Effects; Loss Aversion; Prospect Theory

Further Readings

Heath, T. B., & Chatterjee, S. (1995). Asymmetric decoy effects on lower-quality versus higher-quality brands: Meta-analytic and experimental evidence. *Journal of Consumer Research, 22,* 268–284.

Highhouse, S. (1996). Context-dependent selection: The effects of decoy and phantom job candidates. *Organizational Behavior and Human Decision Processes, 65,* 68–76.

Huber, J., Payne, J. W., & Puto, C. (1982). Adding asymmetrically dominated alternatives: Violations of regularity and the similarity hypothesis. *Journal of Consumer Research, 9,* 90–98.

Huber, J., & Puto, C. (1983). Market boundaries and product choice: Illustrating attraction and substitution effects. *Journal of Consumer Research, 10,* 31–44.

Mishra, S., Umesh, U. N., & Stem, D. E., Jr. (1993). Antecedents of the attraction effect: An information-processing approach. *Journal of Marketing Research, 30,* 331–349.

Parducci, A. (1974). Contextual effects: A range-frequency analysis. In E. C. Carterette & M. P. Friedman (Eds.), *Handbook of perception* (Vol. 2). New York: Academic Press.

Tversky, A., & Kahneman, D. (1991). Loss aversion in riskless choice: A reference-dependent model. *Quarterly Journal of Economics, 106,* 1039–1061.

Wedell, D. H., & Pettibone, J. C. (1996). Using judgments to understand decoy effects in choice. *Organizational Behavior and Human Decision Processes, 67,* 326–344.

Attributable Risk

The concept of attributable risk (AR) is usually used in public health sciences to quantify the population impact of an exposure on overall disease burden. Such a population impact often has two determining factors: (1) strength of an association between the exposure and the disease and (2) the prevalence of exposure in the population of interest.

When exposure is simply binary, that is, exposed versus unexposed, a prototype measure of AR is defined by the so-called AR fraction:

$$\mathrm{AR} = 1 - \Pr\{D\backslash\bar{E}\}/\Pr\{D\},$$

where $\Pr\{D\}$ is the probability of having a disease for anyone in the population, and $\Pr\{D\backslash\bar{E}\}$ is the probability of having a disease only for those unexposed in the population. From a disease prevention perspective, $\Pr\{D\}$ can also be considered as a measure of overall disease burden on the population, while $\Pr\{D\backslash\bar{E}\}$ is considered as the measure of disease burden on the same population but with all exposure eliminated ideally.

For example, information was obtained in Tasmania during 1988 to 1990 from the parents of 2,607 1-month-old infants regarding their baby's usual sleeping positions. Table 1 tabulates the cumulative incidence of crib death (a sudden infant death syndrome) through 1 year of age in these infants by their usual positions: sleep prone (on their

Table 1 Crib death by sleeping position among infants

		Crib Death		
		Yes (*D*)	**No ($\bar{D}$)**	**Total**
Sleep Position	Prone (E)	9	837	846
	Other ($\bar{E}$)	6	1,755	1,761
	Total	15	2,592	2,607

Source: Dwyer, Posonby, Newman, and Gibbons (1991).

stomach) and other position (side or back). In this example, the overall death rate is calculated as $\Pr\{D\backslash\bar{E}\} = 15/2{,}607 = 5.75/1{,}000$, while the death rate of other sleeping positions is calculated as $\Pr\{D\backslash\bar{E}\} = 6/1{,}761 = 3.41/1{,}000$. Therefore, the AR is calculated as $\text{AR} = (1 - 3{,}41/5.75) \times 100\% = 40.7\%$.

An alternative form of the AR defined above is obtained by an application of Bayes's theorem:

$$\text{AR} = \Pr\{E\}(\text{RR} - 1)/[1 + \Pr\{E\}(\text{RR} - 1),$$

where RR is the relative risk measured by the ratio of $\Pr\{D\backslash E\}/\Pr\{D\backslash\bar{E}\}$. From this form, it is clear that AR is determined by $\Pr\{E\}$ and RR jointly. The AR increases as either the prevalence of exposed or the strength of association becomes greater, while it decreases otherwise. A greater RR alone, however, may not necessarily lead to a greater AR, which in fact shall additionally depend on the prevalence of exposure in the population. As a result, AR usually does not have the similar "portability" to the RR in etiologic inferences of a disease association among different populations. AR may vary from population to population. Nevertheless, its ability to jointly assess the association and the prevalence of exposure may serve itself a good measure in policy making to prioritize prevention strategies.

Conceptual Use and Interpretation

In practice, AR is usually used to assess the potential impact of prevention programs aimed to modify the exposure distribution in a target population. It can be used as a guide to evaluate and compare different preventive strategies.

For a specific prevention program, AR can be considered as a measure of impact of modifying multiple risk factors at the same time, although it is seemingly calculated only for one risk factor. It may be interpreted in two ways: one interpretation is that AR measures the potential impact due to modifying the distribution of the exposure and its correlated risk factors in the population, and the other interpretation is that AR for all known risk factors measures what knowledge has been gained about the disease etiology. Specifically in the latter interpretation, AR measures the remaining portion of overall disease burden that is not explained by the known risk 1 – AR factors. For example, most of the researchers consider that an AR less than 50% for known exposure and risk factors means that there is a strong need for further research in a disease area.

However, AR itself does not entail a comparable meaning to any conventional terms for risk. A variety of alternative terms have been used for AR, such as *population attributable fraction, etiologic fraction,* and *attributable fraction.* Many of these terms can be misleading to infer causality. AR itself merely reflects the overall impact of an association and the prevalence of exposure in a population. To avoid such confusion in causal inferences of an association, researchers have used AR to quantify the proportion of disease that can be related or linked, rather than attributed, to an exposure.

Properties

When an exposure is hazardous, that is, with an RR greater than 1, AR as a percentage lies between 0 and 1. When either RR is 1 or no one is exposed,

AR is 0. When an exposure is protective, that is, with RR less than 1, AR is less meaningful with a range of $(-\infty, 0)$. In practice, researchers can recode the exposure by having RR to be always greater than 1 to avoid negative AR.

As a joint measure of prevalence of exposure and RR, AR increases as the prevalence of exposure increases. This means that the value of AR depends on how reference exposure is determined. When an exposure is measured on a continuous scale, more stringent choice of threshold for hazardous exposure usually leads to higher AR. For example, some authors found that the AR was estimated at 38% of esophageal cancer attributed to an alcohol consumption of more than 80 g/day with reference level of 0 to 79 g/day. When the reference level changed to 0 to 39 g/day, the value of AR jumped to 70%. Therefore, when applying AR in prevention of a continuous exposure, it is critical to clearly specify the reference exposure for any meaningful estimation and comparison.

When a reference exposure is determined and the rest of exposure is categorized into several mutually exclusive categories, the sum of ARs for these categories equals the AR for these categories combined. This is called the distributive property of AR. For example, some authors found that the AR estimates are 13%, 6%, and 64% for malignant mesothelioma attributed to moderately low, medium, and high likelihoods of exposure, respectively, and summed up to a total of 83% of nontrivial (moderately low, medium, and high combined) likelihood of exposure. Given this property, some researchers argue that there is no need to break the exposure into finer categories, if the overall AR is of main concern, even when risk appears to increase with higher exposure levels.

Estimation

The estimation of AR usually depends on the study design that dictates how data are collected. In three major types of study design that are often used in public health research, that is, cross-sectional, cohort, and case-control, AR is almost always estimable with proper assumptions.

In cross-sectional studies, all the quantities that define AR are estimable, and the estimation of AR is usually straightforward.

In cohort studies, $\Pr\{D\backslash E\}$, $\Pr\{D\backslash \bar{E}\}$, and $\Pr\{E\}$ are usually estimable from the observed data. When the sampled cohort is a random sample of the population of interest, the estimated $\Pr\{E\}$ is likely comparable with that in the population, and hence the estimated AR is meaningful to the population. When the cohort is sampled with a predetermined proportion of exposure, the value of AR may be less meaningful for the cohort studies.

For case-control studies, researchers often use an alternative form to estimate AR:

$$\mathrm{Ar} = \Pr\{E \mid \mathrm{D}\}(1 - 1/\mathrm{RR}).$$

In this alternative form, $\Pr\{E/D\}$ can be estimated among the cases, that is, individuals with disease, and RR can be estimated by an approximation of the odds ratio (OR), assuming a rare disease.

When only one exposure is of sole concern without taking into account other factors, consider a prototype 2×2 table for observed data in Table 2, regardless of the underlying study design. Then a crude AR can simply be estimated by

$$\mathrm{AR} = (bc - ad)/(nb).$$

Variance of this AR estimator can be estimated by the Delta method for various distributions that are assumed in individual study designs. As a result, $100(1 - \alpha)\%$ -level confidence intervals can be constructed with properly transformed AR, for example, $\log(1 - \mathrm{AR})$, or $\log\{\mathrm{AR}/(1 - \mathrm{AR})\}$. Researchers have discussed extensively in statistical literature the merits of these transformations in confidence interval construction.

Crude AR tends to be biased when it ignores potential confounding factors for the association between exposure and disease. Adjusted AR has been advocated by researchers and methodologists to account for potential confounding factors. Similar to the usual adjustment techniques for RR estimation, adjusted AR can be calculated by a variety of nonparametric and model-based approaches.

One approach is by stratification. It is similar to the Mantel-Haensel approach in estimating OR from several strata. A crucial assumption is that a common RR or OR exists for all the strata. Based on the study designs, either $\Pr\{E\}$ or $\Pr\{D\backslash \mathrm{E}\}$ can be

Table 2 A prototype 2 × 2 table in epidemiologic studies

	Disease (D)	*No Disease ($\bar{D}$)*	
Exposed (E)	a	b	E
Not Exposed ($\bar{E}$)	*c*	*d*	*n – e*
	f	*n – f*	*N*

calculated for each stratum. Then applying an estimate of RR or OR would lead to consistent adjusted AR estimates. Variance of the estimators obtained by this approach tends to be complex but can be computed by the Delta method or the maximum likelihood methods for their large sample asymptotic properties. Empirical simulation has shown that their bias and coverage probability tend to be satisfactory in large samples as well.

A second approach is by calculating the weighted sum of ARs over strata, for example, as in

$$\mathrm{AR} = \sum_{s=1}^{S} w_s \mathrm{AR}_s,$$

where $s = 1,2,\ldots, S$, are stratum indicators, and w_s are the assigned weights for stratum-specific AR_s. This weighted approach usually does not require common RR or OR and yields a variety of types of adjusted AR by choosing different sets of weights. For example, an adjusted AR can be called "case-load" adjusted if w_s is the proportion of cases in stratum s, and "precision-weighted" if w_s is inversely proportional to the variance of stratum-specific AR estimators. Variance of the weighted adjusted AR can be similarly computed by the Delta method in large sample.

Model-based adjusted approaches have been extensively studied as well in statistical literature. For example, one such is in the form of

$$\mathrm{AR} = 1 - \sum_{s=1}^{S} \sum_{e=1}^{E} d_{se} / \mathrm{RR}_{e|s},$$

where $e = 1,2,\ldots, E$ are the levels of exposure, d_{se} are the cases, and $\mathrm{RR}_{e|s}$ are the adjusted *RR*. Note that this form is not exactly a maximum likelihood estimator. Alternative model-based estimators have also been proposed for case-control designs under unconditional logistic regression model and for cohort designs under unconditional logistic regression model and the Poisson model. In practice, these approaches would yield similar results in both small and large samples.

Extensions

The prototype AR is mostly used when both disease and exposure are dichotomous. Extensions of the prototype AR have been studied in various scenarios.

When exposure is not limited to be dichotomous, that is, exposed versus unexposed, those who are exposed can be further categorized into multiple levels of exposure. Then the prototype AR can be extended to the so-called partial or level-specific AR that represents the level-specific AR, which may have practical implication for screening high-risk groups. In literature, extended AR has been developed for continuous exposure.

When there are several types of exposure, researchers have estimated exposure-specific AR and the overall AR for all the exposure types jointly. Usually the sum of exposure-specific AR does not equal the overall AR. When different types of exposures are mutually independent and their effect on disease is multiplicative, then the product of exposure-specific complement AR, that is, 1 – AR, equals the complement overall AR.

When disease outcome is time-to-event outcome, T, say, extensions of time-varying AR have been proposed, such as $\mathrm{AR}(t) = 1 - F(t \mid \bar{E}) / F(t)$at time t, where $F(t \mid \bar{E}) = \Pr\{T \le t \mid \bar{E}\}$ is the cumulative distribution function of unexposed, and $F(t)$ is the cumulative distribution function of overall. When T is subject to censoring, statistical methods have been developed for a similar quantity

$AR(t) = 1 - \lambda(t \mid \bar{E})/\lambda(t)$, where $\lambda(t)$ is the hazard function of unexposed and $\lambda(t)$ is the hazard function of overall, under the widely used Cox proportional hazards model assuming that the relative hazards of $\lambda(t \mid E)|\lambda(t \mid \bar{E})$ is constant. Additional prognostic factors can be included in this model to calculate adjusted $AR(t)$.

AR has also been extended to accommodate ordinal data and recurrent disease events. Other AR-related quantities include the so-called AR in exposed, that is, $AR_e = 1 - \Pr\{D \mid \bar{E} / \Pr\{D \mid E\}$, which essentially plays the same role as RR, and the so-called preventable fraction, that is, $PF = 1 - \Pr\{D\} \mid \Pr\{D \mid \bar{E}\}$, for a protective exposure or intervention, which measures the impact of an association between disease and the protective exposure at the population level.

More generally from a disease prevention perspective, a very important concept that generalizes AR is the so-called generalized impact fraction (*IF*), which is defined as $IF = 1 - \Pr\{D\} / \Pr\{D\backslash\wp\}$, where $\Pr\{D\backslash\wp\}$ is the target disease burden due to modifying the exposure distribution in the population. The generalized *IF* can be used to assess various interventions targeting all subjects, or subjects at specified levels, while aiming at modifying the exposure distribution but not necessarily eliminate exposure. This *IF* can also be extended to censored time-to-event outcomes.

Ying Qing Chen

See also Bayes's Theorem; Cox Proportional Hazards Regression; Logistic Regression; Maximum Likelihood Estimation Methods; Screening Programs

Further Readings

Basu, S., & Landis, J. R. (1993). Model-based estimation of population attributable risk under cross-sectional sampling. *American Journal of Epidemiology, 142,* 1338–1343.

Benichou, J. (2001). A review of adjusted estimates of attributable risk. *Statistical Methods in Medical Research, 10,* 195–216.

Chen, Y. Q., Hu, C., & Wang, Y. (2006). Attributable risk function in the proportional hazards model for censored time-to-event. *Biostatistics, 7,* 515–529.

Drescher, K., & Schill, W. (1991). Attributable risk estimation from case-control data via logistic regression. *Biometrics, 47,* 1247–1256.

Dwyer, T., Posonby, A. L., Newman, N. M., & Gibbons, L. E. (1991). Prospective cohort study of prone sleeping position and sudden infant death syndrome. *Lancet, 337,* 1244–1247.

Greenland, S., & Robins, J. M. (1988). Conceptual problems in the definition and interpretation of attributable fractions. *American Journal of Epidemiology, 128,* 1185–1197.

Levin, M. L. (1953). The occurrence of lung cancer in man. *ACTA Unio Internationalis Contra Cancrum, 9,* 531–541.

Miettinen, O. S. (1974). Proportion of disease caused or prevented by a given exposure, trait or intervention. *American Journal of Epidemiology, 99,* 325–332.

Walter, S. D. (1976). The estimation and interpretation of attributable risk in health research. *Biometrics, 32,* 829–849.

Whittemore, A. S. (1982). Statistical methods for estimating attributable risk from retrospective data. *Statistics in Medicine, 1,* 229–243.

Automatic Thinking

It is commonly said that components of medical diagnosis and decision making, as well as of procedural skills, are executed automatically. This is not always good news. Patients may not appreciate a physician diagnosing their illness without awareness. Physicians, likewise, may not like to view themselves as unconscious automatons. Nonetheless, performing some task components automatically has advantages of speed and efficiency. Indeed, expert performance may depend on this automaticity: Allocating some necessary tasks to unconscious subroutines frees up attention for the more difficult customization of plans accommodating for the particulars of the situation. Automatic thinking may have some disadvantages, however. As it is difficult to reflect on automated thought processes, physicians cannot explain what they are thinking or teach students how to think that way. They may not notice if an automatic process is not going well and hence lose the opportunity to correct an error or improve execution. Insofar as it requires conscious reflection to change the way one executes a skill, the automated aspects of a physician's cognition may not improve with experience. Finally, it is difficult for others, and even for the physicians themselves,

to assess if their automated perception or decision making is biased by self-interest or is influenced by medical advertising or by their relations with manufacturer representatives.

Four varieties of automatic, unconscious thinking that have been described by cognitive psychologists can provide insight for understanding automated processes that occur in medical decision making. These characterizations conceive of automatic thinking (1) as a part of everyday skilled cognition, (2) as a problematic component of expertise, (3) as a characteristic of some motivational components of reasoning, and (4) as a feature of evolutionarily primitive cognition.

A common framework for all these views is the generic cognitive psychology model of knowledge and skill. For medicine, this holds that knowledge structures pertinent to diagnosis and treatment of patients, available in long-term memory, are activated into the physician's working memory when their pattern matches the pattern of the current situation, already attended in working memory. When the degree of fit seems adequate, that is, when the physician is confident that he or she understands the patient's illness, then the physician does the action available in the knowledge structure. When the knowledge does not seem to adequately fit the case, the physician does further work—gathering more case information, seeking more knowledge from other physicians or the literature, or problem solving by consciously reworking the available case information and knowledge. New knowledge structures built in this way are available for later use—whether or not they prove accurate for the present case. A knowledge structure that is useful because it helps explain a patient's disease or guide successful treatment may be more likely to be activated and relied on next time there is a similar patient. With experience, a larger set of specific knowledge structures is built up, so the physician can have a rapid, automatic yet appropriate response to a larger proportion of patients.

Automatic Thinking in the Execution of Everyday Skill

The first of cognitive psychology's explanations of automatic thinking in physicians is the account of ordinary learned knowledge or skill. Just as an experienced driver may arrive at a familiar destination and realize that he or she can't remember making any particular turns today, a physician may realize at the end of a routine day that he or she does not remember examining any of the patients. A surgeon may not remember the details of each layer of stitching. Yet every decision was made adequately and each step of the operation executed competently. Factors that would seem to prevent a physician's work life from being completely automatic in this way include the necessity to explain treatments to patients and helpers, to dictate or type for the medical record, or to explain to students. Yet with enough practice, such acts of communication too can be handled automatically.

Physicians are said to use heuristic strategies to make medical judgments or decisions, although they may not be aware that they do so unless it is called to their attention. This is a distinct concept from automatic thinking. The task of the physician is difficult because of its unavoidable uncertainty. Even when the physician is fully attending to a decision—not thinking "automatically" at all—there remains the problem of how to determine what is best to do. We can articulate very high standards for rational decision making that require extensive computation. Shortcut strategies are unavoidable, and one hopes that physicians use shortcuts that are usually accurate. A physician might consciously apply a heuristic strategy, or that strategy might be well learned and incorporated into a script or knowledge structure that comes to mind as a unit and thus is applied automatically. Thus, the concept of a heuristic strategy is different from automated thinking, even though it might be empirically demonstrable that physicians' behavior is consistent with the use of heuristic strategies more often when they are responding automatically to routine patients than when deliberating about an unusual case.

Automatic Thinking in Expertise

The second account of automatic thinking, while recognizing it may be essential for efficient performance, views it as a barrier to the improvement of performance that is necessary in the attainment of expertise. To put it in perspective, while it may take 50 hours of coached training and self-reflective practice to learn to drive competently enough that one can tune out that familiar route, to attain

a high level of driving skill (as required by a racer or a stunt driver) may require thousands of hours of supervised practice. If we assume that a skill that is executed automatically cannot be changed, then a physician diagnosing or managing a patient without attention cannot improve his or her skill. On the other hand, reaching high levels of skill may require that most of the constituent components of the skill have been overlearned so that they may be executed automatically, freeing up the physician's attention to focus on one particular element that needs to be adjusted and improved. Thus, automatic thinking is both a barrier to and a necessary precondition of the attainment of expertise.

Automatic Thinking in the Motivational Components of Reasoning

The third characterization of automatic thinking focuses on the human motivations and perceptions that imbue the setting in which the physician works. Factors in the external social context, mediated by internal motivations, may influence the thinking process without the physician's intending they do so and without the physician's awareness; these can be characterized as automatic processes. Unlike the automated rationality characteristic of everyday skill or expert cognition, where it is assumed that the physician previously performed these mental operations consciously and intentionally, this motivated automatic thinking may express all-too-human, irrational motives that the physician would not necessarily endorse if they were drawn to his or her attention.

The effect of irrelevant aspects of names is an example of nonrational automatic thinking. If a name carries a value-laden connotation, physicians may unwittingly react with respect to those values. Consider two physicians, Dr. Goode and Dr. Crapp. We might guess that other physicians may be more likely to refer patients to Dr. Goode. At a deeper level, Dr. Goode himself may have been subtly influenced his entire life to live up to his name, acquiring a fundamentally more sound mastery of medicine and a more conventionally upright value system than his colleague, unless the colleague became a gastroenterologist.

A more serious effect of the automatic response to names may be seen in the effect of the labels assigned to ventilation-perfusion scans used when pulmonary embolism is suspected. A patient with a "low probability" scan has a higher probability than normal of having a pulmonary embolism, but the automatic connotations of the "low probability" label—relief, relaxation, having nothing to worry about—may cause a physician to reduce vigilance below what would be appropriate.

Merely being reminded of money can change the way people think and act. Physicians in the United States are daily reminded of its importance, in their regular business meetings, phone calls they must make to protest insurance company denials and justify procedures that patients need, or conversations in which staff threaten to work elsewhere for higher pay. Studies have shown that when people are reminded of money, they tend to help other people less, to hold themselves more separate from others, and to work harder to solve intellectual problems. These automatic changes in their way of thinking can affect physicians' ways of gathering information and using it to make decisions for their patients, in ways that may either promote or impede their Hippocratic values.

The feelings and norms of friendship, social exchange, and mutual obligation can influence physicians' thinking without their awareness. Pharmaceutical detailers cordially and generously provide physicians food and small gifts in exchange for the privilege of delivering brief messages about the advantages of the medications they sell. Physicians, who feel uncomfortable being paid simply for listening, engage the sales representatives in friendly conversations to express their gratitude, and of course the friendliness is reciprocated. As an automatic effect of these friendly exchanges, shown in multiple studies, the physicians increase their use of the products after such visits. This increase occurs even when the physicians declare that they feel no obligation and that they listen to the marketing message with objective skepticism.

Automatic thinking underlies the observation that a physician who does not like a patient cannot be a good doctor for that patient due to the automatic effects of that dislike on each of the parties. The patient perceives some form of disinterest and, without intending, talks less, thus providing less information about the illness. The physician, responding both to sensed patient attitude and to own level of interest, is less likely to dig for more information or to spontaneously ruminate about

the patient's case. These automatic social responses work together to reduce the quality of the cognition the physician applies to a disliked patient.

A similar confluence of automatic responses may underlie the effects of prejudice and stereotyped expectations in the structured apprenticeships of physicians' clinical education. A generation or less ago, for example, some older male surgeons had low expectations of female residents' performance, as well as conflicting social role expectations. This led to offering them less help and coaching and impeded the establishment of warm collegial relations. In this environment, the female residents' automatic thinking led them to make fewer requests for supervision. The result of these enmeshed automatic responses was less educational progress and lower evaluations for the residents. Similar tales have been told regarding African American medical students, residents who graduated from foreign medical schools, and even family medicine residents on specialty rotations.

An important class of automatic influence of the sociomotivational context on thinking is the influence of fear and anxiety. This is manifest in the widespread habit of excessive testing to protect against the remotest possibilities, picked up as a standard operating procedure, often without the physicians recognizing that this part of their script serves more to allay physician anxiety than to protect the patient.

A more vivid manifestation may be seen in physicians' thinking under the real threat of death, in the traditional approach to end-of-life care. When stunned patients and family members deal with the high probability of death, physicians naturally strive to do all they can do, to muster all their power in line with the traditional injunction to preserve life at all costs, despite its futility. Recently, new institutions have been developed to redirect the automatically activated motivations to exercise control in the face of anxiety and to adhere to authority's precepts in the face of the possibility of death. The alternative approach is embodied in the standards of palliative care, which provides a new set of skills to exercise to give the physician something useful to do when the patient is dying. It incorporates a new authoritative framework, including the laws authorizing Do Not Resuscitate orders and living wills, the legitimacy of the Advance Directive, and the proxy decision maker. It is a good example of the design of institutions to cope with the automatic thinking elicited by the most challenging situations physicians face, demonstrating that this form of automatic thinking need not be opposed to rationality.

Automatic Thinking as a Feature of Evolutionarily Primitive Cognition

The fourth account of physicians' automatic thinking attributes the availability of particular types of knowledge structures or reasoning strategies to instincts inherited from our mammalian or reptilian ancestors. Consider, for example, that for eons we have had the capability to hold an object in each hand and sense which is heavier. From this, we speculate, arises our habit of making comparisons between just two treatment options at a time, rather than three or more. This unexamined tendency, traceable to how our minds are embodied, might lead physicians to pay insufficient attention to third or fourth options, the error of premature closure.

The generic cognitive model tells us that physicians may have several cognitive processes going on in parallel, some of them attended and others proceeding automatically. Among the automatic background processes may be evolutionarily primitive processes that scan the environment for danger or for items of appetitive interest. The physician may experience this when an ongoing diagnostic process is interrupted by an involuntary perception of the patient. A holistic assessment of the patient, such as "this patient looks sick," may trump the usual 20 Questions diagnosis game and lead directly to action. This kind of interruption has been characterized as a competition between two systems of thinking, although likely there are more than just two systems. It has been suggested, furthermore, that ideas from the simpler, more primitive system are more likely to control thinking when the physician is tired, distracted, under time pressure, or venturing into unfamiliar territory.

For physicians to make decisions at the optimal standard of rationality could require cognitively intense calculation using all available information, but instead physicians use shortcut strategies that refer to subsets of the information. Such strategies may be traced, it has been suggested, to our evolution in environments with multiple, partially redundant cues, which has endowed us with a

special capability to learn particular types of decision strategy appropriate for such environments. (This is analogous to the claim that humans have an inborn capability to learn the grammar of language.) Thus, while experts might identify 50 signs or symptoms associated with the various causes of chest pain, those symptoms tend to be correlated with each other. One physician could attend to one subset of symptoms, another physician to a different subset, and each could diagnose chest pain accurately. Simple strategies, such as to choose a diagnosis by counting the arguments (features) for each, may be sufficient to support rapid yet accurate decisions that have important consequences.

There is disagreement about whether it is necessary to invoke evolutionary selection to account for the availability of such simple yet effective strategies. In computer simulations, such strategies produce adequate accuracy while using fewer resources and taking less time. With these advantages, even if evolution had not provided the strategies, we would have had to invent them. If physicians adopt such strategies because they are easy and effective, then we may not need to invoke an inbuilt grammar of decision making to explain their use. Nonetheless, the analysis of the fit between the physicians' simplified decision strategies and the structure of the disease environment provides a useful perspective on automatic cognition.

Implications

Physicians' automatic thinking is important because it potentially affects decisions whose outcomes matter, allowing gains in efficiency and providing helpful insights, or perhaps causing the physician to ignore some of the information available about a patient. With the four competing accounts of automatic thinking, there are many opportunities for researchers to describe, explain, and assess its role in physicians' decision making, and to resolve the unknowns concerning the source, role, and malleability of the automatic parts of cognition. Such research will be challenging, however; because automatic cognition is difficult to self-report, experts are reluctant to submit selves to intensive observation, and observing such behavior is likely to change it.

Robert M. Hamm

See also Associative Thinking; Cognitive Psychology and Processes; Context Effects; Decision Making and Affect; Dual-Process Theory; Heuristics; Intuition Versus Analysis; Irrational Persistence in Belief; Pattern Recognition

Further Readings

Abernathy, C. M., & Hamm, R. M. (1995). *Surgical intuition*. Philadelphia: Hanley & Belfus.

Bargh, J. A., & Ferguson, M. J. (2000). Beyond behaviorism: On the automaticity of higher mental processes. *Psychological Bulletin, 126*(6), 925–945.

Betsch, T., & Haberstroh, S. (Eds.). (2005). *The routines of decision making*. Mahwah, NJ: Lawrence Erlbaum.

Bilalić, M., McLeod, P., & Gobet, F. (2008). Inflexibility of experts: Reality or myth? Quantifying the Einstellung effect in chess masters. *Cognitive Psychology, 56,* 73–102.

Bursztajn, H., Feinbloom, R. I., Hamm, R. M., & Brodsky, A. (1981). *Medical choices, medical chances*. New York: Delacorte.

Cain, D. M., & Detsky, A. S. (2008). Everyone's a little bit biased (even physicians). *Journal of the American Medical Association, 299*(24), 2893–2895.

Ericsson, K. A. (2004). Deliberate practice and the acquisition and maintenance of expert performance in medicine and related domains. *Academic Medicine, 79*(Suppl. 10), S70–S81.

Gigerenzer, G. (1996). The psychology of good judgment: Frequency formats and simple algorithms. *Medical Decision Making, 16*(3), 273–280.

Stanovich, K. E., & West, R. F. (2000). Individual differences in reasoning: Implications for the rationality debate? *Behavioral and Brain Sciences, 23*(5), 645–726.

Vohs, K. D., Mead, N. L., & Goode, M. R. (2008). Merely activating the concept of money changes personal and interpersonal behavior. *Current Directions in Psychological Science, 17*(3), 208–212.

Wood, W., & Neal, D. T. (2007). A new look at habits and the habit-goal interface. *Psychological Review, 114*(4), 843–863.

Axioms

Axioms in the days of Euclid (Euclidian Geometry) were self-evident truths within logic and mathematics. Axioms today (as considered within a

system or theory) are sets of rules that are internally consistent. For example, axioms in expected value decision making are a set of internally consistent rules for rational choice. Axioms are often described in terms of how plausible they are, in what sense(s) they may or may not be compelling, and how the axioms are stated. In the latter case, the phrase *elegantly simple* can be directed at the expression of a set of axioms that are viewed positively regarding their statement.

Axioms and Theories

One can start with a set of axioms and then move toward the development of a theory, or one can state a theory and look for or attempt to develop the set of axioms needed to support that theory. In the latter approach, an axiomatic method is proposed that contains a set of postulates (first principles). Within such a set of postulates, the postulates that are stated are "all and only" the necessary definitions and assumptions from which the theory can be derived.

The verb *to axiomatize* suggests that one can take a theory and attempt to derive first principles for or in support of that scientific or social scientific theory. For example, one can have a theory, such as a form of expected utility theory, and take that theory's claims and attempt to axiomatize that form of expected utility theory. Another claim is that an axiomatic method can be used to express all significant theories of any scientific or social scientific discipline and one can further argue that any scientific or social scientific discipline should be capable of such axiomatic expression.

The Axiomatic Versus the Empirical

Some might argue that "the axiomatic" is contrasted with "the empirical," where, for example, the axiomatic refers to logical deduction and the empirical refers to data derived from the real world (by observation alone, observation with measurement, or experimentation) to objectively study and test hypotheses. Here, experimentation includes baseline observation and measurement, introduction of an intervention, and then postintervention reobservation and remeasurement, analysis of data, interpretation of analyzed data, and the drawing of conclusions. Once tested in one environment or setting, a central objective in the development of an empirical science is coherence in the application of theory obtained from this one source to a different source. This coherence then confirms the extensibility of the application of that theory.

The above proposed distinction between what is axiomatically driven and what is empirically driven seems to suggest that the axiomatic is based on consideration of first principles, while the empirical is based on consideration of observations and measurements at a time and over time with or without experimentation. However, this assumption can be challenged as an oversimplification on two grounds. First, a science or a scientific theory itself can be axiomatized. Second, an axiomatic theory can be tested as a framework for understanding a particular science or social science in its application to the real world. For example, as a social science, a theory of human decision-making behavior can be developed and tested in terms of the truth or falsity of its set of axioms as explaining real-world behavior.

Let us take an example of an attempt to explain human behavior on the basis of first principles. A theoretician can be approached about the behavior of humans in the real world and asked the question: Why do humans when placed in this particular setting behave in the fashion that they do? This was a question posed to Daniel Bernoulli: Why do gamblers behave as they do in the St. Petersburg paradox—a coin-flipping game in which most gamblers do not behave like rational bettors? This is a game that has a theoretical return on investment of an infinite sum of money, but it is counterintuitive to realize that. So a typical person won't appreciate the return on investment and will not be willing to pay much in order to play this game. Bernoulli's explanation was that gamblers behave in the St. Petersburg paradox as if they were maximizing the expectation of some utility function of the possible outcomes in the problem facing them. The classical resolution of the paradox involved the introduction of (a) a utility function, (b) the statement of an expected utility hypothesis, and (c) the presumption of diminishing marginal utility of money. Here, the gambling behavior existed (the St. Petersburg paradox was recognized as a problem needing a solution), and Bernoulli was asked to consider an explanation for the behavior

(a solution of the paradox). Bernoulli then developed the set of first principles to explain the behavior (to attempt to solve the paradox). The main contributors to the axiomatic derivation of expected utility theory are John Von Neumann and Oskar Morgenstern, Frank Ramsey, Bruno de Finetti, and Leonard Savage.

Once one has an explanatory hypothesis involving human decision behavior (as in Bernoulli's solution to the St. Petersburg paradox), one can ask whether this hypothesis can be further broken down into a set of axioms that can be tested (one at a time) in the real world as verifiable or falsifiable. It also needs to be recognized that certain axioms, hypotheses, and theories can be true to a specifiable extent under one set of circumstances and false to a specifiable extent under another set of circumstances in terms of the real world.

Axioms of Expected Utility Theory

Axioms and what can be done to these first principles can perhaps be best represented by an "equality" and what can be done to both sides of the equality while still preserving the equality: What can be done to one side of an equal sign in an equality and to the other side of the equal sign of the equality and still maintain the equality? Or in the statement of an equation, it can be asked, "What can be done to both sides of the equation while still preserving the truth value of the equation?"

Two axioms of expected utility theory are (1) comparability (If A and B are in the alternative set S, then either $A > B$ or $B > A$, or both $A = B$); and (2) transitivity (If $A > B$, and $B > C$, then $A > C$).

Testing an Axiom in a System

Real-world settings can be used to test the viability of an axiom of a system or theory to ensure through the testing of the complete set of axioms used to express a system or theory that it is consistent with what is found in the real world. In such a testing setting, one needs to demonstrate that real-world decision makers behave according to the axioms (verification) or do not behave according to the axioms (falsification) of the system or theory. And one can proceed to test each axiom of the set of axioms in the system or theory under consideration.

Testing the Axioms of Expected Utility Theory

In testing the axioms of expected utility theory in the real world, one can start with the axioms and ask in a real-world setting whether humans behave according to the axioms. But how does one go about testing the axioms of expected utility theory in the real world?

Here, one needs to introduce a methodology (technique) for elicitation of preferences, for example, the standard gamble. Once one has the axioms and the methodology, one can test the axioms in terms of verification or falsification.

How would one identify the presence of a falsification of an axiom? If the fundamental value that humans place on any particular health outcome varies according to the position of the outcome in the procedure (i.e., the standard gamble) used to elicit that individual's preferences, then there would be a failure of an axiom in the system and, thus, an internal inconsistency in the system or theory. An internal inconsistency exists when one or more of the main axioms fail to be sustained on real-world testing.

Inconsistencies

What do internal inconsistencies look like in the case of the axioms of expected utility theory? If in a real-world economic setting, there is a divergence between what a human is willing to pay for a good that he or she does not possess and his or her willingness to receive compensation for giving up that same good when he or she does possess it, then there may be an inconsistency in one of the axioms in the system or theory used to explain human decision-making behavior in this economic decision-making setting.

If, in a real-world medical setting, a patient places a greater value on an outcome with a given probability when that outcome is described (framed, presented) in terms of the chance of "survival," then when it is redescribed (reframed, re-presented) in an equivalent way in terms of the chance of "dying," there may be an inconsistency in one of the axioms in the system or theory used to explain human decision-making behavior in this medical decision-making setting. For example, if a patient is willing to accept surgery when the

risk of surgery is described as having a 90% chance of surviving the initial surgery and still being alive 6 months after the surgery but not when that same surgery is described as having a 10% chance of dying at the time of the surgery and not being alive 6 months after the surgery, then there may be an inconsistency in one of the axioms used to explain medical decision-making behavior.

Focusing on the question of the internal consistency of the standard gamble, one can ask, "What is to be done about the internal inconsistencies that are found with the use of the standard gamble?" One research goal is to achieve internal consistency within the standard gamble. Another research goal is an intermediate goal to limit the level of internal inconsistency found within the standard gamble. Here, research studies focus on the attempt to incorporate additional elements (e.g., weighting and probability transformation parameters) to the standard gamble valuation procedure in the attempt to limit internal inconsistency.

Future Directions

It is not the case that a theory is captured by one and only one set of axioms. In the attempt to axiomatize any scientific or social scientific theory, there is always a search under way for the most plausible, the most compelling, and the most elegantly simple set of axioms used to capture a theory in any domain. The search for such axioms is ongoing in every logical, mathematical, scientific, and social scientific field today. In the case of the sciences and social sciences, the search continues for the most plausible, most compelling, and the most elegantly simple set of axioms that can be applied to the real world attempting to successfully explain human behavior in decision making. Such a search should have as its ultimate goal not only describing human behavior but also optimizing that decision making to help people achieve their goals as viewed from their perspectives.

Dennis J. Mazur

See also Expected Utility Theory; Gain/Loss Framing Effects; Risk Aversion

Further Readings

Allais, M. (1979). The foundations of a positive theory of choice involving risk and criticism of the postulates and axioms of the American School. In M. Allais & O. Hagen (Eds.), *Expected utility hypotheses and the Allais paradox*. Dordrecht, the Netherlands: Reidel.

Arrow, K. J. (1971). The theory of risk nearing. In *Essays in the theory of risk bearing*. Chicago: Markham.

Battalio, R. C., Kagel, J., & MacDonald, D. N. (1985). Animals' choices over uncertain outcomes. *American Economic Review, 75,* 597–613.

Camerer, C. (1989). An experimental test of several generalized utility theories. *Journal of Risk and Uncertainty, 2,* 61–104.

Chateauneuf, A., & Wakker, P. (1999). An axiomatization of cumulative prospect theory for decision under risk. *Journal of Risk and Uncertainty, 18,* 137–145.

Chew, S., & Waller, W. (1986). Empirical tests of weighted expected utility theory. *Journal of Mathematical Psychology, 30,* 55–62.

Edwards, W. (1955). The prediction of decisions among bets. *Journal of Experimental Psychology, 50,* 201–214.

Edwards, W. (1962). Subjective probabilities inferred from decisions. *Psychology Review, 69,* 109–135.

Ellsberg, D. (1961). Risk, ambiguity and the savage axioms. *Quarterly Journal of Economics, 75,* 643–669.

Machina, M. (1982). "Expected utility" analysis without the independence axiom. *Econometrica, 50,* 277–323.

Oliver, A. J. (2004). Testing the internal consistency of the standard gamble in "success" and "failure" frames. *Social Science & Medicine, 58,* 2219–2229.

von Neumann, J., & Morgenstern, O. (1944). *Theory of games and economic behavior*. Princeton, NJ: Princeton University Press.

B

Basic Common Statistical Tests: Chi-Square Test, *t* Test, Nonparametric Test

A statistical test provides a mechanism for making quantitative decisions about a process or processes. The intent is to determine whether there is enough evidence to "reject" a conjecture or hypothesis about the process. The conjecture is called the null hypothesis. In medical research, appropriate use of a test, correctly interpreting *p* values, and drawing valid conclusions may help to clarify the confusion between statistical and clinical significance and make judicious decisions. The rest of this entry is organized as follows: Beginning with the chi-square test, the most applicable test for categorical data, this entry introduces *t* test and analysis of variance (ANOVA) for quantitative outcomes, ending with the introduction of nonparametric tests.

Chi-Square Test

Chi-square test is a statistical test commonly used to compare observed data with data a researcher would expect to obtain according to a specific hypothesis. Chi-square tests can be used in tests of goodness of fit, testing if a sample of data came from a population with a specific distribution; or in tests of independence when a researcher wants to see if there is a relationship between categorical variables. In this case, the outcome is categorical—for example, whether people from different regions differ in the frequency with which they report that they support a political candidate.

Pearson Chi-Square

Pearson chi-square is used to assess the above two types of comparison. It is the most common test for significance of the relationship between categorical variables. The chi-square test becomes increasingly significant as the numbers deviate further from this expected pattern. The value of the chi-square and its significance level depend on the overall number of observations and the number of cells in the table. Relatively small deviations of the relative frequencies across cells from the expected pattern will prove significant if the number of observations is large.

The Pearson chi-square inherently tests the underlying probabilities in each cell; and when the expected cell frequencies fall, for example, below 5, those probabilities cannot be estimated with sufficient precision. Therefore, the assumption underlying the use of the Pearson chi-square is that the expected frequencies are not very small.

Maximum Likelihood Chi-Square

Based on maximum likelihood theory, the maximum likelihood chi-square tests the same hypothesis as the Pearson chi-square statistic, and in practice, it is usually very close in magnitude to the Pearson chi-square statistic.

Fisher Exact Test

When conducting a chi-square test in which one or more of the cells have an expected frequency of 5 or less, the Fisher's exact test is used. This test is only available for 2 × 2 tables and is based on the following rationale: Given the margins of the table, and assuming that in the population the two factors in the table are not related (null hypothesis), the probability of obtaining cell frequencies as uneven or worse than the ones that were observed can be computed exactly by counting all possible tables that can be constructed based on the marginal frequencies.

McNemar Chi-Square

This test is primarily used in a before-after design study; that is, it assesses the significance of the difference between two dependent samples. For example, researchers may count the number of students who fail a test of minimal math skills at the beginning of the semester and at the end of the semester. In a 2 × 2 table, the McNemar chi-square tests whether the counts in cells above the diagonal differ from counts below the diagonal. If the two counts differ significantly, this reflects change between the samples, such as change due to an experimental effect between the before and after samples.

t Test

A parametric test is a statistical test that assumes an underlying distribution of observed data. *t* test is one of the most common parametric tests and can be categorized as follows.

One-Sample t *Test*

One-sample *t* test is used to test whether the population mean of the variable of interest has a specific value (hypothetical mean), against the alternative that it does not have this value, or is greater or less than this value. A *p* value is computed from the *t* ratio (which equals the difference of the sample mean and the hypothetical mean divided by the standard error of mean) and the numbers of degrees of freedom (which equals sample size minus 1). If the *p* value is small, the data give more possibility to conclude that the overall mean differs from the hypothetical value.

Two-Sample t *Test*

The two-sample *t* test is used to determine if the means of the variable of interest from two populations are equal. A common application of this is to test if the outcome of a new process or treatment is superior to a current process or treatment.

t Test for Independent Samples

An independent samples *t* test is used when a researcher wants to compare the means of a variable of interest (normally distributed) for two independent groups, such as the heights of gender groups. The *t* ratio is the difference of sample means between two groups divided by the standard error of the difference, calculated by pooling the standard error of the means of the two groups.

t Test for Dependent Samples

If two groups of observations of the variable of interest (that are to be compared) are based on the same sample of subjects who were tested twice (e.g., before and after a treatment); or if the subjects are recruited as pairs, matched for variables such as age and ethnic group, and one of them gets one treatment, the other an alternative treatment; or if twins or child–parent pairs are being measured, researchers can look only at the differences between the two measures of the observations in each subject. Subtracting the first score from the second for each subject and then analyzing only those "pure (paired) differences" is precisely what is being done in the *t* test for dependent samples; and, as compared with the *t* test for independent samples, this always produces "better" results (i.e., it is always more sensitive). The *t* ratio for a paired *t* test is the mean of these differences divided by the standard error of the differences.

Assumptions

Theoretically, the *t* test can be used even if the sample sizes are very small (e.g., as small as 10) so long as the variables of interest are normally distributed within each group, and the variation of scores in the two groups is not reliably different.

The normality assumption can be evaluated by looking at the distribution of the data (via histograms) or by performing a normality test. The equality of variances assumption can be verified

with the *F* test, or the researcher can use the more robust Levene's test.

Analysis of Variance

Analysis of variance (ANOVA) is a statistical test that makes a single, overall decision as to whether a significant difference is present among three or more sample means of the variable of interest (outcome). An ANOVA is similar to a *t* test; however, it can also test multiple groups to see if they differ on one or more explanatory variables. The ANOVA can be used to test between-groups and within-groups differences. There are two types of ANOVAs: one-way ANOVA and multiple ANOVA.

One-Way ANOVA

A one-way ANOVA is used when there are a normally distributed interval outcome and a categorical explanatory variable (with two or more categories), and the researcher wishes to test for differences in the means of the outcome broken down by the levels of the explanatory variable. For instance, a one-way ANOVA could determine whether class levels (explanatory variable), for example, freshmen, sophomores, juniors, and seniors, differed in their reading ability (outcome).

Multiple ANOVA (Two-Way ANOVA, N-Way ANOVA)

This test is used to determine if there are differences in two or more explanatory variables. For instance, a two-way ANOVA could determine whether the class levels differed in reading ability and whether those differences were reflected by gender. In this case, a researcher could determine (a) whether reading ability differed across class levels, (b) whether reading ability differed across gender, and (c) whether there was an interaction between class level and gender.

Nonparametric Test

Nonparametric methods were developed to be used in cases when the researcher knows nothing about the parameters of the variable of interest in the population. Nonparametric methods do not rely on the estimation of parameters (such as the mean or the standard deviation) describing the distribution of the variable of interest in the population.

Nonparametric methods are most appropriate when the sample sizes are small. In a nutshell, when the samples become very large, then the sample means will follow the normal distribution even if the respective variable is not normally distributed in the population or is not measured very well.

Basically, there is at least one nonparametric equivalent for each parametric general type of test. In general, these tests fall into the following categories.

One-Sample Test

A Wilcoxon rank sum test compares the median of a single column of numbers against a hypothetical median that the researcher enters. If the data really were sampled from a population with the hypothetical mean, one would expect the sum of signed ranks to be near zero.

Differences Between Independent Groups

Nonparametric alternatives for the *t* test for independent samples are the Mann-Whitney *U* test, the Wald-Wolfowitz runs test, and the Kolmogorov-Smirnov two-sample test. The Mann-Whitney *U* test, also called the rank sum test, is a nonparametric test assessing whether two samples of observations come from the same distribution. This is virtually identical to performing an ordinary parametric two-sample *t* test on the data after ranking over the combined samples. The Wald-Wolfowitz runs test is a nonparametric test of the identity of the distribution functions of two continuous populations against general alternative hypotheses. The Kolmogorov-Smirnov two-sample test is one of the most useful and general nonparametric methods for comparing two samples, as it is sensitive to differences in both location and shape of the empirical cumulative distribution functions of the two samples.

An appropriate nonparametric alternative to the one-way independent-samples ANOVA can be found in the Kruskal-Wallis test, which is applicable when the researcher has the outcome with two or more levels and an ordinal explanatory variable. It is a generalized form of the Mann-Whitney test method, since it permits two or more groups.

Differences Between Dependent Groups

For the *t* test for dependent samples, the nonparametric alternatives are the Sign test and Wilcoxon's matched pairs test. The sign test can be used to test that there is "no difference" between the continuous distributions of two random samples. The Wilcoxon test is a nonparametric test that compares two paired groups, through calculating the difference between each set of pairs and analyzing that list of differences. If the variables of interest are dichotomous in nature (i.e., "pass" vs. "no pass"), then McNemar's chi-square test is appropriate. If there are more than two variables that were measured in the same sample, then the researcher would customarily use repeated measures ANOVA. Nonparametric alternatives to this method are Friedman's two-way ANOVA and Cochran *Q* test. Cochran *Q* is an extension to the McNemar test and particularly useful for measuring changes in frequencies (proportions) across time, which leads to a chi-square test.

Relationships Between Variables

Spearman *R*, Kendall tau, and coefficient gamma are the nonparametric equivalents of the standard correlation coefficient to evaluate a relationship between two variables. The appropriate nonparametric statistics for testing the relationship between the two categorical variables are the chi-square test, the phi coefficient, and the Fisher exact test. In addition, Kendall coefficient of concordance is a simultaneous test for relationships between multiple cases, which is often applicable for expressing interrater agreement among independent judges who are rating (ranking) the same stimuli.

Li Zhang

See also Analysis of Covariance (ANCOVA); Analysis of Variance (ANOVA); Sample Size and Power

Further Readings

Agresti, A. (2007). *An introduction to categorical data analysis* (2nd ed.). New York: Wiley.

Agresti, A., & Franklin, C. (2007). *The art and science of learning from data*. Upper Saddle River, NJ: Prentice Hall.

Bishop, Y. M. M., Fienberg, S. E., & Holland, P. W. (1975). *Discrete multivariate analysis: Theory and practice*. Cambridge: MIT Press.

Fienberg, S. E. (1977). *The analysis of cross-classified categorical data*. Cambridge: MIT Press.

Kachigan, S. K. (1986). *Statistical analysis: An interdisciplinary introduction to univariate & multivariate methods*. New York: Radius Press.

Kendall, M., & Stuart, A. (1979). *The advanced theory of statistics* (Vol. 2, 4th ed.). London: Griffin.

Mendenhall, W. (1975). *Introduction to probability and statistics* (4th ed.). North Scituate, MA: Duxbury Press.

Runyon, R. P., & Haber, A. (1976). *Fundamentals of behavioral statistics* (3rd ed.). Reading, MA: Addison-Wesley.

Snedecor, G. W., & Cochran, W. G. (1989). *Statistical methods* (8th ed.). Ames: Iowa State University Press.

Bayesian Analysis

Bayes's theorem is often used in decision analysis, so it would be natural to think that *Bayesian analysis* is a generic term to describe decision analyses using the Bayes's theorem. On the contrary, Bayesian analysis refers to a school of thought in statistical analysis. It differs both operationally and conceptually from the two other traditional ways of carrying out statistical analysis: frequentist and likelihood based. Statisticians who adhere to the principles of Bayesian analysis sometimes call themselves Bayesians.

The goal of most statistical analysis is to make inferences about population parameters. These parameters are not observable directly but can be estimated using data. For example, incidence of a particular disease in a given country is a parameter. It is practically impossible to find the true incidence, but it is quite possible to estimate it based on an appropriately chosen sample. In traditional statistical analysis only the information in the sample will be used for the purpose of estimation. In Bayesian analysis, information external to the sample, such as prior related findings, expert information, or even subjective beliefs can be incorporated into the analysis. Results of a Bayesian analysis will reflect a weighted combination of the

information in the sample and the prior information. These weights are intrinsically chosen by the analyst based on the study design (especially sample size) and the precision of prior information.

Example

A simple example might help clarify the concepts and the process. Suppose we want to estimate the mean age of patients seen at a pediatric emergency care facility. Based on our knowledge about the patient profile of this particular institution, we expect the mean to be around 7. We think it could be as low as 5 or as high as 9. We can represent our prior information about the mean age in the form of a normal distribution with mean 7 and standard deviation 1. This means that, a priori, the probability that the mean age is below 5 or above 9 is approximately 5%. The data we collect on a particular day based on 10 consecutive admissions are 7, 6, 8, 12, 15, 10, 4, 8, 11, 9 (sample mean of 9). How can we reconcile our prior information with the observed data?

If we let μ denote the mean age (not the sample mean, but the population mean), and the prior information on μ with $\pi(\mu)$, then $\pi(\mu)$ is a normal distribution with mean 7 and variance 1, to be denoted by $N(7, 1)$. We are assuming that X (the observations) also follows a normal distribution $N(\mu, \sigma^2)$, where σ^2 is the (population) variance of age. Call this distribution $L(X|\mu, \sigma^2)$. For the time being, let us assume that we know that $\sigma^2 = 10$; we will comment later on how to handle the more realistic case of unknown variance. We can now use Bayes's theorem to find the distribution of μ, given X:

$$P(\mu|X) = \frac{\pi(\mu)L(X|\mu, \sigma^2 = 10)}{\int \pi(\mu)L(X|\mu, \sigma^2 = 10)\, d\mu}.$$

This is called the *posterior distribution* of μ (contrast with prior distribution). The fundamental premise of Bayesian analysis is that the posterior distribution contains all the available information about μ and hence should form the basis of all statistical inference.

The analytical evaluation of the integral in the denominator is tedious but possible. It turns out that $P(\mu|X)$ also follows a normal distribution with mean

$$m_p = \frac{(m_\pi/\sigma_\pi^2) + (nx/\sigma^2)}{(1/\sigma_\pi^2) + (n/\sigma^2)},$$

and variance

$$\sigma_p^2 = \frac{1}{(1/\sigma_\pi^2) + (n/\sigma^2)}.$$

Here, n is the sample size, m_π and σ_π^2 are the prior mean and variance, and m_p and σ_p^2 are the posterior mean and variance, Substituting the values from the example, we have

$$m_p = \frac{(7/1) + (10 * 9/10)}{(1/1) + (10/10)} = \frac{16}{2} = 8;$$

$$\sigma_p^2 = \frac{1}{(1/\sigma_\pi^2) + (n/\sigma^2)} = \frac{1}{(1/1) + (10/10)} = .5.$$

By going through the calculations we see that the posterior distribution is $N(8, .5)$. A sensible estimate of the mean patient age in this facility, then, is 8. Notice how the sample mean of 9 is shrunk toward the prior mean 7. In fact, the equation for the posterior mean above can be seen to be a weighted average of prior and sample means, where the weights are inversely proportional to the variances. Since the variance of the sample mean (σ^2/n) decreases with the sample size, increasing the sample size will make the posterior mean closer to the sample mean. For example, if the sample size was 100 with the same mean (9) and variance (10), the posterior mean would be 8.8.

Figure 1 displays the three distributions at work for this example. The prior is shifted right to become the posterior because the bulk of the data lies to the right of the prior. But the variance of the posterior is largely determined by the prior. This suggests a weakness in this analysis, namely, that we were too confident in our prior information to begin with. We were absolutely sure that the mean age would be between 4 and 10, since the prior we chose places negligible mass of probability outside this range. Yet four of our data points were greater than 10. This could of course be due to pure chance, but there could be other reasons. Perhaps the sample was not representative. Since it reflects the experience on a single afternoon, it might have been biased by outside factors that we are not aware of (such as a soccer tournament of 10-plus-year-olds

held nearby). Or we might have judged our confidence in the prior incorrectly. It behooves a good analyst to investigate this further.

We can also form a confidence interval based on this distribution. For example, μ will be within the interval ($\mu_p \pm 2\sigma_p$) approximately 95% of the time, and hence this defines a 95% confidence interval for μ. Confidence intervals are usually called *posterior intervals* if they are calculated from a Bayesian perspective. In this case, the 95% posterior interval for μ is (6.61, 9.39). It is quite wide because it is based on 10 samples only. For purposes of comparison, the standard 95% confidence interval can be calculated as (2.8, 15.3). The Bayesian interval is narrower due to the contributions from the prior information as well as the data. Remember the discussion from the above paragraph suggesting that the prior was perhaps too precise (it had a small variance). If the prior had the same mean but a larger variance, for example, 10, then the posterior distribution would be $N(8.8, 0.9)$ and the 95% posterior interval would be (6.9, 10.7). Notice how the posterior mean shifted closer to the sample mean. This is because the weight that the prior mean received is much smaller now because the prior variance is higher.

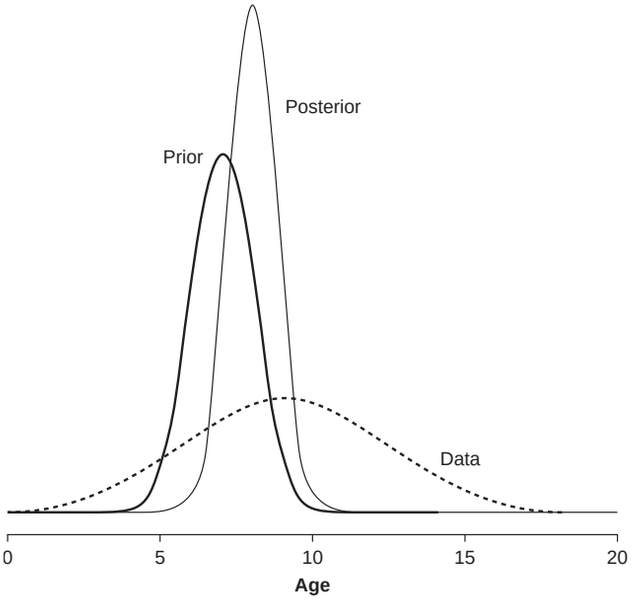


Figure 1 Prior, data, and posterior distributions

Note: Prior (--), data (..), and posterior (__) distributions for the pediatric emergency room example.

Finally, we can compute the posterior probabilities of a hypotheses. If we wanted to test, for example, the hypothesis that H_0: $\mu > 10$, we could simply compute $P(\mu > 10|X)$, that is, the posterior probability that μ is greater than 10. Based on the $N(8, 0.5)$ posterior, this turns out to be .002. Since this probability is very small, we can safely conclude that μ is less than 10. Note that the small prior variance is influencing this inference in the same way in which it influenced the posterior intervals. If the prior variance had been 10, then the posterior probability of this hypothesis would have increased to .103, and it would have been likely for the analyst to conclude that there was not sufficient information to reject the hypothesis that $\mu > 10$ (although it is still less likely than the alternative $\mu \leq 10$). The practice of varying the prior parameters and observing the effects on the posterior is known as sensitivity analysis and is further discussed below in more general terms.

The interpretation of Bayesian findings is very different from the interpretation of traditional statistical results. For example, with a posterior interval, we can conclude that μ lies within the posterior interval with 95% probability. With a confidence interval, we have to resort to the frequentist interpretation that 95% of the intervals constructed in this manner will contain the true μ. Similarly, with hypothesis testing, we can directly conclude that the probability of the hypothesis is low or high. With a *p* value, however, the interpretation is more cumbersome: If the null hypothesis is true, then the probability of observing a result at least as extreme as what is observed is the *p* value. Most people find the Bayesian interpretations more palatable.

Generalization to Multiple Parameters

The ideas in this simple example can be generalized to any statistical model with an arbitrary number of parameters. If we let θ represent the set of parameters (θ will be a vector) we are interested in, and *X* denotes the observations in our sample, the same version of Bayes's theorem holds:

$$P(\theta|X) = \frac{\pi(\theta)L(X|\theta)}{\int \pi(\theta)L(X|\theta)\,d\theta}.$$

This time, both π and *L* will be multivariate distributions. In the example above, if we relax the assumption of known σ^2, despite the fact that we

will have two parameters, we can apply the same principles to arrive at the posterior distribution.

If both π (the prior) and P (the posterior) are from the same family, then we have a *conjugate* family for L. For example, normal distribution is the conjugate for itself (i.e., if π and L are normal, then P will also be normal). A conjugate family makes calculations easier, and its use deserves serious consideration, if it exists. The problem is that for most problems of practical importance, such as regression and analysis of variance, there are no conjugate families. Since the integral in the denominator is often intractable, application of Bayesian methods in practice used to be very limited. The emergence of particular numerical methods that collectively came to be known as Markov chain Monte Carlo (MCMC) enabled statisticians to generate a sample from P without explicitly deriving an equation for it. Since this sample has distribution P, it can be used to mimic the properties of P. For example, one can form a 95% posterior interval from this sample by truncating it at the 2.5th and 97.5th percentiles. In other words, MCMC methods enable an analyst to bypass the integration in the denominator of the Bayes's theorem above. This has fueled an explosion of Bayesian applications, including the ones with very large numbers of parameters that cannot be solved within the regular statistical paradigm.

Objections to Bayesian analysis often include the difficulty of choosing a prior distribution. While most people will agree that some prior information exists in most real-world problems, they rarely agree on how to formulate it in the form of a probability distribution. It is ostensibly true that one can replace the prior of the analyst with another one and repeat the analysis to arrive at his or her own conclusions. This is rarely done, however, leaving Bayesian analysts with the responsibility of choosing a π that will be acceptable to most analysts (in addition to choosing such an L, which is the responsibility of most other statisticians as well) or performing an extensive sensitivity analysis with the hope that wide variations in π will not result in wide variations in conclusions. In the example above, we performed an informal (and highly incomplete) sensitivity analysis by changing the variance of the prior from 1 to 10 and recomputing the posterior distribution. In most cases, sensitivity analysis will define a range within which the inferences are robust to prior specifications, but this range can be unacceptably narrow.

Defenders of Bayesian analysis point to the flexibility it brings as well as the formal incorporation of external information. It is often argued that most scientists are implicit Bayesians, evaluating others' findings in light of their subjective outlook—something that can be explicitly done in the Bayesian framework. Another advantage is that, since θ is random, one can make probability statements about θ and the interpretation of Bayesian intervals and tests is very straightforward. In contrast, most nonstatisticians struggle with the appropriate definition of frequentist statistical results such as p values.

Mithat Gönen

See also Bayes's Theorem; Confidence Intervals; Likelihood Ratio

Further Readings

A Bayesian goes shopping [Editorial]. (1992). *Medical Decision Making, 12,* 1.

Berry, D. (1996). *Statistics: A Bayesian perspective.* Belmont, CA: Duxbury Press.

Berry, D. A. (2006, September). Bayesian statistics. *Medical Decision Making, 26,* 429–430.

Carlin, B., & Louis, T. (2000). *Bayes and empirical Bayes methods for data analysis* (2nd ed.). Boca Raton, FL: Chapman & Hall.

Efron, B. (1986). Why isn't everyone a Bayesian? *American Statistician, 40*(1), 1–5.

Gelman, A., Carlin, J., Stern, H., & Rubin, D. (2004). *Bayesian data analysis* (2nd ed.). Boca Raton, FL: Chapman & Hall.

Gilks, W. R., Richardson, S., & Spiegelhalter, D. J. (Eds.). (1996). *Markov chain Monte Carlo in practice.* London: Chapman & Hall.

Lee, P. (2004). *Bayesian statistics: An introduction.* London: Arnold.

Parmigiani, G. (2002). *Modeling in medical decision making: A Bayesian approach.* Chichester, UK: Wiley.

Bayesian Evidence Synthesis

Evidence synthesis has come to replace *meta-analysis* as a term referring to the statistical

combination of multiple sources of evidence. In its simplest form, each evidence source is represented by a sufficient statistic, which may be, for example, a numerator and a denominator, a mean with its standard error, or a summary estimate such as an estimate of the log odds ratio and its standard error. The evidence synthesis is then the process of finding a suitable weighted average of these quantities. However, meta-analysis need not be restricted to summary statistics and has, for example, been extended to analyses of data from individual patients in multiple studies. Furthermore, evidence synthesis may be used to imply much more general forms of statistical synthesis, involving data sources of multiple types, each perhaps providing information on one or more parameters.

Bayesian evidence synthesis is then the use of Bayesian statistical methods in evidence synthesis. This can be formulated as follows: There are K unknown *basic* parameters, θ, and N data points Y_i, $i = 1,\ldots, N$, each representing, let us assume, a sufficient statistic from study i, in this case consisting of numerators r_i and denominators n_i. We may also define additional *functional* parameters $\theta_{K+1},\ldots, \theta_M$. To rule out recursive definitions, it must be possible to define these as functions $G_{K+1},\ldots, G_M$ of the basic parameters. Finally, each data point provides an estimate of a $G_i(\theta)$ that is some function of parameters.

One approach to computation, the maximum likelihood solution, assuming that the N data points are independent, would be to find values of θ that maximize

$$\boldsymbol{L} = \Pi_{i = 1,\ldots, N}\, \mathrm{L}_i\, (Y_i|\theta_1, \theta_2,\ldots, \theta_K), \qquad (1)$$

bearing in mind that the likelihood contribution from each study might take a different distributional form (normal, binomial, Poisson, etc.).

Bayesian evidence synthesis specifies a prior distribution for the basic parameters only $P(\theta)$. There is no requirement that the parameters be independent, so we may consider this to be a joint prior distribution, if necessary. We then find the joint posterior distribution by application of Bayes's theorem:

$$P(\theta_1,\ldots, \theta_K|Y_1,\ldots, Y_N) \propto P(\theta)\, \boldsymbol{L} \qquad (2)$$

The data to be synthesized form a connected network that can be described in terms of a directed acyclic graph (DAG). However, the synthesis problems are capable of being reparameterized in many different ways, so that items of data that inform, for example, a basic parameter in one parameterization may inform a functional parameter in another. Hence, several DAGs may describe the same network. Some evidence networks may also be described in terms of graphs. One important feature that remains invariant under reparameterization is the inconsistency degrees of freedom, $N - K$. This can be thought of as representing the number of independent ways in which the evidence can be inconsistent under a given model. For example, in the DAG presented in Figure 1 and discussed in more detail below, there are three basic parameters and four independent data items to inform them. The inconsistency degrees of freedom is therefore $4 - 3 = 1$. The Bayesian formulation, which forces the investigator to be explicit about which parameters are basic, and therefore have a prior distribution, and which are functional, yields valuable insights into the structure and dynamics of the data and model.

The Bayesian framework is, of course, essentially the same in evidence synthesis as it is in other areas of statistics, but it takes a slightly different flavor in this context. Instead of combining a "prior" based, formally or informally, on the accumulated evidence so far, together with the likelihood in the latest study, the whole exercise is concerned with combining all the available evidence. For this reason, the priors put on most parameters are typically vague. Nevertheless, the focus on putting together all available evidence, to obtain the best possible estimates with the most realistic assessment of uncertainty, is very much in tune with the Bayesian spirit.

History

The origins and development of Bayesian evidence synthesis lie more in decision making than in traditional statistical inference. On the other hand, Bayesian methods have brought to decision modeling the advantages of formal posterior inference, and of statistical methods for model diagnosis, that have otherwise tended to be lacking. The decision-making context is inevitably associated

with multiple parameters and multiple sources of uncertainty. Probabilistic methods were introduced in the 1980s at the time when the development of computers made it possible to evaluate complex models by Monte Carlo simulation. Each parameter is represented by a statistical distribution, one Monte Carlo cycle value is drawn from each distribution, and the costs and benefits are computed. The expected costs and benefits are then taken as an average over the simulated sequence. This scheme has been regarded as essentially Bayesian in that it focuses on the probability distributions of parameters. Essentially, it samples from a (informative) prior $P(\theta)$ but unlike Equation 2 does not update this with further data.

Forward simulation from a prior is, however, severely limited. Typically, each parameter represented by a separate distribution, as well as each distribution, is informed by a separate item from the data, either from a single study or from a meta-analysis. This means that the number of data sources must equal the number of parameters, no more and no less. A scheme with Bayesian updating, in contrast, can incorporate multiple functions of parameters so that, if the data are available, there can be more sources of data than there are parameters. Furthermore, Bayesian hierarchical models can be deployed to share information over the parameter space and thus to manage situations where there are fewer data items than there are parameters.

A related advantage of full Bayesian updating is that the ability to incorporate data on more functions of parameters than there are parameters represents an opportunity to validate the model. This can also be regarded as a form of probabilistic model calibration. Model diagnostics are available to check the consistency of the different sources of evidence with respect to any parameter.

The above formulation of the Bayesian evidence synthesis model is due to David M. Eddy and his colleagues, whose 1992 book, *Meta-Analysis by the Confidence Profile Method,* appears to have been the first systematic exposition of Bayesian evidence synthesis in the context of medical decision making. Although the book introduced extremely powerful statistical methods and ideas to a wider audience, it failed to have the impact it deserved. This was due, perhaps, to the somewhat stylized examples and the specialized software required.

In fact, Bayesian forms of statistical synthesis seem to have emerged independently in related fields, in each case based on different computational approaches. The Confidence Profile Method was based on a fully Bayesian computation using Monte Carlo simulation and on two further approximate methods that are not always accurate for small sample sizes. Another set of computational methods that have been used for synthesis, named Bayesian Monte Carlo, is based on weighted Monte Carlo sampling, where the weights were given by the likelihood of the data at each set of parameters. This approach became popular in the Environmental Health Risk Assessment field, beginning with simple accept-reject algorithms and then evolving to a fully Bayesian approach. Typical applications included updating prior distributions for contaminant release, environmental transport, and biological effects, with field data on pollution levels.

A further series of methods, called variously Bayesian Synthesis, Bayesian Melding, and Bayesian Pooling, were developed for deterministic models of animal and plant populations. These algorithms are also based on various types of noniterative reweighting schemes.

Bayesian Markov chain Monte Carlo (MCMC) has become the standard software for Bayesian evidence synthesis, certainly in the medical decision-making context. The development of freely available user-friendly MCMC software, such as WinBUGS, has opened the possibilities of Bayesian evidence synthesis to a wide range of researchers. Users of the package need only specify priors and likelihoods and define functional parameters. As a result, a wide and increasing range of applications is appearing in health technology assessment literature. The expression *multiparameter evidence synthesis,* coined by Victor Hasselblad, another founder of the Confidence Profile Method, is often used for applications of this sort.

Many of these are examples of *comprehensive decision analysis,* a term used when a Bayesian statistical synthesis is embedded within a decision analysis, an approach first seen in publications from Duke University in the 1990s. More recently, the increasing adoption of net benefit analysis has made this conceptually appealing, and the simulation format for MCMC, of course, fits in readily with the simulation approach that has become

familiar from probabilistic modeling based on simple Monte Carlo methods.

Examples

This section outlines some examples of Bayesian evidence synthesis. Figure 1 shows a fragment of a model of HIV epidemiology in the form of an influence diagram, in which four sources of data inform three basic probability parameters. The basic parameters would most naturally be given Beta prior distributions. The three surveys that directly inform the basic parameters provide prevalence data and would, therefore, each contribute a binomial likelihood. The fourth source of evidence would provide the observed number of diagnosed cases, which would be represented as a Poisson distribution. This type of evidence structure is quite common in epidemiology applications. Note that the model in effect "calibrates" the basic parameters so that their product is consistent with the routine surveillance data.

As noted earlier, the inconsistency degrees of freedom is 1. Therefore, the investigator can assess whether or not the four sources of data are consistent with each other, under this model. If they are not, this would suggest that one or more of the studies is not estimating the presumed target parameter but is biased. *Which* study is biased, of course, cannot be determined without further data or expert judgment, and very possibly more than one is biased.

Another very common structure for Bayesian evidence synthesis is illustrated in Figure 2. Mixed

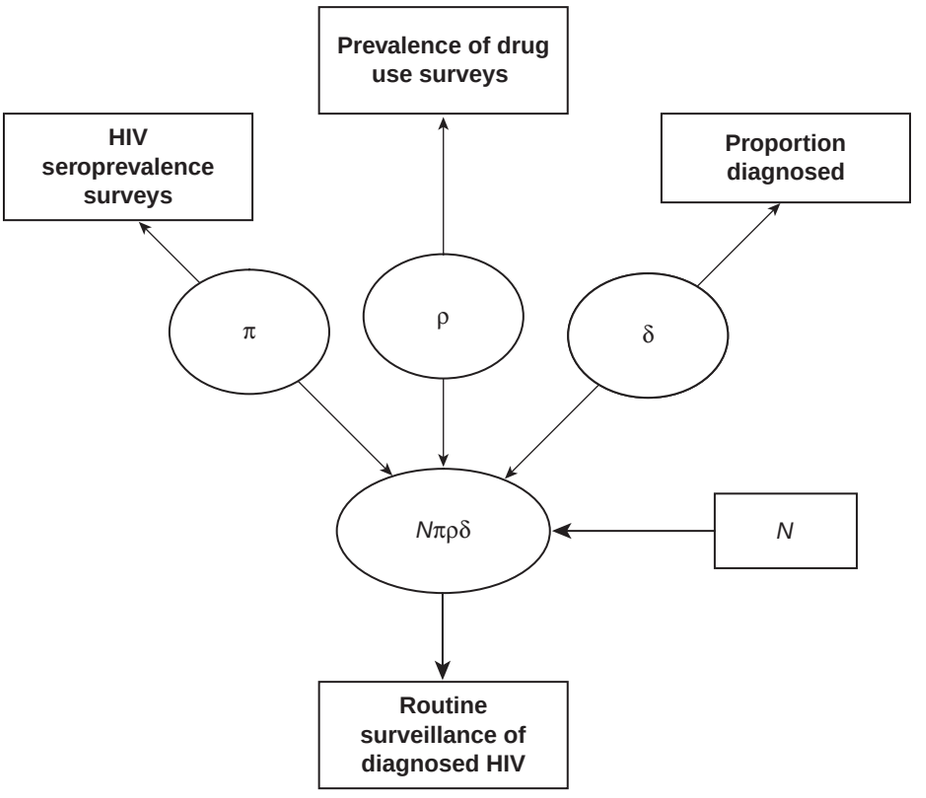


Figure 1 Schematic influence diagram (directed acyclic graph)

Note: Ellipses indicate stochastic nodes, rectangles constants or data, and edges the direction of influence. There are three *basic* parameters: HIV prevalence in injecting drug users (IDUs), π; proportion of the population who are IDUs, ρ; proportion of infected IDUs who are diagnosed, δ. With population size *N*, a constant, these define the product *N*πρδ, a functional parameter, which is the number of diagnosed IDUs. There are four sources of data: Three directly inform the basic parameters; the fourth directly informs a functional parameter.

Treatment Comparison structures allow the synthesis of data from pairwise, or multiarm, randomized controlled trials of treatments. For example, the evidence base may consist of one or more trials making each of the following comparisons: *A* versus *B*, *A* versus *C*, *B* versus *C*, *A* versus *B* versus *D*, *C* versus *D*, and so on. If we arbitrarily choose *A* as the reference point, we may choose the relative treatment effects of *B*, *C*, *D*, and so on, relative to A as the basic parameters. All the other contrasts, d_{XY}, may then be expressed in terms of the basic parameters:

$$d_{XY} = d_{AY} - d_{AX} \quad (3)$$

For example, if we take Streptokinase (SK) as reference treatment, then we can express the relative efficacy of percutaneous transluminal angioplasty (PCTA) relative to accelerated t-PA (At-PA) as follows:

$$d_{\text{PCTA, At-PA}} = d_{\text{SK, PCTA}} - d_{\text{SK, At-PA}} \quad (4)$$

The key assumption being made of the data, of course, is that each of the randomized controlled trials included would, if all the treatments had been included, be providing estimates of the same relative effect parameters. If there are T treatments and information on N pairwise contrasts, then there are $(T - 1)$ basic parameters and the inconsistency degrees of freedom is $(N - T + 1)$. Equations 3 and 4 effectively reduce the parameter space from N unrelated comparisons to $(T - 1)$. In this case, we have $T = 7$ treatments and evidence on $N = 10$ contrasts, giving 4 degrees of freedom for inconsistency, though where multiarm trials are involved, this simple formula requires adjustment.

Bayesian methods have been used to synthesize many other evidence structures. These include, for example, collapsed frequency tables; regression models based on different subsets of variables; surrogate or intermediate endpoints in trials with clinical endpoints; multiple outcomes, or the same outcome reported at multiple time points in clinical trials; Markov rate models; and individual and aggregate data. Collapsed category methods are becoming increasingly common in genetic epidemiology.

Bayesian evidence synthesis in cost-effectiveness analysis is often associated with expected value of information analysis. Because multiple parameters are estimated from a common data set, their posterior distributions are invariably correlated. This can introduce additional complexity in expected value of information calculations.

A E Ades

See also Cost-Effectiveness Analysis; Expected Value of Perfect Information; Meta-Analysis and Literature Review; Net Benefit Regression

Further Readings

Ades, A. E., & Sutton, A. J. (2006). Multiple parameter evidence synthesis in epidemiology and medical decision making: Current approaches. *Journal of the Royal Statistical Society, Series A, 169,* 5–35.

Brand, K. P., & Small, M. J. (1995). Updating uncertainty in an integrated risk assessment: Conceptual framework and methods. *Risk Analysis, 15,* 719–731.

Dominici, F., Parmigiani, G., Wolpert, R. L., & Hasselblad, V. (1999). Meta-analysis of migraine headache treatments: Combining information from heterogenous designs. *Journal of the American Statistical Association, 94,* 16–28.

Eddy, D. M., Hasselblad, V., & Shachter, R. (1992). *Meta-analysis by the Confidence Profile Method: The statistical synthesis of evidence.* Boston: Academic Press.

Goubar, A., Ades, A. E., DeAngelis, D., McGarrigle, C. A., Mercer, C., Tookey, P., et al. (2008). Estimates of HIV prevalence and proportion diagnosed based on Bayesian multi-parameter synthesis of surveillance data (with discussion). *Journal of the Royal Statistical Society, Series A, 171,* 541–580.

Lu, G., & Ades, A. E. (2006). Assessing evidence consistency in mixed treatment comparisons. *Journal of the American Statistical Association, 101,* 447–459.

Raftery, A. E., Givens, G. H., & Zeh, J. E. (1995). Inference for a deterministic population dynamics model for Bowhead whales (with discussion). *Journal of the American Statistical Association, 90,* 402–430.

Spiegelhalter, D. J., Abrams, K. R., & Myles, J. P. (2004). *Bayesian approaches to clinical trials and health-care evaluation.* Chichester, UK: Wiley.

Bayesian Networks

A Bayesian network is a graphical representation of a multivariate probability distribution on a set

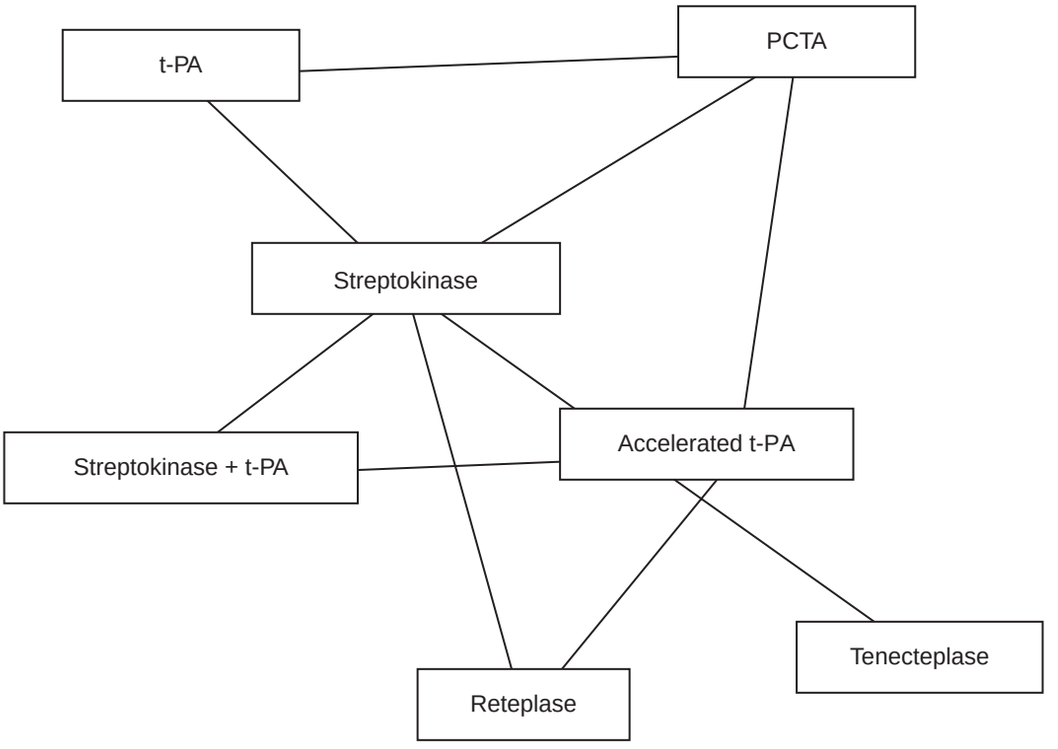


Figure 2 A Mixed Treatment Comparison network involving six thrombolytic treatments following acute myocardial infarction and one surgical treatment, percutaneous transluminal angioplasty (PCTA)

Note: Each edge indicates that the treatments have been compared in at least one randomized, controlled trial.

of discrete random variables. Representational efficiency is achieved by explicit separation of information about *conditional independence* relations between the variables (coded in the network structure) and information about the probabilities involved (coded as a set of numeric parameters or functions). The network structure is expressed as a directed acyclic graph (DAG) that makes the representation amenable to an intuitively appealing, causal interpretation. Algorithms exist for learning both network structure and parameters from data. Furthermore, Bayesian networks allow for computing any marginal or *conditional probability* regarding the variables involved, thus offering a powerful framework for reasoning with uncertainty. Bayesian networks are also called *belief networks* and *causal probabilistic networks*.

Bayesian networks are suited to model the uncertainty that inheres in many biomedical domains and are, therefore, frequently used in applications of computer-assisted decision making in biomedicine. Furthermore, extensions of Bayesian networks (called *influence diagrams*) can be used to perform decision analyses.

This entry first sketches the historical background of Bayesian networks. Subsequently, it elaborates on model structure, approaches for network construction, inference methods, medical applications, and software.

Historical Background

Bayesian networks originated in the mid-1980s from the quest for mathematically sound and computationally tractable methods for reasoning with uncertainty in artificial intelligence. In the preceding decade, the first applications of computer-assisted decision making had found their way to the medical field, mostly focusing on the diagnostic process. This had required the development of methods for reasoning with uncertain and incomplete diagnostic information.

One popular method was the naive Bayesian approach that required specification of positive and negative predictive values for each of a set of predefined diagnostic tests and a prior (i.e., marginal) probability distribution over possible diagnostic hypotheses. The approach assumed that all test results were mutually independent markers of disease and used Bayes's theorem to compute posterior (i.e., conditional) probabilities on the hypotheses of interest. The approach is simple and fast and requires a relatively small number of marginal and conditional probabilities to be specified. However, the assumption of independence is mostly wrong and leads to overly extreme posterior probabilities.

Another approach arose in the field of expert systems, where algorithms had been devised to reason with so-called certainty factors, parameters expressing the strength of association in if-then rules. The underlying reasoning principles were mostly ad hoc and not rooted in probability theory, but large sets of if-then rules allowed for a domain representation that was structurally richer and more complex than naive Bayesian models. Bayesian networks bring together the best of both approaches by combining representational expressiveness with mathematical rigor.

Model Structure

Bayesian networks belong to the family of probabilistic graphical models (PGMs), graphs in which nodes represent random variables, and the (lack of) arcs represent conditional independence assumptions. Let $G = (V(G), A(G))$ be a directed acyclic graph, where the nodes $V(G) = \{V_1, \ldots, V_n\}$ represent discrete random variables with a finite value domain. For each node $V_i \in V(G)$, let π_i denote the set of parent nodes of V_i in graph G. A Bayesian network now is a pair $B = (G, \Theta)$, where $\Theta = \{\theta_i | V_i \in V(G)\}$ is a set of parametrization functions. The function θ_i describes a local model for node $V_i \in V(G)$ by specifying a conditional probability $\theta_i(v|s)$ for each possible value v of variable V_i and all possible value assignments s to its parents π_i. The Bayesian network B defines a unique multivariate probability distribution Pr on $V_1, \ldots, V_n$ using the factorization

$$\Pr(V_1, \ldots, V_n) = \prod_{i=1}^{n} \theta_i(V_i|\pi_i).$$

An example Bayesian network is shown in Figure 1. This network has eight variables and is a simplified representation of diagnosing a patient presenting to a chest clinic, having just come back from a trip to Asia and showing dyspnea. This symptom may be caused by tuberculosis, lung cancer, or bronchitis. In this example, the local model for the variable "dyspnea" specifies that there is a .80 probability that dyspnea is present when the patient has bronchitis but no tuberculosis or lung cancer and a .70 probability of dsypnea when the patient does have tuberculosis or cancer but no bronchitis.

It follows from the definition of Bayesian networks that each variable is conditionally independent of its nondescendants in the graph given its parents; this is called the *local Markov condition*. It induces a more general notion of conditional independence, the *global Markov condition*, which builds on the graphical criterion of *path blocking*. Let X, Y, and Z be nonintersecting sets of nodes in $V(G)$, and consider an arbitrary path from a node in X to a node in Y. The path is blocked by the set Z if it includes a node such that either (a) the arrows on the path meet head-to-tail or tail-to-tail at the node, and the node is in the set Z, or (b) the arrows meet head-to-head at the node, and neither the node, nor any of its descendants, is in the set Z. For example, the set {smoking, bronchitis} blocks the path lung cancer—smoking—bronchitis—dyspnea. Sets X and Y are conditionally independent given Z in probability distribution Pr if each path from a node in X to a node in Y in graph G is blocked by Z. In words, this means that once Z has been observed, knowing X will not influence our beliefs about Y and vice versa.

A Bayesian network represents the conditional independence relations between a set of variables,

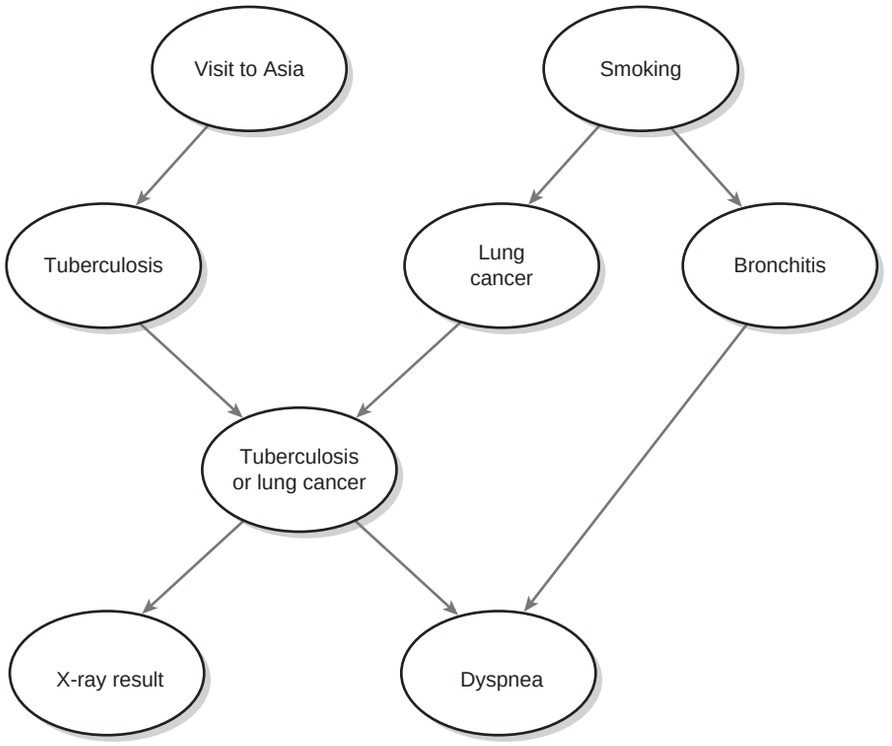


Figure 1 Example Bayesian network for diagnosing dyspnea after a visit to Asia

Source: From Table 2 (p. 164) of Lauritzen, S. L., & Spiegelhalter, D. J. (1988). Local computations with probabilities on graphical structures and their application to expert systems (with discussion). *Journal of the Royal Statistical Society, Series B, 50*(2), 157–224. Reprinted with permission of Wiley-Blackwell.

not their causal dependencies. Under certain conditions, however, one can assume that the arcs in a Bayesian network permit a causal interpretation. The crucial observation is that the occurrence of an uncertain event is independent of its noneffects, given its direct causes. Therefore, if a directed graph accurately depicts causality, it will also obey the local Markov condition, and one can use it as the graphical part of a Bayesian network. In such cases, one can basically think of conditional independence relations as byproducts of causality. This applies to the example network from Figure 1.

In the most basic representation, the parametrization function θ_i of each node $V_i \in V(G)$ is simply stored as a contingency table. The size of such a table, however, grows exponentially in the number of parents of V_i in the graph. There exist various ways of reducing the size of the representation. Popular examples are the noisy OR gate, which assumes that the influence of each parent on V_i is independent of other parents, and local tree structures for representing the table.

Construction

Two different approaches to developing Bayesian networks can be distinguished. The first one is manual construction in collaboration with domain experts and was frequently applied in early medical applications of Bayesian networks. The second approach learns the network from data; this approach has become more feasible in the medical field with the large amounts of patient data that are currently recorded in information systems and is, consequently, being applied in more recent medical applications.

Manual construction of Bayesian networks involves the use of knowledge engineering techniques

and, thus, resembles the manual construction of knowledge bases and decision models such as decision trees. In the development process, a number of stages can be distinguished that are iterated, inducing further refinement of the network under construction. The first stage is the selection of relevant variables that form the nodes in the network. Variable selection is generally based on expert opinion (interviews) and descriptions of the domain. Subsequently, dependency relationships among the variables are identified and added as arcs in the network. For this purpose, the notion of causality is generally employed in interviews with domain experts by asking for the causes and consequences of manifestations. In the third stage, qualitative probabilistic constraints (e.g., the probability of an adverse outcome with severe comorbodities is at least as high as with moderate comorbidities) and logical constraints (e.g., the occurrence of pregnancy is limited to females) among the variables are identified. These constraints are also helpful in the next stage of assessing the local (conditional) probability distributions for each variable and in their verification. Elicitation methods that originate from the field of medical decision making (e.g., for translation of verbal expressions of probabilities to numbers) can be used for deriving subjective probabilities from domain experts.

In the second approach of constructing Bayesian networks, both the graphical structure and the conditional probability distributions are learned from data. As an exhaustive search through the space of all possible network structures (DAGs) is computationally prohibitive, most algorithms apply a heuristic search strategy, starting with a single structure (either an empty or a randomly chosen graph) and incrementally modifying this structure until a termination condition is reached. There are two main types of algorithms. The first type of algorithm evaluates the current network structure and its closest resemblers using a goodness-of-fit scoring function and continues with the structure having the highest score. The second type of algorithm employs statistical independency tests on the data to determine, for each pair of variables, whether an arc should be added between them in the graph. After the network structure has been established, the parametrization functions are estimated from the data using maximum likelihood estimation; the expectation maximization (EM) algorithm can be used in case of incomplete data. In addition to network learning in a frequentist approach, Bayesian statistical methods can be used in which prior probabilities are assigned to the network structure and the probability distributions.

In practice, often mixtures of the above approaches are used for construction of Bayesian networks, for example, by employing data to estimate or update conditional probability estimates in an otherwise manual construction process or by placing constraints on the network structure based on domain knowledge before inducing a network from data. In each approach, construction of Bayesian networks is completed with the evaluation of the performance of the network constructed, preferably on independent data.

Inference

Given a Bayesian network $B = (G, \Theta)$, we can identify a number of probabilistic inference tasks. Let Pr denote the multivariate probability distribution that is defined by B, let e denote evidence (i.e., observed states) on a subset $V' \subset V(G)$ of network variables, and let $V'' \subset V(G) \backslash V'$ be a subset of variables of interest. Inference tasks fall into two categories.

Evidence Propagation: Given evidence e, what is conditional probability distribution on the set V''? Special cases are computation of the conditional probability $\Pr(h|e)$ of a particular state h of the set V'' and of the marginal (i.e., unconditional) probability $\Pr(e)$.

Maximum a Posteriori (MAP) Assignment: Given evidence e, what is the most likely state h of V''; that is, $h = \text{argmax}_s \{\Pr(s, e)\}$, where s ranges over all possible states of V''? The special case of MAP assignment, where V'' consists of all network variables except those in V', that is, $V'' = V(G) \backslash V'$, is called *most probable explanation*.

Generally speaking, both inference categories become more complicated when V'' gets larger because the associated set of possible states grows exponentially in size. The main difference between the categories is that evidence propagation infers probabilities for given states, while MAP assignment infers a state. In both cases, it does not matter

whether the evidence variables V' are located above or below the variables of interest V'', because information can travel both in the direction of network arcs and against it. The former situation corresponds to causal (or predictive) reasoning, the latter situation to diagnostic reasoning.

Both inference categories constitute challenging computational problems. Much research effort has therefore been devoted to designing efficient inference methods. It is common to distinguish between *exact* and *approximate* inference. The most popular method for exact inference, the join tree algorithm, converts the Bayesian network into a tree structure in which each junction represents a cluster of network variables. Evidence propagation proceeds by applying a message-passing scheme to the join tree. The join tree representation may take a long time to construct and become very large when the network is densely connected. It need only be constructed once, though, and the message-passing phase is fast. Other exact inference methods are variable elimination and recursive conditioning.

Approximate inference methods can be used when exact inference methods lead to unacceptable computation times because the network is very large or densely connected. Popular approaches are simulation methods and variational methods. Simulation methods use the network to generate samples from the conditional probability distribution $\Pr(V''|e)$ and estimate conditional probabilities of interest when the number of samples is sufficiently large. Variational methods express the inference task as a numerical optimization problem and then find upper and lower bounds of the probabilities of interest by solving a simplified version of this optimization problem.

Medical Applications and Examples

In a medical context, Bayesian networks are mainly developed to support three types of problem solving: (1) diagnostic reasoning, (2) prognostic reasoning, and (3) therapy selection. Bayesian networks form a suitable formalism for modeling the uncertainties in diagnostic tests due to false-positive and false-negative findings and enable the computation of the conditional probability $\Pr(h|e)$ of a diagnostic hypothesis h given the evidence of diagnostic test results e by evidence propagation. Early examples of diagnostic applications of Bayesian networks are the MUNIN system for diagnosis of peripheral muscle and nerve diseases, the Pathfinder system for diagnosis of lymph node diseases, and a network for diagnosis in internal medicine and neurology, a reformulation of the rule-based expert system, INTERNIST-1/QMR. Diagnostic Bayesian networks are often equipped with methods for determining the optimal order of diagnostic tests for reducing the uncertainty in a patient's differential diagnosis.

Prognostic Bayesian networks have a pronounced temporal structure with the outcome variable as final node in the network and pretreatment variables and treatment variable as its ancestor nodes. With MAP assignment, most probable prognostic scenarios can be determined using the network. In the literature, a relatively small number of prognostic applications of Bayesian network have been described, including applications for non-Hodgkin lymphoma, for malignant skin melanoma, and for cardiac surgery.

The medical task of therapy selection involves both diagnostic and prognostic reasoning. Bayesian networks for therapy selection are, therefore, usually extended to decision-theoretic models that include utility functions to guide the choice among different decisions. A suitable extension of Bayesian networks for representing probabilistic knowledge, decisions, and utility information are influence diagrams. Examples of this type of medical application include Bayesian networks for therapy selection for esophageal cancer and for treating infectious diseases in intensive care medicine.

Software

A large number of software packages are available for inference with Bayesian networks, manual construction, and network induction from data. A widely used package is Hugin, which includes algorithms for both inference and network induction (structure and parameter learning). Similar functionality is provided by the BayesiaLab software and the Bayes Net Toolbox that can be used within the Matlab mathematical software package. The Netica software supports parameter learning only. A Web directory of Bayesian network software is available at http://directory

.google.com/Top/Computers/Artificial_Intelligence/Belief_Networks/Software.

Niels Peek and Marion Verduijn

See also Bayes's Theorem; Causal Inference and Diagrams; Computer-Assisted Decision Making; Conditional Independence; Conditional Probability; Diagnostic Process, Making a Diagnosis; Diagnostic Tests; Expert Opinion; Expert Systems; Frequentist Approach; Influence Diagrams; Markov Models; Probability; Probability, Verbal Expressions of; Problem Solving; Subjective Probability

Further Readings

Cowell, R. G., Dawid, A. P., Lauritzen, S. L., & Spiegelhalter, D. J. (1999). *Probabilistic networks and expert systems*. New York: Springer.

Darwiche, A. (2008). Bayesian networks. In F. van Harmelen, V. Lifschitz, & B. Porter (Eds.), *Handbook of knowledge representation* (pp. 467–509). Amsterdam: Elsevier.

Jensen, F. V., & Nielsen, T. D. (2007). *Bayesian networks and decision graphs*. New York: Springer.

Lauritzen, S. L., & Spiegelhalter, D. J. (1988). Local computations with probabilities on graphical structures and their application to expert systems. *Journal of the Royal Statistical Society, Series B, 50,* 157–224.

Lucas, P. J. F., Van der Gaag, L. C., & Abu-Hanna, A. (2004). Bayesian networks in biomedicine and health care. *Artificial Intelligence in Medicine, 30,* 201–214.

Murphy, K. (2005). Software packages for graphical models/Bayesian networks. Retrieved January 23, 2009, from www.cs.ubc.ca/~murphyk/Bayes/bnsoft.html

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems: Networks of plausible inference*. San Francisco: Morgan Kaufmann.

Bayes's Theorem

A Bayesian approach to inference implies combining prior judgment with new information to obtain revised judgment. Prior judgment is expressed in a prior probability that a hypothesis is true. The prior probability is subsequently updated with new data that become available to yield the revised posterior probability of the hypothesis. Bayesian updating can be applied to the results of diagnostic tests (which is explained here), to a research hypothesis under investigation, or to a parameter being estimated in a study.

Bayesian Updating of the Probability of Disease

Estimates of probabilities of disease conditional on diagnostic test results are usually not readily available. One is more likely to have an assessment of the probability of a test result among patients with or without the disease. Converting conditional probabilities of the latter type (test results given disease) to probabilities of the type needed for decision making (disease given test results) should take into account the pretest (or prior) probability of disease, $p(D+)$, the test characteristics (sensitivity and specificity), and the test result (positive or negative) to obtain a posttest (revised or posterior) probability of disease, $p(D+|T+)$ or $p(D+|T-)$. This process is called Bayesian probability revision and can be done using one of several methods.

As an example, consider a 47-year-old female patient who presents with atypical angina in whom you would like to exclude coronary artery disease (CAD). Based on the literature, her pretest (prior) probability of having CAD is 13%. You refer her for a CT to determine her coronary calcium score (CTCS), which is 0 (i.e., a normal/negative test result). Now you wonder whether she still could have CAD in spite of the negative CTCS result. CTCS has a sensitivity of 96% and specificity of 60% for the diagnosis CAD.

Bayesian Probability Revision With a 2 × 2 Table

Given the prior probability of disease $p(D+)$, sensitivity $p(T+|D+)$, and specificity $p(T-|D-)$, we can construct a 2 × 2 table of a hypothetical population and, with the numbers of TP, FN, FP, and TNs, calculate the posttest (revised) probabilities. The steps are as follows:

1. Pick an arbitrary number n for the total hypothetical population (e.g., $n = 10{,}000$).
2. Using the prior probability $p(D+)$, partition the total number of patients across those with and

without the disease, that is, $n(D+) = n \cdot p(D+)$ and $n(D-) = n \cdot (1 - p(D+))$.

3. Using sensitivity $p(T+|D+)$, determine the number of patients with disease who have a true-positive versus a false-negative test result, that is, $TP = n(D+) \cdot p(T+|D+)$ and $FN = n(D+) \cdot (1 - p(T+|D+))$.
4. Using specificity $p(T-|D-)$, determine the number of patients without disease who have a true-negative versus a false-positive test result, that is, $TN = n(D-) \cdot p(T-|D-)$ and $FP = n(D-) \cdot (1 - p(T-|D-))$.
5. Calculate the posttest (revised or posterior) probabilities as follows:
   - Postpositive test probability of disease = $p(D+|T+) = TP/(TP + FP)$.
   - Postpositive test probability of absence of disease = $p(D-|T+) = FP/(TP + FP)$.
   - Postnegative test probability of disease = $p(D+|T-) = FN/(TN + FN)$.
   - Postnegative test probability of absence of disease = $p(D-|T-) = TN/(TN + FN)$.

Note:

Postpositive test probability of disease = positive predictive value.

Postnegative test probability of absence of disease = negative predictive value.

For our example patient, the 2 × 2 table is as follows:

	CAD+	CAD–	
CTCS+	1,248	3,480	4,728
CTCS–	52	5,220	5,272
	1,300	8,700	10,000

The postnegative CTCS probability of CAD = 52/5272 = 1%. In other words, with a 0 calcium score on CT, the likelihood of CAD in this patient is really very low.

Probability Revision Using Bayes's Formula

Consider a test result R, which may be any finding, for example, a positive or negative test result for dichotomous tests or a particular result on a categorical, ordinal, or continuous scale for tests with multiple results. Consider the true disease status D_j, which indicates a particular disease status j, one of a set of disease statuses $j = 1, \ldots, J$. From the definition of a *conditional probability* we know that

$$p(D_j|R) = p(R, D_j)/p(R);$$

that is, the probability of D_j (the disease status j) among patients with a test result R equals the proportion of those with R that also have D_j. Test result R can occur among patients with any disease status $j = 1, \ldots, J$; that is,

$$\begin{aligned} p(R) &= p(R, D_1) + p(R, D_2) \\ &\quad + p(R, D_j) + \ldots + p(R, D_J) \\ &= p(R|D_1) \cdot p(D_1) + p(R|D_2)\, p(D_2) + \\ &\quad p(R|D_j) \cdot p(D_j) + \ldots + p(R|D_J)\, p(D_J). \end{aligned}$$

Substituting the expression for $p(R)$ in the first equation, we get the generalized version of Bayes's formula:

$$p(D_j|R) = \frac{p(R|D_j)p(D_j)}{\sum_j p(R|D_j)p(D_j)}.$$

For a dichotomous (+ or –) test, R becomes either $T+$ or $T-$, and for disease present versus disease absent, D_j becomes $D+$ or $D-$, in which case Bayes's formula becomes

$$p(D+|T+) \quad \frac{p(T+|D+)p(D+)}{p(T+|D+)p(D+) + p(T+|D-)p(D-)},$$

which is the same as

$$\begin{aligned} &\text{Postpositive test probability} \\ &= \frac{\text{Sensitivity} \times \text{Pretest probability}}{\text{Sensitivity} \times \text{Pretest probability} + [1 - \text{Specificity}]} \\ &\quad \times [1 - \text{Pretest probability}]. \end{aligned}$$

For our example, we are interested in the postnegative test probability of CAD, so the appropriate equation is

$$\begin{aligned} p(D+|T-) &= \frac{p(T-|D+)p(D+)}{p(T-|D+)p(D+) + p(T-|D-)p(D-)} \\ &= \frac{(1 - .96) \cdot .13}{(1 - .96) \cdot .13 + .60 \cdot (1 - .13)} \\ &= .01. \end{aligned}$$

Probability Revision With the Odds-Likelihood-Ratio Form of Bayes's Formula

Consider a test result R and true disease status $D+$ and $D-$. From the definition of a *conditional probability,* we know that

$$p(D+|R) = p(R, D+)/p(R)$$
$$= p(R|D+) \cdot p(D+)/p(R),$$

and

$$p(D-|R) = p(R, D-)/p(R)$$
$$= p(R|D-) \cdot p(D-)/p(R).$$

Dividing the first by the second equation, we get

$$\frac{p(D+|R)}{p(D-|R)} = \frac{p(D+)}{p(D-)} \times \frac{p(R|D+)}{p(R|D-)},$$

which is Bayes's in odds-likelihood-ratio form and can also be rewritten as

Posttest (posterior) odds = Pretest (prior) odds × Likelihood ratio for R.

In plain English this means the following:

Our judgment that the patient has the disease after doing the test (posterior odds) equals our judgment that the patient has the disease before doing the test (prior odds), updated with the information we get from the test result R (likelihood ratio for R).

The *likelihood ratio* (LR) for test result R summarizes all the information we need to know about the test result R for purposes of revising the probability of disease. LR for test result R is the ratio of the conditional probability of R given the disease under consideration and the probability of R given absence of the disease under consideration.

The posttest (posterior) *odds* can be converted back to a probability using

$$\text{Probability} = \frac{\text{Odds}}{1 + \text{Odds}}.$$

In our example, we have a 0 calcium score, so we need to use the LR for a negative test result:

$$\text{Probability} = \frac{\text{Odds}}{1 + \text{Odds}}.$$

$$= (1 - .96)/.60 = .067.$$

Prior probability = .13

Prior odds = .13/(1 − .13) = .15

Posterior odds = Prior odds × LR(CTCS−) = .15 × .067 = .0100

Posterior probability = .01/(1 + .01) = .0099

Note that the posterior odds and posterior probability are practically equal because the probability is very low.

M. G. Myriam Hunink

See also Conditional Probability; Diagnostic Tests; Likelihood Ratio; Odds and Odds Ratio

Further Readings

Hunink, M. G. M., Glasziou, P. P., Siegel, J. E., Weeks, J. C., Pliskin, J. S., Elstein, A. S., et al. (2001). *Decision making in health and medicine: Integrating evidence and values.* Cambridge, UK: Cambridge University Press.

BENEFICENCE

In biomedical research, generally, the success of new therapeutic approaches relies on three conditions: specificity, efficacy, and lack of toxicity. These conditions are often tested in cell cultures, mouse models, and clinical trials before a drug is offered to patients. Hence, if biomedical approaches are to be used therapeutically, one should balance the possible harms and the possible benefits of these methods (perform a *risk-benefit analysis*). The terms *harms* and *benefits* are ethically relevant concepts, since ethical obligations or principles about not inflicting harm (*nonmaleficence*) and promoting good (*beneficence*) are generally accepted. The ethical principles of nonmaleficence and beneficence form part of several different ethical theories. For instance, they are the foundation

of the utilitarian theory, which says that ethically right actions are those that favor the greatest good for the greatest number. Another example is the Hippocratic Oath, which expresses an obligation of beneficence and an obligation of nonmaleficence: I will use treatment to help the sick according to my ability and judgment, but I will never use it to injure or wrong them.

This entry analyzes the ethical principles of beneficence and nonmaleficence in biomedicine by drawing on the bioethical theory of principles of the American bioethicists Tom L. Beauchamp and James F. Childress. These ethicists have published their theory in several editions of the book, *Principles of Biomedical Ethics*.

Risk-Benefit Analysis

According to Beauchamp and Childress, the evaluation of risk in relation to possible benefit in biomedicine is often labeled *risk-benefit analysis*. They say that the term *risk* refers to a possible future harm, where *harm* is defined as a setback to interests, particularly in life, health, and welfare. Statements of risk are both descriptive and evaluative. They are descriptive because they state the probability that harmful events will occur, and they are evaluative because they attach a value to the occurrence or prevention of the events. Commonly in the field of biomedicine, the term *benefit* refers to something of positive value, such as life or health. Beauchamp and Childress state that the risk-benefit relationship may be conceived in terms of the ratio between the probability and magnitude of an anticipated benefit and the probability and magnitude of an anticipated harm. Use of the terms *risk* and *benefit* necessarily involves an evaluation. Values determine both what will count as harms and benefits and how much weight particular harms and benefits will have in the risk-benefit calculation.

Risk and benefit identifications, estimations, and evaluations are all stages in risk-benefit analysis; the next step is *risk management*, which Beauchamp and Childress define as the set of individual or institutional responses to the analysis and assessment of risk, including decisions to reduce or control risks. These ethicists believe that while risk-benefit analysis may seem like a technical issue, in which risks and benefits are defined, quantified, and compared, the definition of risk and benefits and the evaluation of how much risk is acceptable (risk management) are clearly ethical issues. Beauchamp and Childress offer an example: Risk management in hospitals includes establishing policies aimed at reducing the risk of medical malpractice suits.

Required Actions

According to Beauchamp and Childress, the balancing of the general ethical principles of nonmaleficence and beneficence is not symmetrical, since our obligation not to inflict evil or harm (nonmaleficence) is more stringent than our obligation to prevent and remove evil and harm or to do and promote good (beneficence). These authors state that our obligation of beneficence requires taking action (positive steps) to help prevent harm, remove harm, and promote good, whereas our obligation of nonmaleficence only requires intentionally refraining from actions that cause harm; hence, nonmaleficence usually involves omissions. Thus, according to Beauchamp and Childress, possible harms associated with potential therapeutics are given more weight in a risk-benefit analysis than the possible benefits. For clarity, Table 1 presents a brief formulation of the principles of beneficence and nonmaleficence of Beauchamp and Childress.

Different Kinds of Beneficence

The question remains, however, whether we are obligated to sacrifice ourselves to benefit others. Beauchamp and Childress believe that there are limits to the demands of beneficence. They distinguish between *obligatory beneficence* (in the forms of general beneficence and specific beneficence) and *optional beneficence* (in the form of ideals of beneficence).

General Beneficence

According to Beauchamp and Childress, a person *X* has a determinate obligation of beneficence toward Person *Y* if and only if each of the conditions listed in Table 2 is satisfied (assuming *X* is aware of the relevant facts).

Specific Beneficence

Beauchamp and Childress state that obligations of specific beneficence usually rest on special moral

Table 1 Two of the four principles of biomedical ethics: beneficence and nonmaleficence (a brief formulation of the bioethical principles of beneficence and nonmaleficence of Beauchamp and Childress)

The principle of beneficence
One ought to prevent and remove evil or harm.
One ought to do and promote good.
One ought to weigh and balance the possible goods against the possible harms of an action.
The principle of nonmaleficence
One ought not to inflict evil or harm. Or, more specifically, one ought not to hurt other people mentally or physically.

Table 2 Conditions determining the obligation of general beneficence of Beauchamp and Childress

1. *Y* is at risk of significant loss of or damage to life or health or some other major interest.
2. *X*'s action is needed (singly or in concert with others) to prevent this loss or damage.
3. *X*'s action (single or in concert with others) has a high probability of preventing it.
4. *X*'s action would not present significant risks, costs, or burdens to *X*.
5. The benefit that *Y* can be expected to gain outweighs any harms, costs, or burdens that *X* is likely to incur.

relations (e.g., in families and friendships) or on special commitments, such as explicit promises and roles with attendant responsibilities (such as healthcare professional and patient).

Ideal Beneficence

Beauchamp and Childress make a distinction between ideal beneficence and obligatory beneficence in terms of the costs and the risks to the agents of beneficence. Ideals of beneficence involve severe sacrifice and extreme altruism in the moral life (e.g., giving both of one's kidneys for transplantation). According to Beauchamp and Childress, persons do not have an obligation of ideal beneficence; other persons can admire those who fulfill the ideal, but they cannot blame or criticize those who do not practice it.

Strength of Principles

According to Beauchamp and Childress, ethical issues of biomedicine not only include the balance of the possible harms and the possible benefits (risk-benefit analysis), it also includes considerations about respecting the autonomy of the patient or the human subject and justice considerations regarding healthcare allocation. They argue that the four ethical principles of (1) beneficence, (2) nonmaleficence, (3) respect for autonomy, and (4) justice are central to and play a vital role in biomedicine. Table 3 presents a brief formulation of the bioethical principles of respect for autonomy and justice of Beauchamp and Childress.

According to Beauchamp and Childress, no one principle ranks higher than the others. Which principles should be given most weight depends on the context of the given situation. Beauchamp and Childress consider the four principles as *prima facie binding*; that is, they must be fulfilled, unless they conflict on a particular occasion with an equal or stronger principle. These ethicists believe that some acts are at the same time prima facie wrong and prima facie right, since two or more principles may conflict in some circumstances. Agents must then determine what they ought to do by finding an actual or overriding principle. This means that the agents must find the best balance of right and wrong by determining their actual obligations in such situations by examining the respective weights of the competing prima facie principles. For instance, in modern medicine, patients' right to make judgments about treatment is valued. It is discussed in biomedical ethics whether respect for

Table 3 Two of the four principles of biomedical ethics: respect for autonomy and justice (a brief formulation of the bioethical principles of respect for autonomy and justice of Beauchamp and Childress)

The principle of respect for autonomy

- As a negative obligation: Autonomous actions should not be subjected to controlling constraints by others.
- As a positive obligation: This principle requires respectful treatment in disclosing information, probing for and ensuring understanding and voluntariness, and fostering autonomous decision making.

This principle does not count for persons who are not able to act autonomously: Infants and drug-dependent patients are examples. However, these persons are protected by the principles of beneficence and nonmaleficence.

The principle of justice

Beauchamp and Childress examine several philosophical theories of justice, including egalitarian theories that emphasize equal access to the goods in life that every rational person values. Beauchamp & Childress propose that society should recognize an enforceable right to a decent minimum of healthcare within a framework for allocation that incorporates both utilitarian and egalitarian principles.

the autonomy of patients should have priority over professional beneficence directed at those patients; hence, there are conflicts between beneficence and respect for autonomy (the problem of paternalism).

Beauchamp and Childress believe that the principles find support across different cultures. They claim that the principles are part of a cross-cultural common morality and that in all cultures people who are serious about moral conduct accept the norms of this common morality. However, even though these principles are generally acknowledged, this does not mean that there is consensus about what is good and bad; the principles are to be specified, balanced, and interpreted in different cultural settings.

Although Beauchamp and Childress's theory is widely used and outstanding in bioethics, it is also subject to much philosophical discussion. For example, in an attempt to criticize philosophical bioethics in general, the ethicist Adam M. Hedgecoe points to Beauchamp and Childress's theory in his 2004 article, *Critical Bioethics: Beyond the Social Science Critique of Applied Ethics*, because principlism is the dominant way of doing bioethics. Hedgecoe claims that philosophical bioethics gives a dominant role to idealized rational thought and tends to exclude social and cultural factors. He believes that principlism defends abstract universal principles without empirical evidence and that principlism develops and justifies theories without paying attention to the practical application of those theories. As an alternative to principlism, Hedgecoe defends the position of what he calls *critical bioethics,* where the results of empirical research feed back to challenge and even undermine the theoretical framework of bioethics.

Some ethicists do not think that Hedgecoe's critique of Beauchamp and Childress's theory is justified. First of all, according to Beauchamp and Childress, there is no straightforward movement from principles to particular judgments. Principles are only the starting points and, as such, general guidelines for the development of norms of appropriate conduct. The principles need to be supplemented by paradigm cases of right action, empirical data, organizational experience, and so on. Beauchamp and Childress state that rights, virtues, and emotional responses are as important as principles for ethical judgment. Secondly, in his 2003 article, *A Defense of the Common Morality*, Beauchamp stresses the importance of empirical research for ethical principles. He claims that the usefulness of the four principles can be tested empirically and that the question of whether they are part of a cross-cultural common morality can be explored. Beauchamp does not present any empirical data generated systematically by qualitative

research to support this position. But he does invite the design of an empirical research study to investigate the issue. For example, a Danish empirical study by Mette Ebbesen and B. D. Pedersen shows that the four bioethical principles of Beauchamp and Childress are reflected in the daily work of Danish oncologist physicians and Danish molecular biologists. Empirical research can likely improve the bioethical theory of principles by bringing it into concord with practice.

Mette Ebbesen

See also Bioethics; Risk-Benefit Trade-Off

Further Readings

Beauchamp, T. L., & Childress, J. F. (1989). *Principles of biomedical ethics* (3rd ed.). Oxford, UK: Oxford University Press.

Beauchamp, T. L., & Childress. J. F. (2001). *Principles of biomedical ethics* (5th ed.). Oxford, UK: Oxford University Press.

DeGrazia, D. (1992). Moving forward in bioethical theory: Theories, cases, and specified principlism. *Journal of Medicine and Philosophy, 17,* 511–539.

Ebbesen, M., & Pedersen, B. D. (2007). Using empirical research to formulate normative ethical principles in biomedicine. *Medicine, Health Care, and Philosophy, 10*(1), 33–48.

Ebbesen, M., & Pedersen, B. D. (2008). The principle of respect for autonomy: Concordant with the experience of oncology physicians and molecular biologists in their daily work? *BMC Medical Ethics, 9, 5*.

Engelhardt, H. T., Jr. (1998). Critical care: Why there is no global bioethics. *Journal of Medicine and Philosophy, 23*(6), 643–651.

Frankena, W. (1973). *Ethics* (2nd ed.). Englewood Cliffs, NJ: Prentice Hall.

Hedgecoe, A. M. (2004). Critical bioethics: Beyond the social science critique of applied ethics. *Bioethics, 18*(2), 120–143.

Holm, S. (1995). Not just autonomy: The principles of American biomedical ethics. *Journal of Medical Ethics, 21*(6), 332–338.

Lustig, B. A. (1998). Concepts and methods in recent bioethics: Critical responses. *Journal of Medicine and Bioethics, 23*(5), 445–455.

O'Neill, O. (2001). Practical principles & practical judgment. *Hastings Center Report, 31*(4), 15–23.

Pellegrino, E. (2008). *The philosophy of medicine reborn: A Pellegrino reader*. Notre Dame, IN: University of Notre Dame Press.

Strong, C. (2000). Specified principlism: What is it, and does it really resolve cases better than casuistry? *Journal of Medicine and Philosophy, 25*(3), 323–341.

BIAS

In statistics, *bias* generally refers to a systematic distortion of a statistical result. Bias can occur in both the process of data collection and the statistical procedures of data analysis. Very few studies can avoid bias at some point in sample selection, study conduct, and results interpretation. Analysis of results without correcting the bias can be misleading and harmful in decision making. With careful and prolonged planning, researchers may reduce or eliminate many potential sources of bias. Collaboration between the statistician and the domain expert is very important, since many biases are specific to a given application area. This entry discusses two different aspects that the term *bias* is commonly used to describe.

Bias in Sampling

Bias in sampling is the tendency that the samples differ from the target population from which the samples are drawn in some systematic ways. A few important concepts include the following.

Biased Sample

Most biases occur during data collection, often as a result of taking observations from an unrepresentative subset of the population rather than from the population as a whole. A sample is said to be a *biased sample* if the probability of a member in the population being sampled depends on the true value(s) of one or more variables of interest of that member. The sampling process that leads to a biased sample is called *biased sampling*. For example, if women with a family history of breast cancer are more eager to join a mammography program, the sample of women in the mammography program is a biased sample of all women. If the variable(s) is important to a study, conclusions based on biased samples may not be valid for the population of interest.

Sample weights can sometimes be used for correcting the bias if some groups are underrepresented in the population. For instance, a hypothetical population might include 50 million men and 50 million women. Suppose that a biased sample of 100 patients included 70 men and 30 women. A researcher can correct for this imbalance by attaching a weight of 5/7 for each male and 5/3 for each female. This would adjust estimates to achieve the same expected value as a sample that included exactly 50 men and 50 women.

Response Bias

Response bias is a type of cognitive bias that occurs when the sampled members from a population tend to produce values that systematically differ from the true values. It happens frequently in survey studies and affects the results of a statistical survey, especially when the questions on a survey are not properly worded or if the question relates to some variables that are sensitive to the members being surveyed, such as household income or drug history. In such situations, respondents answer questions in the way they think the questioner wants them to answer rather than according to their true beliefs.

Nonresponse Bias

Nonresponse bias is an extreme form of biased sampling. Nonresponse bias occurs when responses are not obtainable from all members selected for inclusion in the sample. Nonresponse bias can severely affect the results if those who respond differ from those who do not respond in important ways. Online and phone-in pools may be subject to nonresponse biases because many members in the target population may not have a phone or access to the Internet.

Measurement Bias

The term *measurement error bias* usually refers to systematic deviation from the true value as a result of a faulty measurement instrument, for instance, an improperly calibrated scale. Several measurements of the same quantity on the same experiment unit will not in general be the same. This may be because of natural variation in the measurement process. In statistical analysis, measurement error in covariates has three main effects: (1) It causes bias in parameter estimation for statistical models; (2) it leads to a loss of power, sometimes profound, for detecting interesting relationship among variables; and (3) it masks the features of the data, making graphical model analysis difficult.

Censoring Bias

Censoring bias occurs when a value occurs outside the range of a measuring instrument. Limitations in censoring at either end of the scale can result in biased estimates. For example, a bathroom scale might only measure up to 250 pounds. If a 320-pound individual is weighed using the scale, the observer would only know that the individual's weight is at least 250 pounds. Censoring bias is also common in survival analysis. Special techniques may be used to handle censored data.

Bias in Estimation

Another kind of bias in statistics does not involve biased samples but does involve the use of a statistic whose average value differs from the value of the quantity being estimated. In parameter estimation, bias refers to the difference between the expected value of an estimator and the true value of the parameter being estimated. An estimator with zero bias is called an *unbiased estimator,* and an estimator having nonzero bias is said to be a *biased estimator*.

Suppose a researcher is trying to estimate the parameter μ using an estimator $\hat{\mu}$ (i.e., a certain function of the observed data). The bias of the estimator μ is defined as the expected value of the difference between the estimator and the true value. This can be written mathematically as

$$\text{Bias}(\hat{\mu}) = E(\hat{\mu} - \mu) = E(\hat{\mu}) - \mu.$$

A famous example of a biased estimator is the sample variance. Suppose $X_1, X_1, \ldots, X_n$ are independent and identically distributed (i.i.d.) random variables with expectation μ and variance σ^2. The sample mean is defined as

$$X = \frac{X_1 + X_2 + \cdots + X_n}{n},$$

and the sample variance is defined as

$$S^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - X)^2.$$

It can be shown that the sample mean is an unbiased estimator, while the sample variance is a biased estimator, where

$$E(X) = \mu,\ E(S^2) = \frac{n-1}{n}\sigma^2 \neq \sigma^2.$$

Although biased estimators sound pejorative, they may have desirable statistical properties. For example, they sometimes have a smaller mean squared error than any unbiased estimator. Biased estimators are used in some special cases of statistical analysis.

Xiao-Feng Wang and Bin Wang

See also Bias in Scientific Studies; Probability Errors

Further Readings

Indrayan, A., & Sarmukaddam, S. B. (2001). *Medical biostatistics*. New York: Marcel Dekker.

Stewart, A., & McPherson, K. (2007). *Basic statistics and epidemiology: A practical guide* (2nd ed.). New York: Radcliffe.

Biases in Human Prediction

In the language of cognitive psychology, the ability to predict is the ability to infer, estimate, and judge the character of unknown events. By this definition, a large part of clinical medicine requires that physicians make medical predictions. Despite its importance, it remains subject to many biases. There are a number of important biases affecting medical prediction in diagnosis, prognosis, and treatment choices. This is particularly true in emotionally intense medical circumstances at the end of life. Physicians, patients, and policy makers should be aware of these biases when confronted with decisions in all these circumstances to help avoid their consequences. This entry outlines ways in which cognitive biases often prevent accurate medical predictions across a number of decision-making situations.

Medical Prediction

One major type of medical prediction is the diagnosis of patients' disease. Diagnosis involves gathering and integrating evidence, testing hypotheses, and assessing probabilities. This requires that a clinician be able to generate accurate predictions from incomplete data about the underlying cause(s) of the patient's symptoms. For example, the symptom "pelvic pain" might be caused by a urinary tract infection, a sexually transmitted infection, or by cancer, among other possible diagnoses. A physician who sees a patient with this symptom must accurately predict the likelihood of multiple possible underlying causes to effectively gather evidence (i.e., ask about other possible symptoms and order appropriate tests), cognitively integrate that evidence, and determine the most probable diagnosis.

Once the physician has made a diagnosis, he or she must, along with the patient, make another medical prediction when they decide together on a treatment decision. Selecting the optimal treatment from multiple options requires that a clinician be able to predict which treatment will provide the patient with the best possible health outcome, accounting for both positive and negative effects. For example, a patient with localized prostate cancer has multiple treatment options available, including surgery, radiation therapy (of two types), hormone deprivation therapy, and surveillance. To make a treatment recommendation, a physician must predict the patient's response to various treatments, both in terms of disease control and potential burden from treatment side effects. The physician must also consider the patient's overall health, comorbidities, resources, social support, and preferences for possible health states.

Physicians also make medical predictions when necessary to provide prognoses, which are predictions of the likely duration, course, and outcome of a disease based on the treatment chosen. This is particularly important in diseases, such as terminal cancer, where patients and their families wish to form appropriate timelines for goals of care and to have access to certain types of care, such as hospice, when they would most benefit from them. Unfortunately, as Nicholas Christakis has shown, prognosis is particularly difficult in emotionally intense situations such as this.

Given the centrality of accurate predictions to medical decision making and the common assumption that medical training improves physician's decisions, it is disheartening that research has repeatedly shown that physicians' medical predictions are as susceptible to cognitive biases as others are in nonmedical domains. The mistakes are systematic, not random, errors that are likely due to the difficulty of the prediction task combined with human psychology. Thus, these biases are not significantly reduced by current medical training. As Reid Hastie and Robyn Dawes argue, one of the most persistent of these biases is overconfidence concerning one's predictions. The danger of overconfidence is that one cannot begin to correct other biases affecting the quality of one's predictions; simply recognizing their existence is something that overconfidence prevents. For example, an overconfident surgeon might regularly predict better surgical outcomes for his or her patients and perform surgeries on patients who are poor candidates for surgery. This overconfidence bias will go uncorrected because it is unrecognized as a systematic error.

Biases Affecting Diagnosis

The way in which possible diseases are represented has been shown to give rise to systematically different probability predictions in diagnosis. In one study, house officers were given a case description of a 22-year-old woman with right lower quadrant abdominal pain of 12 hours' duration. Half were asked to estimate the probabilities of the competing diagnostic possibilities gastroenteritis, ectopic pregnancy, or neither. The other half were asked to estimate the probabilities of five diagnoses: (1) gastroenteritis, (2) ectopic pregnancy, (3) appendicitis, (4) pyelonephritis, (5) pelvic inflammatory disease, or "none of the above." Although physicians in both groups were told that their probabilities must sum to 100%, the judged probability of "none of the above" was significantly smaller in the shorter list (50%) of diagnoses than in the longer list (69%). Logically, the opposite should be true, since the additional choices decrease the chances of none of the available diagnoses being correct.

This suggests that physicians do not think enough about diagnoses that are either not listed or that are not what they are currently thinking that the underlying problem might be, and that they don't pay close enough attention to probabilistic information such as the "base rate" of the disease in the population. In medicine, this can lead to inappropriate confirmatory testing, where physicians increase costs without increasing the likelihood of a correct diagnosis. That is, they order tests that will confirm what they already know, making them overly confident of their diagnoses without actually providing any new information. In the long term, overconfidence and failure to correct for cognitive biases cause more experienced physicians to be more confident, but not more accurate, than less experienced physicians. It is easy to see how this can perpetuate a pattern of misdiagnosed and inappropriately treated patients because, as findings from social psychology have demonstrated, less confident and experienced people are more likely to defer to more experienced experts than to try to find alternative explanations.

The first step in correcting such biases is to demonstrate their existence and to alert physicians to their presence. However, another common cognitive bias, the hindsight bias, makes it difficult to learn from cases that show the errors of others. This bias has been demonstrated experimentally in physicians. In the relevant experiment, five different groups of physicians were presented with a challenging diagnostic case describing a patient with a mix of symptoms along with four potential diagnoses. Physicians in the control group were asked to predict the likelihood of each of the four diagnoses given the symptoms. Those in the other four groups were told which of the potential diagnoses was the "actual" one (each group was given a different one) and asked for the probabilities that they would have assigned to each of the four diagnoses. Physicians in each of the four "hindsight" groups inflated the probability that they would have assigned to the diagnosis they were told was correct. This has an important clinical implication because of the similarity of the experimental conditions to teaching rounds presentations. Challenging diagnostic cases presented authoritatively at teaching rounds may seem far more obvious than they really are because of hindsight bias, leading medical team members to fail to learn the difficulty of prediction illustrated by the case because they "knew it all along."

Biases Affecting Prognosis

Physicians also often commit the value-induced bias in medical prediction in which they unknowingly

distort relevant probabilities regarding patient prognosis so as to justify poorly justified treatment choices. This bias helps explain why such a high percentage of the U.S. healthcare budget is spent on patients in the last 6 weeks of life. No physician wants to give up on a desperately ill patient (who may be cured) by stopping treatment, so physicians exaggerate the likelihood of success from treatment. This often leads to prolonging invasive (and often painful) treatments in the face of overwhelming odds against success in the belief that the patient might benefit, even though statistics are clearly against such an outcome.

Accurate prognosis is most valuable to patients and their families, and most difficult for physicians, in life-limiting illnesses such as terminal cancer. Physicians are remarkably poor at predicting the life expectancies of patients with terminal illnesses. In his book on medical prognosis at the end of life, Christakis points out that while prognosis can be a technically difficult task for physicians in many circumstances, the emotional difficulties associated with prognosis at the end of life make their prognoses in such cases even worse.

One of the biases that Christakis emphasizes is the superstitious belief in self-fulfilling prophecies of prognoses. In these situations, physicians seem to feel that by acknowledging a limited prognosis and treating appropriately with palliation, physicians will hasten a patient's demise. Evidence demonstrates that this is not the case; patients undergoing palliative care live just as long (or longer) as similarly diagnosed patients who have been given overly optimistic prognostic information. He demonstrates that physicians are far more likely to ameliorate patient pessimism or provide encouragement, even when unwarranted, than they are to correct unrealistic optimism. Although the motivation to provide hope to one's patients that motivates this response is largely a positive and compassionate one, the responsibility to provide accurate prognostic information and appropriate treatment planning is equally important, something this bias prevents.

Biases Affecting Treatment

To select the best treatment for a patient, a physician must predict the patient's adherence to the therapy. This is more challenging than it might at first appear, because even though patients generally do want to adhere to their treatment regimes, it is much easier to talk about changing a future behavior (e.g., taking a medicine regularly) than it is to actually do it. This "empathy gap" between one's current situation and one's future situation makes it very hard for physicians to appreciate the power of various visceral factors causing patients to make choices that they know are not good for their health.

For example, studies show that adherence to a medication schedule drops with the number of pills a patient is to take. So a physician might prescribe a blood pressure medication to a hypertensive patient (who is already taking several other medications), only to find out at a follow-up visit that the patient's hypertension has not improved. The problem may be that the patient needs a higher dose of the medication. However, there is also a substantial probability that the patient has not managed to adhere to the medication schedule. Nevertheless, studies comparing physician-prescribing behavior with patient prescription-filling behavior indicate that physicians almost always respond by prescribing the higher dosage, even when the prescription for the lower dosage is not being refilled. Physicians fail to predict that patients are not taking the currently prescribed dose.

Most people, including physicians, tend to underweight statistical evidence relative to other forms of evidence such as personal experience. Even though physicians are now trained in evidence-based medicine (EBM), which emphasizes following statistical guidelines based on the medical literature, they often fail to apply relevant statistical data. For example, statistical data tell us that there is no survival advantage to using pulmonary artery catheterization in the intensive care unit to guide fluid management for patients, making this an unnecessary procedure for guiding treatment choices. However, it is still commonly practiced because it is experientially convincing to closely monitor a patient's pulmonary artery pressures, even knowing that doing so does not improve patient outcomes. Before EBM, monitoring pulmonary arterial pressures via catheterization seemed like a logical thing to do based on pathophysiology; however, data do not support this practice. Thus, the practice continued beyond its statistical justification.

William Dale, Liz Moliski, and Joshua Hemmerich

See also Diagnostic Process, Making a Diagnosis; Physician Estimates of Prognosis; Prediction Rules and Modeling

Further Readings

Chapman, G. B., & Sonnenberg, F. A. (Eds.). (2000). *Decision making in health care: Theory, psychology, and applications.* Cambridge, UK: Cambridge University Press.

Christakis, N. (1999). *Death foretold: Prophecy and prognosis in medical care.* Chicago: University of Chicago Press.

Glare, P., Virik, K., Jones, M., Hudson, M., Eychmuller, S., Simes, J., et al. (2003). A systematic review of physicians' survival predictions in terminally ill cancer patients. *British Medical Journal, 327*, 195–198.

Hastie, R., & Dawes, R. M. (2001). *Rational choice in an uncertain world: The psychology of judgment and decision making.* Thousand Oaks, CA: Sage.

Loewenstein, G., Read, D., & Baumeister, R. F. (Eds.). (2003). *Time and decision: Economic and psychological perspectives on intertemporal choice.* Thousand Oaks, CA: Sage.

The National Heart, Lung, and Blood Institute Acute Respiratory Distress Syndrome (ARDS) Clinical Trials Network. (2006). Pulmonary-artery versus central venous catheter to guide treatment of acute lung injury. *New England Journal of Medicine, 354,* 2213–2224.

Stone, P. C., & Lund, S. (2007). Predicting prognosis in patients with advanced cancer. *Annals of Oncology, 18,* 971–976.

Bias in Scientific Studies

In empirical science, bias is any factor that may systematically distort quantitative or qualitative conclusions and recommendations. Psychological sources of biases have separate encyclopedia entries.

Delimitation

Bias must be distinguished from fraud, oversights, misunderstandings, and nonsense arithmetic. It must further be distinguished from the field of statistical pitfalls, illusions, and paradoxes, though each of these, when unrecognized, may bias perceptions and recommendations.

The classical borderline between random error and bias is sometimes fuzzy. The label "bias" is often used about poor data recording, regardless of whether it will affect conclusions and, if so, how. Moreover, blunt procedures (imprecise measurements) may delay the recognition of a health hazard, or benefit, and in that sense pure randomness is itself "biased" against public interests.

Recognition of Bias

Just as, while there is no checklist for the quality of poems, one can develop one's flair for good poetry, the field of bias is open ended. Notwithstanding attempts, it is impossible to devise an exhaustive list of mutually exclusive bias types. Even broad categories such as selection bias and information bias meet at hazy frontiers. But everybody can train his or her flair for detecting bias.

Overly critical readers sometimes find bias where it isn't (*bias bias*), or reject investigations on grounds of bias even when the bias is obviously negligible or purely hypothetical.

Texts often explain a bias by means of hypothetical examples from which all unnecessary adornment has been peeled off. This is the strength, not the weakness, of such examples. "Real patients do not look like that!" is an often-heard but invalid objection. Precisely, the complexity of clinical data often lies behind an investigator's failure to realize that his or her research procedure is biased.

The Estimand

One cannot discuss hits and misses without a bull's-eye. So any discussion of bias presupposes a defined target, the estimand. Not until agreement about the estimand has been reached can the statistician and client proceed to discuss bias and, subsequently, random uncertainty. Key questions are as follows: What do we want to measure? What is a rational measure thereof? For example, What is a rational measure of successful rehabilitation after multitrauma? What precisely is meant by "the waiting time for liver transplantation in 2006"?

There are four rules of thumb for establishing the estimand. (1) It should be conceptually well-defined (often by imagining an ideal method being applied to 10,000 truly representative cases). (2) Its definition should be detached from study design

(i.e., it should parameterize the object process, not the inspection process, with its potential sources of bias). (3) In predictive settings, prospectivity should be built into the definition. This calls for a notion of "a population of naturally occurring identical-looking instances" (*case stream*), to which predictions are meant to apply. Anything that requires hindsight should be weeded out. Care must be taken to define the right units of prediction (women vs. pregnancies; bladder tumors vs. control cystoscopies). (4) Biased and data-driven agendas should be avoided. These remarks apply, *mutatis mutandis*, to qualitative research questions as well.

In studies whose key purpose is comparative, *internal validity* refers to the comparison being fair and *external validity* to the comparison's matching an envisaged target population. *Generalizability* (a broader term) refers to the applicability of study results outside the population sampled.

Formal Definitions of Bias

In theoretical statistics, the bias of a data-summarizing estimator is defined as the amount by which its expected value departs from the population (object-process) parameter, the estimand:

$$\text{Bias} = E(\text{Estimator}) - (\text{Estimand}).$$

An estimator is unbiased when the departure is zero for all values of the parameters in the statistical model of the object process. As an example, for the purpose of estimating a population mean, any average of independent observations is—provably—unbiased, no matter what their common distribution looks like. Their median, on the other hand, is rarely unbiased, except when the distribution is symmetric.

The logarithm of an unbiased estimator is not an unbiased estimator of the log estimand; the same holds true for other nonlinear transformations. In practical biostatistics, many ratio statistics, such as epidemiological odds ratios (OR), are unbiased on log scale, at least approximately. Neither the estimated OR nor its inverse is then unbiased for its estimand. However, unlike the biases engendered by methodological flaws, these "mathematical" biases are often small and tend to zero as sample sizes increase (*asymptotic unbiasedness*) typically faster than the associated standard error (*SE*).

Bias-variance trade-off: In moderately complicated statistical models, the analyst may face a choice between two or more estimator statistics, one of which has little bias, another a low variance. If the loss associated with misestimation is proportional to squared error, $[(\text{Estimator}) - (\text{Estimand})]^2$, one would choose the estimator that minimizes the mean square error (*MSE*). The way this quantity depends on estimator bias and variance is simple:

$$MSE = E\{[(\text{Estimator}) - (\text{Estimand})]^2\} = \text{Bias}^2 + \text{Variance}.$$

Typical applications are those with an inherent risk of overfitting the data: probability density estimation, multivariate discrimination (statistical diagnosis), recursive partitioning trees, and so on.

Conditional bias and unbiasedness, given a (hidden or observable) event *E,* refer to the conditional distribution of the estimator given that *E* is present.

A *median-unbiased* estimator overrates and underrates the estimand equally often.

A *significance test* is said to be *biased* if the probability of rejecting the null hypothesis is sometimes smaller when it is false than when it is true: With a test at the 5% level, certain alternatives enjoy a power <5% (Type II error risk > 95%). Everyday statistical tests are designed to maximize power and have little or no bias.

Confidence limits: Confidence intervals around a biased estimator typically inherit the bias, but there is no standard notion of bias in connection with confidence limits. Relevant concerns include the following. Incorrect coverage is when a nominal 95% interval will straddle the true value of the estimand either more or less than 95 times out of 100. More appropriately, an upper 97.5% limit can be said to be biased if the probability that it exceeds the estimand is not .975: The limit is either misleadingly large or does not offer the claimed protection. Analogous remarks apply to the lower limit. Confidence intervals offer protection against random variation only; protection against bias must rest on plausible assumptions concerning systematic errors, preferably built into sensitivity analyses. If a data summary is beset with a bias of unknown magnitude, the protection offered by a confidence interval is spurious, unless

the bias is obviously small relative to the width of the interval.

Biases in the Scientific Process

Biased Agendas

Sticking to easy, noncontroversial, and fundable aspects of a health problem could count as biased question asking. Health outcomes, for example, are easier to handle than questions of patient-physician rapport.

Data-Driven Agendas

The decision of what questions to answer and what parameters to estimate should be made beforehand (frequentist statistical theory presupposes that estimates and tests are reported no matter how the observations turn out). When clinical trialists selectively report those outcomes that have produced statistically significant differences, we have an instance of data-driven question asking, and their report should be received with skepticism due to the perils of *multiplicity* (multiple tests bombarding the same null hypothesis) and data dredging. Selective reporting also presents a severe obstacle to meta-analyses that try to amalgamate several studies.

Data Dredging

Data dredging is when researchers keep ransacking their data set until something "significant" turns up. One can have equally little trust in diagnostic indices, or indices of therapeutic success, constructed by combining the variables on file in myriad ways and choosing the "best" fit (almost certainly an overfit). *Repeated peeking* at the data as they accrue is similar: When trends in the data themselves codetermine when to stop and produce a report, we have an "*informative*" (bias-prone) *stopping rule*. Stopping as soon as, but only when, the data look sufficiently promising will bias results in an optimistic direction.

Conceptual Bias

The interpretation of a given data set is restricted—one may say biased—by narrowness of theoretical outlook (or Kuhnian paradigm). One straightjacket is the idea of the natural course of a disease process. Cholecystectomy was once suspected of causing gastrointestinal cancer. This reflected a failure to realize that premonitory cancerous dyspepsia plus silent gallstones sometimes triggered a cholecystectomy: The association was the diagnosticians' own fault! In sum, disease processes—and the hypotheses formed about them—are shaped, in part, by healthcare culture, its imperfections, and its self-image.

Publication Bias

Investigations having something new or significant to tell are more promptly and widely published. Hence the published literature on any particular date remains biased relative to the body of data analyses actually completed. *Double publication* adds a further slant, as does *citation bias*: Not only are investigators likely to cite preferentially the studies they agree with, but there also appear to be much-cited papers that become cited just because everybody else cites them. The net effect is a self-perpetuating body of knowledge, or prejudice, with insufficient built-in bias correction.

Additional Biases

Unlike the preceding "sociological" topics, the flaws that follow are primarily the responsibility of the individual research team. Again, dishonest action and simple oversights will be bypassed, as will breaches of good research practice.

Bias-prone handling of numerical data includes rounding problems (e.g., age on last birthday vs. exact age). *Misleading design of graphs and tables* should be caught by senior authors or at peer review. Narrow-minded interpretation, or an attempt to save words, may lie behind *misleading conclusions*. A statistical association may easily become "Young taxi drivers were more accident prone," suggesting causality. "Analgesic A proved superior to B" deprives readers of a chance to question the choice of doses.

Blunt analyses are biased toward the "nothing new" conclusion:

1. Unnecessary dichotomization is wasteful of information.
2. Chopping up the data set, with the laudable aim of comparing like with like (*stratification*), may

produce several nonsignificant tests and leave a cross-stratum effect undocumented.

3. Investigators taught to observe certain rules, such as reserving *t* tests for normally distributed data, give up halfway due to "violated assumptions" instead of exploiting the robustness of most statistical procedures. Again, results will be valid but vague. Replace the virtuous stance with a valiant one, and the data will speak.

Biased Data Collection

Overall distortions of answers or measurements may or may not bias results, as in the case of observer-interviewer effects, including interactions (elderly people may balk at the jargon and manners of young interviewers); Hawthorne effects (disciplined behavior when under observation); framing effects (wording of questions); or effects of embarrassing questions and forced choices.

Changes in data collection over time lead to treacherous biases. Examples include training and fatigue effects when interviewers have to conduct interviews over several months and unnoticed slippage or change of reagents in the lab.

Selection bias is a common term for bias due to the sampling of study units being misaligned with the intended population (despite a random or exhaustive sampling scheme). The little fish slip through the net; the big fish tear it apart. Death notices in newspapers are a biased source of longevity data.

Ascertainment bias refers to the data source: Telephone interviewing was notorious for reaching mostly middle-class people. Clinical materials that comprise a woolly mix of prevalent and incident cases are not representative of any recognizable population and, therefore, are biased regardless of study purpose. People who volunteer are self-selected and probably special.

In clinical studies, consecutive enrollment is the primary safeguard against selective forces. Randomized allocation serves to prevent skewed recruitment of comparison groups, and, by facilitating blinding, it helps prevent other types of bias. Concealment of allocation extends the veil of blinding backward to cover enrollment deliberations.

Chronic-disease trials preferentially recruit those who are dissatisfied with their current treatment; the result is a potential bias in favor of any new drug and a selection skew relative to an unselected stream of "my next patient" (first-time as well as chronic cases).

Healthy-worker, or healthy-survivor, *effects* refers to the notion that those who do not give in or succumb to occupational and other stresses are the strong and healthy; even after years of toil and exposure, they may still be healthier, or appear sturdier, than others.

Once selected, cases may be subjected to flawed intervention or flawed data recording. Flawed interventions lead to *performance bias* (think of unequal surgical skills) and *collateral treatment bias* due to secret use of supplementary medication.

As to data recording, *information bias* arises when the study objects "influence" the amount, kind, or quality of their data records. When chemical exposure is documented through labor union records, comparisons may end up being misleading because some trades are associated with less organized lifestyles (even in the absence of solvent-induced brain damage). *Recall bias* in a narrower sense would exist if those with neuropsychological impairment were, or were helped to become, more aware of past exposure. *Missing data* will cause bias if reluctance to provide data is somehow related to the study question (if an eligible subject's very existence also remains unrecorded, a selection problem is added).

Unequal contact with healthcare providers may skew the records (*surveillance, detection, verification, workup, access bias*). The risk of endometrial neoplasia in women on menopausal hormone replacement therapy once seemed high, but the excess was explained by occasional curettage prompted by bleeding.

Attrition bias: Dropouts from clinical trials pose a major problem, whether unbalanced or not, as the strict intention to treat (ITT) paradigm requires all outcomes to be recorded and utility assessed, preferably with equal precision.

Investigator-induced information bias: In the context of diagnostic test evaluation, *discrepant analysis* consists in trying to resolve discrepancies between the study test and the reference test by appealing to one or more arbiter tests in the hope of proving the reference test wrong; cases of agreement are not similarly challenged. An optimistic bias ensues.

Purity bias: Clinical investigators are sometimes obsessed with purity. They are reluctant to delve into the muddy waters of everyday case streams. Patients are thrown out arbitrarily when they look "atypical," even retrospectively. The resulting biases mostly involve violations of prospectivity. Downright short-circuits may occur, for example, when drugs intended to reduce infarct size suppress the infarct markers and lead to a final diagnosis of no infarction. Here, the prospective indication (clinical problem) was, and should remain, presumptive infarction.

In a quest for precision—a variant of purity—an oncologist in one study chose to disregard cancer recurrences not datable within ±2 months. As X-ray intervals rose to 6 months after 2 years without recurrence, many late recurrences were discarded, biasing not just the frequency and timing of recurrences but also the ratio between symptomatic and silent ones.

Uncertain Predictors and Covariates

Borrowing a term from radio engineering, statisticians use the word *noise* as a shorthand for unwanted random variation, regardless of source and cause. Noise affecting predictors or covariates gives rise to bias problems quite different from those connected with noisy response variables. One distinguishes between nondifferential and differential misclassification/distortion; that is, given the true predictor, is the noise independent of the response, or not?

Nondifferential Distortion

In a linear regression context, proportional misrepresentation of the predictor causes a proportional change in the apparent regression coefficient (in the opposite direction), whereas a fixed additive term is innocuous; tests are unaffected. Independent measurement variation (additive noise) attenuates the regression coefficient by a factor $S^2/(S^2 + s^2)$, where S is the *SD* of the true predictor and s the noise *SD*; tests also lose power. Other regression models are affected in roughly the same way. Multivariate-covariate adjustments also bias regression coefficients, but there is no general rule about the direction. In two-group comparisons, group differences are also attenuated or destroyed by misclassification.

However, additive noise in predictors may interact with data selection to produce insidious biases, especially when the protocol requires a predictor, such as fasting blood glucose, to stay within the normal range.

Differential Misclassification

Differential misclassification is serious and always subtle. A dietary habit that has been falsely accused of being harmful is given up by those who want to lead, and do lead, a healthy life. The incidence of a disease now proves higher among those who confess to the habit: False suspicion = Conviction (due to population [self-]manipulation). Related is the *treatment paradox*: When known danger signals during pregnancy prompt referral and special care, neonatal outcomes are equalized, falsely suggesting that referrals are unnecessary.

Dependent Observations

Dependent observations may bias comparisons, in addition to invalidating the *SE* formulae. For example, in an individually randomized trial of intensive versus standard poststroke support, patients in the gym share their experiences, producing correlated follow-up interviews; cross-talk between the randomization arms weakens the apparent intervention effect. So neither the observed effect nor its nominal *SE* can be trusted.

Special Bias Mechanisms

Time-related phenomena cause bias if ignored, *censoring* being a familiar example. Similar information biases arise when what happens outside a time window is unobservable or when enrollment is conditional on some event occurring within the window (a *truncation*). For example, conditionally on giving birth within a study window, women are interrogated concerning time to conception; subfertile women are thereby preferentially excluded.

Length bias, *size bias*: The larger the stone on the beach, the more seagull droppings, but not because the gulls take aim. The longer the duration of a condition, the less likely it is that the case will escape admission or notification. Chronic cases dominate cross-sectional snapshots (*prevalence*).

Conversely, rapidly growing cancers are unlikely to be caught at screening.

Cross-sectional surveys of outpatient clienteles conducted on January 1 will be dominated by chronic and late-autumn cases, whereas an analysis of treatments begun and ended within a calendar year are dominated by quick recoveries or by winter cases (in the northern hemisphere). A September case-control study comparing miscarriages with childbirths would show that hay fever at the time of conception is a cause of fetal loss. Even in all-year studies, seasonal variation is warped by weekends and holidays, both on the patient and on the healthcare side.

Lack of *observational synchronization*: Screen-detected cancer patients live longer after diagnosis than other patients even if death dates are unaffected by therapy (*lead bias*). Responders to cancer therapy live longer than nonresponders, if only because the predicate "responder" takes some time to acquire.

Berkson's fallacy concerns a bias in hospital studies. To statisticians, the distinguishing feature is this: An attempt to compare the frequencies of some property *A* given presence and absence of disorder *B* is made in a clinic that receives patients with *B* or *C* (note the *disjunction*!), thereby effectively comparing $P\{A|B\}$ with $P\{A|(C \text{ but not } B)\}$ instead of $P\{A|\text{not } B\}$.

Regression toward the mean: An outlying lab value probably holds a random swing, so when the test is repeated, a less extreme value is normally obtained. The pitfall is that of thinking that the change requires a biological explanation. Patients with fluctuating diseases are seen during exacerbations: Improvement follows, even without treatment; causal speculations are misplaced.

Noisy stratification inherits the regression problem: A drug made the heart beat faster in some subjects and slower in others, due to random fluctuation. Convinced that it was a real difference between two classes of people, the pharmacologist documented his result with a *t* test. Not biology, however, but his sorting of subjects into high and low had made the no-difference hypothesis false. Had he repeated the experiment, the two groups would have slipped back toward a joint mean. (Had another inactive drug been added in the second round, he would have discovered an antidote!) Groupings based on noisy criteria are dangerous.

Biased matching: In designs with individual matching, a tiny person will often get a somewhat taller control, and so on. A "sister closest in age" control scheme preferentially picks the middle sister of three, so mid-sib characteristics will be prevalent among controls. With "best friend" schemes, friends-making personalities will be overrepresented among controls.

Jørgen Hilden

See also Bias; Biases in Human Prediction; Confidence Intervals; Confounding and Effect Modulation; Hypothesis Testing; Numeracy; Worldviews

Further Readings

Andersen, B. (1990). *Methodological errors in medical research*. Oxford, UK: Blackwell.

Armitage, P., & Colton, T. (2005). *Encyclopedia of biostatistics* (2nd ed.). Chichester, UK: Wiley Interscience.

Gluud, L. L. (2006). Bias in clinical intervention research. *American Journal of Epidemiology, 163,* 493–501.

Gøtzsche, P. C. (1990). *Bias in double-blind trials*. Copenhagen, Denmark: Lægeforeningens Forlag.

Gøtzsche, P. C. (1990). Bias in double-blind trials. *Danish Medical Bulletin, 37,* 329–336.

Hill, A. B. (1961). *Principles of medical statistics*. London: Lancet.

Mainland, D. (1963). *Elementary medical statistics*. Philadelphia: Saunders.

Porta, M. S. (2008). *A dictionary of epidemiology* (5th ed.). New York: Oxford University Press.

Sackett, D. L. (1979). Bias in analytic research. *Journal of Chronic Diseases, 32,* 51–63.

Bioethics

Ethics or moral philosophy is the branch of philosophy that concerns itself with the analysis of moral propositions and judgments. Bioethics, a neologism first used in the late 1960s, is currently used to describe two slightly different fields of applied ethics: (1) as a broad term covering the ethics of the life sciences and all their applications, including environmental and animal ethics—this is the common usage in Europe—and (2) as a narrower term covering the ethics of new

biotechnological developments and medical/healthcare ethics—this is the common usage in North America. It is the second, narrower, use of the term that is adopted in this entry.

Bioethics differs from traditional medical professional ethics, or medical deontology, in its emphasis on the role of the patient in decision making and the need to respect the patient's self-determination. Ethical considerations play a role in many medical decisions and form part of the background to many kinds of healthcare regulation. Areas where bioethics play a major role in decision making include reproductive medicine, end-of-life decision making, decision making for incompetent patients, and research ethics.

The involvement of bioethics and bioethicists in the development of healthcare regulation may sometimes lead to confusion, especially if the eventual regulation or regulatory body has "ethics" as part of its title. This does not necessarily mean that all the regulations are justified by good ethical reasoning. Research ethics as a regulatory system does, for instance, contain many elements that are not easily derivable from ethical analysis of research practices.

In the present entry, the focus is on bioethics as a branch of applied moral philosophy of use to individual healthcare professionals in their clinical decision making.

One specific feature that sets bioethics somewhat apart from other fields of applied ethics is the development of a number of bioethical frameworks specifically designed to be of direct use in clinical decision making. The most prominent of these is the "four principles" approach.

The Four Principles Approach

The four principles approach was initially developed in the United States. The impetus for the development of this approach was the observation that people can often agree on what should be done, that is, agree on a specific course of action, without being able to agree on why this course of action is the right one.

The basic idea in the four principles approach, or *principlism,* as it is often called by its critics, is that a healthcare professional should consider four ethical principles when making a clinical decision:

1. Respect for autonomy
2. Nonmaleficence (do not cause harm)
3. Beneficence (do good)
4. Justice

The principles are not ranked and none of them is absolute. They are all prima facie in the sense that they can be overridden if there are stronger reasons for following one of the other principles.

When making a decision with ethical implications, a healthcare professional should consider the following: (a) which of these principles are engaged in the decision, (b) how the principles are engaged, and (c) if two or more principles are engaged, whether they point to the same decision or whether they are in conflict and have to be balanced against each other.

In a conflict situation, three questions need to be answered: (1) Does the situation really fall within the scope of the principles? (there may, for instance, be no autonomy to respect if the patient is a fetus or is in a coma), (2) What is the exact entailment of each principle? (What does it tell us to do?), and (3) What is the right decision when the principles are weighed against each other? These three steps are referred to as determining *scope, specification,* and *balancing.*

Within moral theory, the four principles occupy a space in between overarching moral theories and specific moral judgment, and they are, in this sense, midlevel principles (see Figure 1). They can be derived top-down from moral theory. Any serious moral theory must support some version of these principles. No moral theory could, for instance, claim that harming others was not bad. The principles can also be derived bottom-up from the concrete judgments of everyday, common morality. If we reflect on these judgments and try to systematize them, we will also reach the four principles. After the derivation of the principles, we can then dispense with both in-depth consideration of moral theory and the messiness of common morality and use the principles instead.

The claim for the four principles is thus that they can resolve or mediate two kinds of moral disagreement, disagreement at the theoretical level and disagreement at the level of concrete judgments. They are furthermore useful for structuring

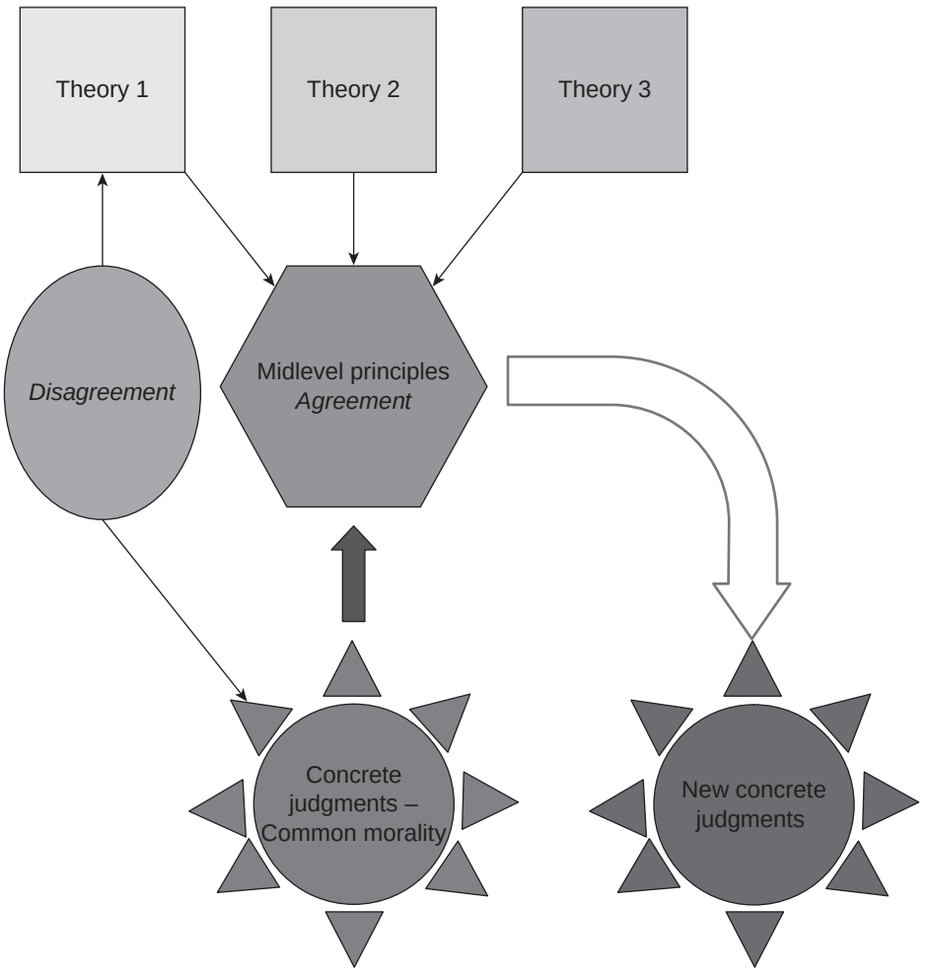


Figure 1 The justification of the four principles

moral reflection and discussion concerning specific decisions.

Critiques of Principlism

The four principles approach has been the subject of considerable criticism focusing on two issues: (1) The principles may hide various forms of moral disagreement and (2) the decision process when two or more principles cannot be satisfied at the same time is unclear.

The first set of criticisms point out that the level at which we can agree on the principles is the level of contentless labels but that if we dig deeper disagreement reappears. Whereas we can all agree that we should be beneficent, that is, that there is some obligation to help others in need, we disagree concerning how strong this obligation is. How much of my wealth should I give to disaster relief, or how strong is my obligation to be a good healthcare Samaritan outside working hours? Many critics link this point to an ambiguity in the bottom-up derivation of the principles from common morality. Is common morality the same everywhere, and will we get similar content in the principles if derived from the common morality of the United States as we get when derived from the common morality of one of the Scandinavian welfare states? Or, put more strongly: Are the four principles really the principles of *American* bioethics?

The second set of criticisms focuses on the decision procedure when principles are in conflict.

There are many situations in healthcare where, for instance, the principle of respect for autonomy will be in conflict with considerations of justice, and to be action-guiding, the four principles approach needs an unambiguous decision procedure to resolve such conflicts. It has been argued that the three steps of determining scope, specification, and balancing are neither individually nor in combination sufficiently clear to provide unambiguous and unbiased decisions. It has especially been pointed out that it is unclear how principles are to be balanced against each other and that intuitions about this may be culturally specific. This links to a further criticism that, although the proponents of the four principles claim that there is no intrinsic or explicit ranking of the principles, there is an implicit ranking, with respect for autonomy and nonmaleficence trumping beneficence and justice. This trumping effect comes about because respecting the autonomy of others and not harming others are what moral philosophers call perfect duties—duties where it is possible to fulfill them completely. But doing good and being just are imperfect duties; there is always something more that can be done, and it is difficult to say precisely when a person is doing too little in respect of one of these duties. In any given situation, it is therefore easier to identify and estimate the importance of a breach of one of the perfect duties than of one of the imperfect duties.

Proponents of the four principles approach respond to these criticisms by claiming that (a) there is actually substantial agreement concerning the content of the principles, despite the protestations of the critics and (b) healthcare professionals find the approach helpful in making decisions, so it must be sufficiently action-guiding despite any inherent vagueness.

Variants of Principlism and Other Frameworks

Several variations on the principlist theme have been developed. These include the Ethical Grid and a transposition of the four principles into an ethics of love.

One impetus behind the development of the Ethical Grid is the argument that it is not enough to respect autonomy. An important element of healthcare practice is to create, promote, and support autonomy in patients and clients. Another is that nonmaleficence and beneficence are two sides of healthcare's central focus on needs and not wants. Based on these considerations and the perceived need to provide more guidance concerning how to think through an ethical problem, a graphical aid—the Ethical Grid—has been developed for analyzing ethical problems in clinical practice and healthcare policymaking. In the grid, core ethical values in healthcare are at the center, more specific rules in the next level, considerations of beneficence in the third level, and more general considerations at the outer level. In using the grid, the first step is to identify which boxes are engaged in the problem at hand. If a problem, for instance, has no resource implications and no other repercussions for anyone else than the patient and the healthcare team, a number of boxes are irrelevant and can be left out of further consideration. In the second step, the implications of the possible choices are then considered for each relevant box in light of the core values in the center. This will identify the reasons for and against each possible choice. Based on this, it should then be possible to reach a conclusion concerning which action is best supported in the present context.

The Ethical Grid has been developed into a more comprehensive Web-based tool for exploring values, The Values Exchange.

Because of the focus on creating and promoting autonomy, the Ethical Grid has become popular in nursing and other professions allied to medicine, where care is seen as equally important to treatment.

Another variation on the principles theme proceeds from the following arguments: (a) that the basis for any ethics must be love in both its emotional and cognitive sense and (b) that the four principles as originally proposed appeal exclusively to the cognitive elements of our relationship with ourselves and with others. It is suggested that we will gain a better understanding of the scope and importance of the principles by understanding them as four different aspects of love, according to the following transposition:

1. Respect for autonomy = Love of self
2. Nonmaleficence = Love of life
3. Beneficence = Love of good
4. Justice = Love of others

Other frameworks that have been proposed for clinical bioethics are the 10 so-called moral rules:

1. Do not kill.
2. Do not cause pain.
3. Do not disable.
4. Do not deprive of freedom.
5. Do not deprive of pleasure.
6. Do not deceive.
7. Keep your promises.
8. Do not cheat.
9. Obey the law.
10. Do your duty.

The primary difference between the 10 moral rules and the four principles is that the rules are more specific and that positive obligations to benefit others are less prominent. The implications of the moral rules for healthcare practice have been explicated in a number of publications.

Liberal Utilitarianism

Within academic bioethics, there is considerable skepticism toward the use of bioethics decision frameworks. Many professional academic bioethicists suggest that the search for midlevel principles or similar devices is misguided and that any proper bioethical decision making needs to be based on moral theory. There is, however, significant disagreement concerning which moral theory to choose.

In what can broadly be described as Anglo-American bioethics (including the north of Europe, Canada, Australia, and New Zealand), liberal utilitarianism has become the preferred approach. Utilitarianism, which is a type of consequentialism, states that the morally right action is the one that maximizes net good consequences. It is, however, well known that unmodified utilitarianism can lead to strongly counterintuitive and very illiberal results.

In contemporary bioethics, utilitarianism is therefore almost always combined with some form of liberal restriction on allowable state or societal action, most often in the form of John Stuart Mill's so-called harm principle:

> That principle is, that the sole end for which mankind are warranted, individually or collectively in interfering with the liberty of action of any of their number, is self-protection. That the only purpose for which power can be rightfully exercised over any member of a civilized community, against his will, is to prevent harm to others. His own good, either physical or moral, is not a sufficient warrant.

If this principle is accepted as a restriction on allowable actions by the state or by individual actors, then individuals have liberty to pursue their own projects as long as they do not harm others, although they may still be morally obligated to sacrifice their own interests for the maximization of good consequences. This has the desirable consequence for the liberal that most decisions made by patients are protected from interference even if they do not maximize good consequences overall. In the healthcare setting, this means that a healthcare professional should respect a patient's choices even if they seem to be to the detriment of the patient.

Liberal utilitarianism does, however, face problems in the context of resource allocation or priority setting in healthcare. Standard utilitarianism is broadly consistent with welfare economics and with health economics approaches to resource allocation, for instance in the form of maximization of quality-adjusted life years (QALY maximization). But the consistency with welfare economics is lost in liberal utilitarianism because of its emphasis on the liberty rights of persons. This has led some liberal utilitarians to argue that any allocation that deprives a person of a treatment that has health benefits is problematic, unless the allocation is done through some kind of lottery that provides everyone a chance to get the treatment that will benefit them, irrespective of resource implications.

Bioethics, the Embryo, and the Fetus

Reproductive decision making has been considerably influenced by bioethical analysis of the status and moral importance of the human embryo and fetus. It is traditionally assumed that embryos and fetuses are morally important in themselves or intrinsically, but this is denied by many writers in bioethics, who hold that they are not morally

important and that there is nothing morally problematic in terminating them.

The arguments for the view that embryos and fetuses have no intrinsic moral importance but are only important if others (e.g., their progenitors) value them vary in their details. The main line of argument is, however, fairly constant and based on the idea that what is wrong with killing an entity is that it frustrates a preference or conscious interest that that entity has. It is thus only wrong to kill people who do not want to be killed, and voluntary euthanasia is by implication acceptable. But embryos and fetuses have no preferences or conscious interests concerning their future existence, either because they are not conscious at all (embryos or early fetuses) or because they do not have the concept of a future existence (late fetuses).

On this view, the creation and destruction of embryos for good reasons, for instance, as part of assisted reproduction or for stem cell research, is morally neutral as is abortion on demand. Although no country has legislation on reproductive medicine that is as liberal as this view of embryos and fetuses requires, it has influenced the move toward liberalization in many countries. Critics of this line of argument point to the fact that it has very wide application. Not only does it entail that abortion on demand is acceptable at any time during a pregnancy but also that infanticide, or the killing of normal infants, on the request of their parents becomes a morally neutral action, since infants are unlikely to have the conscious concept of a future existence. It also entails that persons with severe cognitive deficits are without intrinsic moral value.

Søren Holm

See also Cultural Issues; Rationing; Religious Factors; Shared Decision Making

Further Readings

Beauchamp, T. L., & Childress, J. F. (2001). *Principles of biomedical ethics* (5th ed.). New York: Oxford University Press.

Gert, B. (1973). *The moral rules: A new rational foundation for morality*. New York: Harper & Row.

Gert, B., Culver, C. M., & Clouser, K. D. (1997). *Bioethics: A return to fundamentals* (2nd ed.). New York: Oxford University Press.

Häyry, M. (1994). *Liberal utilitarianism and applied ethics*. London: Routledge.

Holm, S. (1995). Not just autonomy: The principles of American biomedical ethics. *Journal of Medical Ethics, 21,* 332–338.

Macer, D. R. J. (1998). *Bioethics is love of life: An alternative textbook*. Retrieved from http://www.eubios.info/BLL.htm

Mill, J. S. (1879). *On liberty and the subjection of women*. New York: Henry Holt. Retrieved January, 23, 2009, from http://oll.libertyfund.org/title/347

Seedhouse, D. (1998). *Ethics: The heart of health care* (2nd ed.). Oxford, UK: Wiley/Blackwell.

Steinbock, B. (Ed.). (2007). *The Oxford handbook of bioethics*. New York: Oxford University Press.

The Values Exchange: http://www.values-exchange.com/news

BIOINFORMATICS

We are on the cusp of an explosion of biological data generated by the human genome project and sequencing projects in multiple organisms, coupled with advances in both experimental and information technologies. All this is contributing to a new era of personalized medicine, using a deeper understanding of our bodies and their diseases at the molecular level. The huge demand to manage, analyze, and interpret these various data has led to the growing stature of the field of information science that is called bioinformatics.

Bioinformatics encompasses all aspects of biological information—acquisition, processing, storage, distribution, analysis, and interpretation—and combines the tools and techniques of mathematics, computer science, and biology with the aim of furthering the understanding of diseases. The National Institutes of Health defines bioinformatics as "research, development, or application of computational tools and approaches for expanding the use of biological, medical, behavioral or health data, including those to acquire, store, organize, archive, analyze, or visualize such data."

Bioinformatics can be viewed as a bottom-up approach, working with molecular data to determine physiological information. In contrast, medical informatics can be viewed as a top-down approach, working with patient clinical data to

determine underlying physiological processes. Together, bioinformatics and medical informatics are key methods shaping the future of personalized medicine. This means that, due to bioinformatics analysis of genomic data, medical decision making is evolving to be based on a person's individual genomic information instead of on studies relying on statistics about the general population.

History

As a field of biological and information science, bioinformatics has been present since the discovery of DNA, when proteins and cell forms became known as the building blocks of life. The cardinal functions of bioinformatics have been (a) handling and presentation of nucleotide and protein sequences and their annotation; (b) development of databases to store, analyze, and interpret these data; and (c) development of algorithms for making predictions based on available information. To address these topics, the field drew from the foundations of statistics, mathematics, physics, computer science, and molecular biology. Bioinformatics still reflects this broad base.

The Human Genome

The genome's language is a DNA code containing an alphabet of just four letters, or bases: G, C, A, and T. Remarkably, the entire human genome contains 3 billion of these DNA bases. While sequencing the human genome to decode these billions of bases in multiple people from different ethnicities, bioinformatics technologies were used and improved to view, combine, compare, and find patterns across this enormous amount of data. By comparing sequences of known and unknown genes, bioinformatics programs were developed that used probabilities and modeling to predict the function and roles of previously unknown genes. This vast amount of completed genomic sequence data and individual gene information now also needed to be stored, bringing about the creation of various gene databases that are publicly available.

During this genomic era, bioinformatics tools played a pivotal role in allowing researchers to generate and compare the DNA sequences of many genes to identify their roles and to determine whether a particular gene sequence has different DNA bases than seen normally. This information has provided insights into many biochemical, evolutionary, and genetic pathways. It has also provided an important building block for potential medical decision making by making it possible to identify whether a patient's specific gene is normal or mutated.

Bioinformatics in the Postgenomic Era

The map of the human genetic code provides information that allows researchers and physicians to pursue new options for diagnosing and eventually treating many diseases, symptoms, and syndromes. Bioinformatics has enabled these discoveries via analysis and comparison of the various data sets of the genomic era.

Advances in experimental technologies for detecting the multiple levels of biological organization (DNA, RNA, or protein) on a high-throughput scale have required the bioinformatics field to develop increasingly more sophisticated methods and systems for analyzing and storing data. The emerging era of medicine depends strongly on a broad array of these new technologies, such as DNA sequencing, gene expression profiling, protein profiling, and developing new algorithms for finding patterns across large, sometimes dissimilar data sets. Bioinformatics methodologies are useful due to their ability to sift through this vast array of information to converge on a few relevant facts. Together, these new high-throughput technologies and bioinformatics analyses are providing the ability to understand and predict the behavior of complex biological systems, giving rise to the field of systems biology. We have arrived at a point in biology where the underlying mechanisms behind diseases are becoming known.

Gene expression microarrays and single nucleotide polymorphism (SNP) genotyping are two major areas where bioinformatics plays a vital role in interpreting the data generated from these high-throughput technologies. Analysis of the gene-expression profiles from healthy and diseased persons can provide the identification of what genes may be responsible for that disease, which can be investigated further using several technologies. Genotyping identifies an individual's DNA sequence, and bioinformatics analysis across genotypes provides a measurement of the genetic variation between

those genotypes, or between members of a species. SNPs are changes in a single base of the DNA sequence, and these are often found to be the etiology of many human diseases and are becoming particularly important in pharmacogenetics. SNPs can also provide a genetic fingerprint for use in identity testing.

Scientific advances coupled with novel bioinformatics algorithms have helped uncover other functional elements of the genome such as miRNAs (microRNAs), RNAi (RNA interference), and so on, depicting the complex nature of the genome and its regulation. As newer molecules are discovered, the need to manage, analyze, and interpret them is also being addressed, using bioinformatic tools.

In the postgenomic era, bioinformatics has helped create integrated resources of databases, pathways, functions, and visualization tools to assimilate and integrate multiple layers of a person's molecular data. These resources and capabilities are enabling researchers to understand how the molecular processes of cells are linked to higher physiological functions.

Applications of Bioinformatics in Medical and Health-Related Research

Bioinformatics has the potential to influence a wide range of medical and health-related research, with subsequent downstream effects translated into more individualized medical decision making.

The mining, or comparison, of similar sets of patient gene expression data from microarray chips can find which genes are differentially expressed in patients as compared with the normal population. To further understand how changes in certain genes are linked to the clinical outcome, bioinformatics can be leveraged to provide information on the genes of interest, such as their processes, functions, interactions with other genes or proteins, genetic pathways, and any known drug targets associated with them.

Genotyping and bioinformatics play a key role in the search for genes that increase the susceptibility to specific diseases; for their genetic variations (SNPs); for SNP patterns that can be used to predict patient response to medicines; for identifying tractable drug targets; and for defining the function of the genes and proteins they produce.

Understanding the relationship between genetic variation and biological function on a genomic scale is expected to provide fundamental insights into the biology, evolution, and pathophysiology of humans and other species. Analyzing and comparing the genetic material of different species is an important method for studying the functions of genes and the mechanisms of inherited diseases. For example, by comparing the sequences in newly identified genes with those of genes whose functions are already known, scientists can make educated interpretations about which genes might be related to specific biochemical pathways in the body and how they might affect the occurrence or treatment of the disease. This information can also be used to experimentally model those sequence changes to verify these gene functions and to test if there is a better or worse response to drug treatment. Based on the differences in the genetic variants among ethnic groups, one can predict the appropriate dosage for a drug to be effective or to avoid serious side effects. This growing body of information from bioinformatics analysis is the basic foundation of the field of pharmacogenomics.

Pharmacogenomic approaches, which involve the study of how an individual's genetic inheritance affects the body's response to drugs, are emerging across broad classes of therapeutics to assist practitioners in making more precise decisions about the correct drugs to give to the appropriate patients to optimize their benefit-to-risk ratio. Bioinformatic analysis of data helps eliminate false-positive leads in the early stages of the drug discovery process, thus substantively compressing the time, resources, and costs needed in the drug discovery efforts. These approaches are continuously evolving to address the complexity and multivariate nature of increasing amounts of data. There are many clinical drug trials, sponsored by the pharmaceutical industry, that leverage bioinformatics with medical informatics, which will undoubtedly continue to change and improve therapeutic decisions for patients.

Classification of clinical syndromes by bioinformatics molecular profiling can advance the use of *gene* testing in the broadest sense (as a molecular diagnostic tool) in the diagnosis, therapy, and counseling of individuals affected with genetic disorders. For these advances to have real use, there needs to be equally robust phenotypic data that are

meticulously mapped to DNA, RNA, and protein genotype. Databases have been and are being made available for storing, indexing, and querying patient data, clinical phenotypes, and genotypic data; however, an understanding of the types of data to be objectively documented, and the standards to be adopted regarding terminologies and storing data in query-compatible forms, is still evolving. With the proliferation of all these biological databases, tools, and resources, there is a high likelihood of compromised data quality and reliability; thus, caution is the watchword while using these resources judiciously for decision making.

Personalized Medicine

Personalized medicine is the use of information and data from a patient's genotype, or level of gene expression, to stratify disease, select a medication, provide a therapy, or initiate a preventive measure that is particularly suited to that patient at the time of administration. In addition to genetic information, other sources of information, including imaging, laboratory tests, and clinical knowledge about the disease process and the patient play equally important roles.

Translational bioinformatics has emerged as a field that bridges bioinformatics and medical informatics, with the potential of immense benefit for the medical community in reaching the personalized medicine era. This field uses translational tools and techniques to analyze and integrate the data resulting from high-throughput technologies to facilitate smooth translation of important information. This brings the information from bench to bedside by enabling medical providers to incorporate information into their routine medical practice of diagnosis and treatment.

Together, these tools will enable a paradigm shift from genetic medicine—based on the study of individual inherited characteristics, most often single genes—to genomic medicine, which by its nature is comprehensive and focuses on the functions and interactions of multiple genes and gene products, among themselves and with their environment. The information gained from such analyses, in combination with clinical data, is now allowing us to assess individual risks and guide clinical management and decision making, all of which form the basis for genomic medicine.

As medical technology has advanced rapidly over the past century to cure major diseases and discover drugs and therapies, there also has been major variability in therapeutic responses and ensuing side effects. The new insights from studying human diseases from an information science perspective helps in understanding that humans are a system of interconnected and dynamically organized cells, proteins, and genes. The new molecular data have given evidence that the variability in drug response is genetically determined, with age, sex, nutrition, and environmental exposures also playing contributory roles. Thus, classifying patient data among these various parameters and studying genetic distinctions in different subclasses of relevant data, via bioinformatics, will facilitate a more direct route to a patient's wellness and disease prevention than has yet been possible.

Future Direction

Sequencing of the human genome has ushered in prospects for personalized care. There is growing evidence that the practice of medicine might soon have a new toolbox to predict and treat disease more effectively. The Human Genome Project has spawned several important "omic" technologies that allow "whole genome" interrogation of sequence variation ("genomic"), transcription ("transcriptomic"), proteins ("proteomic"), and metabolites ("metabolomic"), which all provide more exacting detail about the disease mechanisms being investigated. In the field of molecular imaging, researchers are developing chemical and biological probes that can sense molecular pathway mechanisms that will allow medical professionals to monitor health and disease on an individual basis.

As genetic and genomic data proliferate from various public and government efforts worldwide, notably in the United States, Europe, and Japan, the push to cull meaningful insights from these mountains of data has also gathered speed, necessitated by seeking cures for elusive diseases and by pharmaceutical companies' desire for breakthroughs in drug discovery. This gold hunt for the perfect molecule and perfect drug target is largely facilitated by bioinformatics tools and technologies employed in the early phases of the drug discovery and development process.

As innovators race toward getting a person's DNA sequenced for $1,000, down from $100 million a decade ago, the field of bioinformatics has paralleled this rapid technology advancement. Leveraging bioinformatics and medical informatics is crucial for giving medicine and medical care a preventive, predictive, and personalized form in the near future.

Banu Gopalan and Petra Platzer

See also Genetic Testing

Further Readings

Augen, J. (2005). *Bioinformatics in the post-genomic era: Genome, transcriptome, proteome, and information-based medicine*. Boston: Addison-Wesley.

Bioinformatics Definition Committee. (2000). *NIH working definition of bioinformatics and computational group*. Retrieved January 23, 2009, from http://www.bisti.nih.gov/docs/CompuBioDef.pdf

Kanehisa, M. (2000). *Post-genome informatics*. Oxford, UK: Oxford University Press.

Kuhn, K. A., Knoll, A., Mewes, H. W., Schwaiger, M., Bode, A., Broy, M., et al. (2008). Informatics and medicine: From molecules to populations. *Methods of Information in Medicine, 47*(4), 283–295.

Liebman, M. N. (2002). Biomedical informatics: The future for drug development. *Drug Discovery Today*, *7*(20 Suppl.), S197–S203.

Maojo, V., & Kulikowski, C. A. (2003). Bioinformatics and medical informatics: Collaborations on the road to genomic medicine? *Journal of the American Medical Informatics Association, 10*(6), 515–522.

Martin-Sanchez, F., Iakovidis, I., Nørager, S., Maojo, V., de Groen, P., Van der Lei, J., et al. (2004). Synergy between medical informatics and bioinformatics: Facilitating genomic medicine for future health care. *Journal of Biomedical Informatics, 37*(1), 30–42.

National Center for Biotechnology Information: http://www.ncbi.nlm.nih.gov

Teufel, A., Krupp, M., Weinmann, A., & Galle, P. R. (2006). Current bioinformatics tools in genomic biomedical research (Review). *International Journal of Molecular Medicine, 17*(6), 967–973.

Valafar, F. (2003). Techniques in bioinformatics and medical informatics. *Annals of the New York Academy of Sciences, 980,* 41–64.

Yan, Q. (2008). The integration of personalized and systems medicine: Bioinformatics support for pharmacogenomics and drug discovery. *Methods in Molecular Biology, 448,* 1–19.

Boolean Algebra and Nodes

A Boolean, or logical, variable is one that can take the values **T** (true) or **F** (false); and the Boolean, or logical, algebra pioneered by George Boole (1815–1864) holds the formal machinery that allows such *truth values* to be logically combined. Boolean principles offer a framework for handling questionnaire and symptom data of the common *binary* kind (yes vs. no, normal ["negative"] vs. abnormal ["positive"], etc.), for clinical decisions, even measurements, probabilities, and so on, often have to be *dichotomized* (*binarized*). Library searches exploit Boolean AND, OR, and NOT, and the digital computer is essentially a huge number of electronic switches (on vs. off) connected in a Boolean manner, marching to the beat of a clock. Boolean principles also underlie logical checking of rule-based decision support systems for inconsistencies, incompleteness, and redundancy.

The Algebra

Let A, B, C,… be diagnostic tests or, more precisely, the Boolean variables that hold the answers to "Did test A come out positive?" and so on. Boolean *negation*, alias NOT, swaps **T** and **F**, indicating, in our example, whether a test came out negative:

$$\neg A = (\text{not } A) = (\text{false if } A \text{ is true; true if } A \text{ is false}) = (\mathbf{F} \text{ if } A, \text{ otherwise } \mathbf{T}).$$

Note that $\neg(\neg A) = A$. Other basic operations are AND and OR:

AND ("Did both A and B come out positive?"):

$$A \wedge B = (A \text{ and } B) = (\mathbf{T} \text{ if both } A \text{ and } B, \text{ otherwise } \mathbf{F});$$

OR ("Did A, B, or both, come out positive?"):

$$A \vee B = (A \text{ or } B) = (\mathbf{F} \text{ if neither } A \text{ nor } B, \text{ otherwise } \mathbf{T}) = \neg(\neg A \wedge \neg B).$$

The mirror image of the rightmost identity also works:

$$A \wedge B = \neg(\neg A \vee \neg B) = (\mathbf{F} \text{ if one or both of } A \text{ and } B \text{ are false, otherwise } \mathbf{T}).$$

Set theory involves set intersection ($\cap$), union ($\cup$), and complementing, which are the precise analogs of $\wedge$, $\vee$, and $\neg$, respectively.

OR and AND are *associative* operations: More than two terms can be ORed (or ANDed), in arbitrary order, to reflect "at least one true" ("all true"). *Distributive* properties include

$$(A \wedge B) \vee (A \wedge C) = A \wedge (B \vee C),$$
$$(A \vee B) \wedge (A \vee C) = A \vee (B \wedge C).$$

The former may be read: To be a "female diabetic or female with hypertension" means to be a "female with diabetes or hypertension." Self-combination:

$$(A \wedge A) = (A \vee A) = A.$$

Finally, test A must be either positive or negative, but cannot be both:

$$(A \vee \neg A) = \mathbf{T}, \quad (A \wedge \neg A) = \mathbf{F}.$$

The former expression is a *tautology*, that is, a necessarily true proposition.

The OR described so far is the inclusive OR, as in the "or both" above. Informatics (checksums, cryptography) makes frequent use of the *exclusive* OR, abbreviated EXOR. *Equivalence* ($\equiv$), in the sense of having the same truth value, and EXOR are each other's negations:

$$(A \equiv B) = (A \text{ and } B \text{ are both true or both false});$$

$$(A \text{ EXOR } B) = \neg(A \equiv B) = (\text{one of } A \text{ and } B \text{ is true, not both}).$$

Now, ((A EXOR B) EXOR C) is true if just one or all three terms are true. Extending this rule to repeated EXORs of multiple Boolean terms, one finds that the result is true if the number of true terms is odd and false if it is even.

The Implication Symbol

The implication symbol ($\rightarrow$) is a treacherous abbreviation:

$$(A \rightarrow U) = (A \text{ "implies" } U) = (U \vee A) = (\text{if } A, \text{ then } U; \text{ otherwise } \mathbf{T}).$$

That is, if A, then the expression reproduces the truth value of U; if not A, then the result is **T**—*regardless* of U!

This is different from "A causes U," however construed. It is also different from the language of decision recommendations and deductions in rule-based decision support systems. When you come across a statement such as "If symptom A is present, diagnosis U is applicable," you take it to be telling you nothing about patients without symptom A. If told that the statement is untrue, you take that to mean that A alone should not trigger label U. This is exactly how a rule-based system would react to cancellation of "If A, then U." The Boolean negation $\neg(A \rightarrow U) = (\neg U \wedge A)$, on the other hand, would claim that, despite symptom A, diagnosis U was not made (in a particular case), or every patient has symptom A and diagnosis U is never made (when read as a general rule).

Physical Analogs

A and B may be gates. When they must be passed one after the other, entry is possible if and only if both gates are open. Symbolically, $E = A \wedge B$. If, on the other hand, A and B are alternative routes of entry, $E = A \vee B$ (Is at least one gate open?). Electrical switches connected "in series" or "in parallel" are analogous.

Boolean Data in Programming: Boolean Nodes

Most programming languages make available a Boolean, or logical, data type. It may look thus:

```
  Boolean L; #declares L to be a Boolean variable#
  L: = (n < 20); #L becomes true or false depending on the value of n#
  if(L)print("n  small"); #to  be  printed only if n is below 20#
```

Programming languages have different ways of writing AND and OR. Often, however, it is convenient to let **T** and **F** be represented by 1 and 0, leading to

$$\neg A = (1 - A),$$

$$A \wedge B \wedge C \wedge \Lambda = ABC \quad [\text{an ordinary product of 0s and 1s}]$$

$$= \min(A, B, C, \ldots),$$

$$A \vee B \vee C \vee \Lambda = \max(A, B, C, \mathrm{K}).$$

Conversely, mathematicians often write $\wedge$ and $\vee$ for min and max; so $x \wedge y = \min(x, y)$, the smaller of x and y. Checksum calculations, involving EXORing of multiple 0 or 1 digits, produce the result 1 precisely when the number of component 1s is odd, as stated above; that is, when their ordinary sum is odd:

$$\mathrm{EXOR}(A, B, C, \mathrm{K}) = (1 \text{ if } A + B + C + \wedge \text{ is odd; otherwise, } 0).$$

Boolean Nodes

A standard decision tree has a stem (clinical problem) and two types of splits: chance nodes and decision nodes. Between the stem and the leaves, each of which represents a possible clinical diary, the tree gets increasingly bushy because there is no way branches can rejoin. In practice, many subtrees are nearly identical and can be drawn, and programmed, just once. They can then be entered from various points as subroutines, with arguments that let them inherit patient features from the calling point.

Numerical arguments allow probabilities and utilities to be context dependent (e.g., age dependent); Boolean arguments allow the subtree to be structurally modified. They govern special *Boolean nodes*, which act as switches. The Boolean argument "Has the patient already had a hemicolectomy?" may govern a switch that blocks a decision branch involving hemicolectomy. Likewise, in modeling an annual screening program, the same annual subtree may have to be entered recursively: A Boolean switch may serve to prevent endless recursion.

Boolean nodes allow compact, minimally redundant trees to be drawn—sometimes at the expense of intelligibility. (Some authors have used the term for other kinds of two-way nodes.)

Boolean Matrices

Consider a graph of N interconnected nodes (*vertices*). An $N \times N$ matrix $C = \{C_{ij}\}$ of Boolean elements may represent the interconnections (*arcs, edges*),

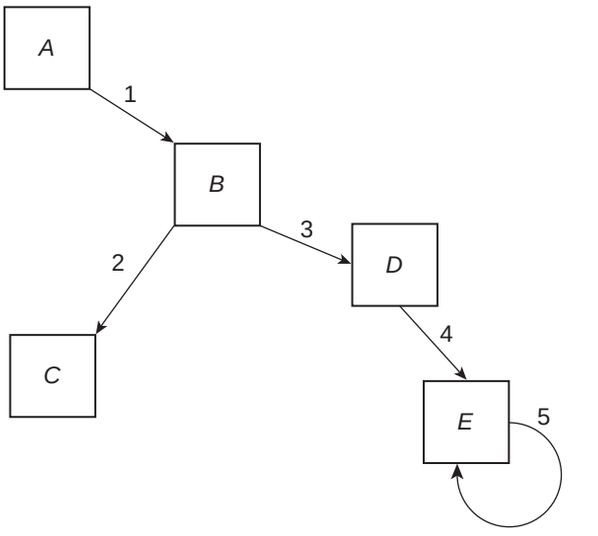


Figure 1 Boolean graph

C_{ij} being **T** if and only if node i is directly connected to node j. Both directed and nondirected graphs may be represented in this way; in the latter case, the matrix is symmetric ($C_{ij} = C_{ji}$). The diagonal elements, C_{ii}, are set to **F** unless self-referring nodes make sense, as they do in state-progression models, including Markov chains. Matrix C is called the *adjacency matrix* of the graph.

The matrix product $D = CC$, with addition and multiplication replaced with OR and AND, now answers the question of whether one can pass in precisely two steps from node i to node k. Readers familiar with matrices will see that

$$D_{ik} = \bigvee_{j = 1, \mathrm{K}, N}\{C_{ij} \wedge C_{jk}\} = (C_{i1} \wedge C_{1k}) \vee (C_{i2} \wedge C_{2k}) \vee \Lambda.$$

$$= (\text{Is there at least one node } j \text{ that can be used as a stepping stone?}).$$

This idea can be elaborated to answer many types of connectedness questions. For example, in the directed graph in Figure 1, the arcs are numbered for convenience, and in Figure 2, the elements in matrix C that go into the calculation of $D_{A,D}$ are shaded.

The Boolean Approach to Diagnostic Tests

The 2^k possible outcomes of k binary tests and the $k!$ possible sequences of execution may render

Adjacency matrix C (edge absent marked with a dot; edge numbers shown for clarity)

To / From	A	B	C	D	E
A	.	T (1)	.	.	.
B	.	.	T (2)	T (3)	.
C	.	.	.	.	.
D	.	.	.	.	T (4)
E	.	.	.	.	T (5)

Matrix D = CC (by aligning the "From A" row with the "To D" column in matrix C, both shaded, one sees that the only match is in the second position, implying that node B is the only stepping stone from A to D, i.e., route 1–3)

To / From	A	B	C	D	E
A	.	.	T (12)	T (13)	.
B	.	.	.	.	T (34)
C	.	.	.	.	.
D	.	.	.	.	T (45)
E	.	.	.	.	T (55)

Figure 2 Boolean tables

ordinary decision trees unmanageable: One ends up with $2^k k!$ leaves, or 48 when $k = 3$ (80 when tests may be executed simultaneously). A shortcut is offered by the following, much simpler, Boolean procedure, supplemented, if necessary, by an ad hoc analysis to find the least costly or risky execution scheme (*flowchart*).

After tabulating the 2^k outcomes and the associated clinical actions found by utility maximization, suppose it turns out that intervention U is recommended when, and only when, Tests A, B, and C have the result patterns: $(+ + +)$, $(+ + -)$, $(- + +)$, or $(- - +)$. The Boolean representation $U = (A \wedge B \wedge C) \vee (A \wedge B \wedge C) \vee (A \wedge B \wedge C) \vee (A \wedge B \wedge C)$ reduces to $U = (A \wedge B) \vee (A \wedge C)$, suggesting that test A should be performed first and, depending on the result, B or C will decide. However, it also reduces to $U = (B \wedge C) \vee ((B \text{ EXOR } C) \wedge (A \equiv B))$, suggesting that one should begin with B and C and, if they disagree, check whether A sides with B.

The former execution scheme is attractive when A is inexpensive and risk-free, the latter when A is costly or risky. Unless the choice is obvious, one must go through all contingencies and calculate expected money and utility costs. These concerns, as well as the constraints that arise when tests are technically intertwined, can be handled in a decision tree, but the answer found by the Boolean procedure is otherwise exactly the one a decision tree would give.

In other words, the two procedures lead to the same *test interpretation scheme*, but the Boolean procedure may require ad hoc supplementary calculations to find the optimal *test execution scheme*. *Warning:* Speaking of Tests A and B being applied in parallel (in series) may refer to the tests being *executed* simultaneously (vs. one after the other). Some authors, thinking of the gate analogy above, therefore use the parallel versus series terminology to characterize two *interpretation* schemes, namely, $U = (A \vee B)$ versus $U = (A \wedge B)$.

As a by-product, the Boolean procedure reveals whether any tests are mandatory (needed in all cases) or redundant (dispensable). In the small artificial example, this did not happen.

The 1986 paper that popularized these techniques also illustrated some geometric features that the ROC diagram will possess when the *lattice* of all Boolean combinations of several binary tests is plotted.

Jørgen Hilden

See also Causal Inference and Diagrams; Diagnostic Tests; Markov Models; Receiver Operating Characteristic (ROC) Curve; Subtrees, Use in Constructing Decision Trees

Further Readings

Boole, G. (1854). *An investigation of the laws of thought.* London: Macmillan.

Garcia-Remesal, M., Maojo, V., Laita, L., Roanes-Lozano, E., & Crespo, J. (2007, August). An algebraic approach to detect logical inconsistencies in medical appropriateness criteria. *Proceedings of the 29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, 1,* 5148–5151.

Glasziou, P., & Hilden, J. (1986). Decision tables and logic in decision analysis. *Medical Decision Making, 6,* 154–160.

Lau, J., Kassirer, J. P., & Pauker, S. G. (1983). Decision Maker 3.0: Improved decision analysis by personal computer. *Medical Decision Making, 3,* 39–43.

Bounded Rationality and Emotions

The rational decision maker with limitless capacities to process information does not exist. People's cognitive and emotional resources are bounded (limited), thereby motivating them to engage in strategies to maximize effective use of these resources. It has been increasingly recognized that deliberation and intuition are two resources essential for adequate decision making. This entry discusses not only rationality and emotions but also the role they play in decision making and the strategies people use to approach decision problems.

Bounded Rationality

In 1957, Herbert Simon proposed the notion of *bounded rationality* to account for the fact that perfectly rational decisions are often not feasible in practice due to the finite information-processing capacities of humans. Simon points out that most people are only partly rational and are in fact emotional or irrational in the remaining part of their actions. Human beings are limited in their capacity for storing and processing information. As a consequence, people process information sequentially and use heuristics, or rules of thumb, to keep the information-processing demands of complex tasks within the bounds of their cognitive capacities. These heuristics are procedures for systematically simplifying the search through the available information. The use of heuristic strategies improves the performance of the individual as a limited information processor and is, as Simon argues, "at the heart of human intelligence." In this vein, Gerd Gigerenzer and colleagues argue that simple alternatives (i.e., heuristics) to a full rational analysis as a mechanism for decision making frequently lead to better decisions than the theoretically optimal procedure. These heuristics are generally successful, but in certain situations they lead to systematic cognitive biases.

Heuristics

People use heuristics for multi-attribute decision problems such as choosing a car or choosing a hospital for treatment as well as for risky decision making such as the choice between an operation and a wait-and-see policy with different mortality risks. Two approaches to the use of heuristics in decision making can be distinguished: the *accuracy/effort approach* and the *heuristics-and-biases approach*. According to the accuracy/effort approach, people process and evaluate only a part of the information and use noncompensatory decision rules to limit the information-processing demands of multi-attribute decision problems. For instance, people eliminate options because of an unsatisfactory score on one attribute, as, for example, when choosing among cars, a decision maker eliminates all options above a certain price, irrespective of the evaluation of the other attributes. John Payne and colleagues see humans as adaptive decision makers who weight the benefits of a decision strategy (i.e., the probability that a strategy will select the best alternative) against the costs (i.e., the mental effort, time, and money needed).

Another approach is the heuristics-and-biases approach. This approach is most prominently represented by the research of Amos Tversky and Daniel Kahneman and emphasizes the biases and errors in human judgment that are due to the use of heuristics. Contrary to the earlier approach, it does not consider the use of heuristics as a rational trade-off between accuracy and effort but as a failure to recognize the "correct" solution. Tversky and Kahneman consider these heuristics as highly economical and usually effective but add that in some cases they may lead to systematic and predictable errors. An example of such a heuristic is the availability heuristic: Objects or events are judged as frequent and probable or causally efficacious to the extent that they are readily available in memory. This heuristic is likely to erroneously affect the evaluation of information whenever some aspect in the environment is made disproportionally salient or available to the perceiver.

These two approaches are both related to the limited information-processing capacities or the

bounded rationality of people. The heuristics-and-biases approach focuses on the first stage of the decision process, that is, the editing phase, which occurs in a more or less automatic way. The accuracy/effort approach is concerned with the stages after the initial coding process, which seem more controlled. According to this approach, people resort to the use of heuristics if an analytical approach to decision making becomes too demanding. A more complete understanding of decision making should include the accuracy/effort as well as the heuristics-and-biases approach, as decision behavior is likely to consist of multiple systems that interact in various ways.

Dual-Processing Theories

It is an old idea in psychology that human processing of information takes place on many levels that can operate simultaneously and relatively independently. These dual-process models of human reasoning and decision making have become more popular in the past decade or so. In a recent article, Kahneman relates the heuristics-and-biases research to the dual-processing theories about reasoning and decision making. This ancient idea that cognitive processes can be distinguished into two main categories, roughly corresponding to the everyday concepts of intuition and reason, is now widely embraced under the general label of dual-process theories. These two categories of cognitive processes can be distinguished by their speed, their controllability, and the contents on which they operate. The dual-process models distinguish cognitive operations that are quick and associative from others that are slow and governed by rules. Intuitive or System 1 thinking is closely related to perception and quickly proposes intuitive answers to judgment problems as they arise. Operations are fast, automatic, associative, and effortless, and they are often emotionally charged. They are also governed by habit and are therefore difficult to control or modify. Deliberative or System 2 reasoning monitors the quality of these proposals, which it may endorse, correct, or override. The processes of System 2 are slower, serial, effortful, and deliberatively controlled. They are also relatively flexible and potentially rule governed. A characteristic of System 2 is that it is limited or bounded by working memory capacity and is assumed to be linked to general intelligence, while System 1 functions independently of working memory. These characteristics that have been attributed to the two modes are related to consciousness (unconscious and holistic vs. conscious and analytic) and functionality (e.g., associative, automatic, and parallel vs. rule-based, logical, and sequential). In a recent article, Jonathan Evans gives an overview of the several dual-processing theories of reasoning, judgment, and social cognition.

Emotions: Immediate Emotions

Emotional processing, although not included in all dual-processing theories, is placed in System 1 rather than in System 2. In several theories, a fast emotional basis for decision making is contrasted with a slower and more deliberative cognitive basis. The emotional side of judgment and decision making has recently received more attention in judgment and decision research. Research shows that every stimulus evokes affective evaluation that is not always conscious. *Affective valence* is a natural assessment and can, according to Paul Slovic and colleagues, be used as a heuristic attribute for making complex decisions. They propose that representations of objects and events in people's minds are tagged, to varying degrees, with affect. When making a judgment or a decision, people consult or refer to an "affect pool" containing all the positive and negative tags consciously or unconsciously associated with the representations. Affect may serve as a cue to judgments in the same way as availability.

Slovic and colleagues argue that the affect heuristic guides the perception of risk and benefits in the sense that a positive affect generalizes to other aspects of the activity or technology. Thus, when benefits of a technology are seen as high, this positive affective evaluation generalizes to a positive evaluation of the risk associated with this technology, that is, to a lower perceived risk. Conversely, technologies with low perceived individual benefits are associated with higher perceived risks. The affect heuristic may also work with other heuristics. Slovic and colleagues suggest that the availability heuristic may work not only through ease of recall or imaginability but also because remembered images are associated with

affect. As is the case with other heuristics, the affect heuristics, presented by Slovic and colleagues as the centerpiece of experiential or System 1 thinking, may also have drawbacks leading to erroneous judgments. When emotional responses to risky situations (e.g., worry, fear) diverge from cognitive evaluations, these may have a greater impact on risk-taking behavior than do cognitive evaluations. Risks of positive-valued activities, such as fast driving (for some people), may be underestimated, while negative-valued activities, such as flying in airplanes, may lead to an overestimation of risks. When emotions are more intensive, they can even overwhelm deliberative decision making altogether. Some people experience intense fear when they think about flying in airplanes, even though they recognize that the risks are low.

Immediate emotions can have a direct effect, as is the case in the affect heuristic in which affect is used as information for judgment, or an indirect effect. The indirect influence of immediate emotions occurs by influencing people's judgments of expected consequences and their emotional reactions to these outcomes. For instance, when people are not hungry or not in pain, they underappreciate what it will feel like to be hungry or in pain. Furthermore, immediate emotions can bias the interpretation of information in such a way that decision makers selectively attend to and retrieve emotionally relevant information. Studies have found that negative emotions narrow attentional focus, while positive emotions broaden attentional focus. Negative emotions are also found to trigger more systematic processing than positive emotions. One explanation given by George Loewenstein and Jennifer Lerner for this is that negative emotions alert the individual to the possibility that something is wrong and action has to be taken. Happiness or positive mood may have the meaning that everything is all right and, therefore, may lead to more heuristic processing. It has been found, for instance, that happiness increased reliance on stereotypes, which indicates a categorical, holistic way of processing information rather than an analytical way.

Emotions: Anticipatory Emotions

The effect of immediate emotions on decision making should be distinguished from anticipatory emotions. People often compare the consequences of their decisions with what could have happened under different circumstances, which results in counterfactual emotions. One of these emotions is anticipatory regret that results from the comparison between the outcome one will experience as a consequence of a decision and the outcome one would experience if one were to choose differently. For instance, women may choose to attend public health screening for breast cancer in spite of very low chances of having breast cancer because not going may result in strong negative feelings in case they might have a cancer that would then be detected much later. Anticipatory emotions may also explain the finding that patients having to make decisions that are emotionally charged, such as whether to go for prenatal testing or have an operation, usually do not take into account the probabilities. An explanation of people's lack of responsiveness to probabilities is that anticipatory emotions arise as reactions to mental images of the outcome of a decision. Such images are discrete and not very much affected by probabilities. The impact of the image of having a probability of 1 out of 50 of carrying a child with Down syndrome may be the same as the impact of the image of having a probability of 1 out of 500. The decision of women to opt for or against prenatal testing may therefore be more influenced by how they feel about having a child with Down syndrome than by the probabilities. These anticipatory emotions are partly cognitive, in the sense that people may think of them consciously and take them into account when weighing the pros and cons of several options. On the other hand, as they are affect laden, they are part of System 1 thinking and largely intuitive.

Danielle R. M. Timmermans

See also Bias; Dual-Process Theory; Emotion and Choice; Heuristics; Intuition Versus Analysis

Further Readings

Evans, J. St. B. T. (2008). Dual-processing accounts of reasoning, judgment and social cognition. *Annual Review of Psychology, 59*, 255–278.

Gigerenzer, G., & Todd, P. M. (1999). *Simple heuristics that make us smart*. Oxford, UK: Oxford University Press.

Kahneman, D. (2003). A perspective on judgment and choice. Mapping bounded rationality. *American Psychologists, 58,* 697–720.

Loewenstein, G., & Lerner, J. S. (2003). The role of affect in decision making. In R. J. Davidson, K. R. Scherer, & H. H. Goldsmith (Eds.), *Handbook of affective sciences* (pp. 619–642). Oxford, UK: Oxford University Press.

Payne, J. W., Bettman, J. R., & Johnson, E. J. (1992). Behavioral decision research: A constructive processing perspective. *Annual Review of Psychology, 43,* 87–131.

Slovic, P., Finucane, M. L., Peters, E., & MacGregor, D. G. (2002). The affect heuristic. In T. Gilovic, D. Griffin, & D. Kahneman (Eds.), *Heuristics and biases: The psychology of intuitive judgment* (pp. 397–420). Cambridge, UK: Cambridge University Press.

Brier Scores

Brier scores are used to assess the precision of probability predictions. For an event that can only occur in a set of mutually exclusive categories, the Brier score is the sum of the squared differences between the predicted probabilities that the event will occur in a specific category (numbers in the interval from 0 to 1) and the observed outcomes (1 if the event occurs in a specific category, 0 otherwise). Brier scores were originally proposed as a means to describe the precision of probabilistic weather forecasts (e.g., categories "rain," "no rain"). Here, they were appreciated because they allow for a finer assessment of a forecaster's ability to generate accurate predictions than mere counts of numbers of correct predictions. For the same reason, Brier scores have been proposed in the medical context as an alternative to receiver operating characteristic (ROC) methods in diagnostic testing for the calibration of the quality of medical decision makers, for tuning statistical prediction rules, and for the assessment of predictions in survival analysis.

General

Consider an event that can only occur in one of r distinct categories. For example, a medical decision maker might assign probabilities to each category, where the probabilities should sum up to 1. Let $\pi_j \in [0,1]$ denote the prediction for the probability that the event occurs in category j, for $j = 1,\ldots, r$, and let Y_j denote the random outcome, where $Y_j = 1$ if the event occurs in category j and $Y_j = 0$ if it does not. The Brier score is a loss function that has been proposed as a measure to quantify the loss incurred if π is predicted, and Y is the outcome. It is the squared difference $(\pi - Y)^2$. In a sample of size n where π_{ij} and y_{ij} are the ith prediction and the ith actually observed outcome for category j, respectively, the *empirical Brier score* is given by $(1/n)\sum_i \sum_j (\pi_{ij} - y_{ij})^2$. For example, when the events "relapse" versus "no relapse" are of interest, and there are two patients, the first with a relapse and the second without, a naive predicted probability of .5 for both patients results in a Brier score of $1/2 \times (((.5 - 0)^2 + (.5 - 1)^2) + ((.5 - 1)^2 + (.5 - 0)^2)) = .5$. If, however, for the patient with relapse, the predicted probability of relapse is .6, and for the patient without relapse the predicted probability of relapse is .3, then the Brier score reduces to $1/2 \times (((.4 - 0)^2 + (.6 - 1)^2) + ((.7 - 1)^2 + (.3 - 0)^2)) = .25$.

Brier Scores With Dichotomous Data

In a setting with dichotomous data, that is, with only $r = 2$ categories (e.g., when one predicts whether a patient will survive for a certain period of time), it is common to consider only one of the categories for calculation (dropping the subscript j). While in the original formulation, the Brier score takes values in the range from 0 to 2 when there are only 2 categories, the modified version ranges from 0 to 1; that is, the resulting value is only half of the original Brier score.

Brier scores can generally be applied to predictions π_i with $0 \leq \pi_i \leq 1$. These predictions may have been derived from a careful statistical model-building process. They can, however, stem from diverse sources and might also, for example, constitute a summary of expert guesses. For calculating the Brier score, predicted probabilities are needed. When only classifications are available and these are taken as predicted probabilities, so that all π_i are 0 or 1, the squared differences take only the values 0 and 1, and the Brier score is the proportion of observations where classification and outcome are identical. Therefore, in these cases, the empirical Brier score coincides with the misclassification rate.

Properties

The Brier score is a *strictly proper scoring rule*. This means that, when prediction error is quantified by the Brier score, the best predictor is the true probability of the event: Let $p = P(Y = 1)$; then $E(\pi - Y)^2$ attains its unique minimum for $\pi = p$. This can be seen by the following decomposition of the expected Brier score, $E(\pi - Y)^2 = E(p - \pi)^2 + E(p - Y)^2$, which means that inaccuracy (measured by the Brier score) can be split into imprecision and inseparability. Using the Brier score to judge prediction error, thus, forces forecasters to give their best probabilistic predictions, not just classifications. However, imprecision and inseparability primarily are theoretical quantities that cannot be measured directly.

Comparison With ROCs

Another popular technique for judging the performance of medical decision makers is ROC curves. These are obtained from the predicted probabilities π_i by using a cutoff for arriving at actual predictions and then varying this cutoff. For each value of the cutoff, the proportion of correctly classified observations with outcome $y_i = 1$ is recorded and plotted against the proportion of wrongly classified observations with $y_i = 0$ for that cutoff. The area under the resulting ROC curve is an indicator for the performance of the decision maker, with larger values indicating better performance. However, since ROC curves are a rank-based method, two decision makers who rank the observations in the same order will have identical ROC curves, even if their actual predicted probabilities differed. Since it has been argued that the predicted probabilities might be even more important than the actual classification in medical settings, a technique for evaluating performance should be more focused on the former, and therefore, the Brier score should be preferred over ROC curves.

Calibration of Medical Decision Makers

Given predicted probabilities from decision makers, feedback should not only be given on accuracy, that is, on how effective objects could be assigned to the correct category (using a cutoff on the probabilities), but also on precision, that is, on how close the predicted probabilities are to the true probabilities. While the misclassification rate only gives information on accuracy, the Brier score also considers precision. For the dichotomous setting, it is a sum of a measure of imprecision (a property of the decision maker) and a measure of inseparability (a property of the situation at hand). The difference of the Brier scores for two decision makers in the same situation therefore gives their difference in precision.

There are several decompositions of the Brier score that result in explicit values for precision, which in this context is also called reliability or calibration. These decompositions, therefore, also provide for estimates of the theoretical quantities of imprecision and inseparability. Similar to the procedure used for constructing calibration plots, the predictions are grouped by the value of the predicted probabilities; that is, predictions with the same or similar predicted probability are combined into groups j, where $j = 1, \ldots, J$, for which the predicted probability is d_j and the proportion of events is p_j. With n_j being the number of observations in group j, the Brier score can be decomposed into a reliability component $(1/n)\sum_j n_j (d_j - p_j)^2$ and a resolution component $(1/n)\sum_j n_j p_j (1 - p_j)$. The former indicates how close the predicted probabilities are to the true probabilities (with smaller values indicating better calibration of the decision maker), while the latter indicates the ability of the decision maker to sort the observations into categories such that the proportions of outcomes p_j are maximally diverse. From this decomposition, it can again be seen that the Brier score takes its minimum value when the true probabilities are used as predictions.

Assume that there are 20 patients, where a predicted probability of relapse is wanted and that 8 of these patients actually suffer a relapse. If the predicted probability is .5 for all patients, there is only one group ($J = 1$) with predicted probability $d_1 = .5$ and proportion of events $p_1 = .5$. The reliability component therefore is $1/20 \times 20 \times (.5 - .4)^2 = .01$, which seems to be very good. However, the resolution, which is $1/20 \times 20 \times .4 \times .6 = .24$, is rather poor, resulting in a Brier score of .25. If, in contrast, a decision maker provides two predicted probabilities, .6 for a group of 10 patients, where 5 have an event, and .2 for the remaining 10 patients, where 3 suffer relapse, the value of the

reliability component will also be equal to .01, that is, very good. However, the resolution now is $1/20 \times ((10 \times .5 \times .5) + (10 \times .3 \times .7)) = .23$, resulting in a Brier score of 0.24, indicating better overall prediction performance.

Reliability, either by explicitly reporting its value or by means of calibration plots, which graph p_j against d_j, has often been used for giving feedback to decision makers, that is, for improving calibration, while resolution seems to have been neglected. In addition, there exist several other decompositions that allow for detailed analysis of decision-maker performance.

Evaluation of Statistical Prediction Rules

Besides analyzing the performance of decision makers, the Brier score can also be used for evaluating statistical prediction rules. The basis for the latter is formed by statistical models, which are typically built from some training data, where, in addition to the outcome for each observation, a set of informative variables is given (e.g., "age" and the concentration of some biomarker for patients, for which survival up to some time is the event of interest). Many statistical models can not only provide classification for new observations (given the information from the variables), thus comprising prediction rules, but also predicted probabilities. Using a cutoff for classification on the latter, performance could be judged by misclassification rate on new data. However, many statistical models require choice of some tuning parameters, which should be selected to maximize performance. Optimizing by means of misclassification rate may easily result in overfitting; that is, use of tuning parameters results in more complexity than is supported by the data, as this criterion is most sensitive to the fit of a statistical model for observations with large ambiguity (i.e., which have true probabilities close to the classification cutoff). When it is expected, for example, that the cutoff for classification might be changed later, the Brier score is a more reasonable criterion for selecting tuning parameters, as it is sensitive to the model fit for all observations, regardless of their true probability. Therefore, it is also more appropriate if interpretation of the prediction rule is wanted, as the structure of the fitted model will be equally valid for all observations.

Brier Scores With Survival Data

In survival analysis, the outcome of interest Y is the time until a specific event (e.g., death) occurs. Here, probability predictions often refer to the survival status $Y(t^*)$ at a specific time t^*, $Y(t^*) = 0$ meaning dead/event occurred, $Y(t^*) = 1$, alive/event not yet occurred at t^*. Let $\pi(t^*)$ denote the prediction for the survival status at t^*.

Brier Score at a Specific Time and Integrated Over Time

The *Brier score at* t^* is $[\pi(t^*) - Y(t^*)]^2$, if survival status $\pi(t^*)$ is predicted and $Y(t^*)$ is the outcome. When predictions are to be assessed over a period from time 0 to time t^* rather than for one specific time t^*, the prediction error at t can be averaged over this interval, yielding the *integrated Brier score,* $\int_0^{t^*}[\pi(t) - Y(t)]^2 dW(t)$, where $W(t)$ is a suitable weight function, for example, t/t^*. For a sample of size n, the empirical versions are given by $(1/n)\sum_i[\pi_i(t^*) - y_i(t^*)]^2$ and $(1/n)\sum_i\int_0^{t^*}[\pi_i(t) - y_i(t)]^2 dW(t)$, respectively.

Censoring

In studies of survival time data or time-to-event data, a common problem called *censoring* occurs when some, but not all, individuals can be followed up until death (or until the event of interest occurs). This may happen for many reasons, for example, when a medical study on mortality is evaluated before all patients have died. In that case, the outcome data are (Y, δ), where Y denotes the observation time when the individual was observed to survive, and δ is the event indicator containing the censoring information: $\delta = 1$ indicates that death was observed after Y time units, whereas $\delta = 0$ means that the individual was observed to survive for Y time units before it was censored, so that the exact time of death is unknown.

Empirical Brier scores can be devised to estimate the true expected Brier score even in the presence of censoring, which, however, needs to be accounted for. To this end, the individual contributions to the empirical Brier score are weighted according to the censoring information. Thus, the empirical Brier score at t^* in the presence of censoring is $(1/n)\sum_i w_i(t^*)\,[\pi_i(t^*) - y_i(t^*)]^2$, where $w_i(t^*)$ is the weight for individual i. In the simplest case, where censoring

can be assumed not to depend on the individual's survival chances, the weights incorporate the Kaplan-Meier estimator G of the censoring or potential follow-up distribution, which is obtained from $(Y, 1 - \delta)$ by exchanging the roles of censored and uncensored observations. Then $w_i(t^*) = \delta_i/G(Y_i)$ if $y_i \leq t^*$, and $w_i(t^*) = 1/G(t^*)$ if $y_i > t^*$. With this approach, individuals whose survival status at t^* is unknown due to censoring receive weight 0 (these individuals contribute indirectly to the empirical Brier score, because they are used in the calculation of G). Individuals whose survival status at t^* is known receive weights >1, so that they represent the contributions of individuals whose Brier score is unobservable in addition to their own contribution. Again, to average prediction error at t over an interval from 0 to t^*, an integrated version of the empirical Brier score can be used: $(1/n)\sum_i \int_0^{t^*} w_i(t)[\pi_i(t) - y_i(t)]^2 dW(t)$.

Dynamic Predictions

Similar techniques can be used to devise empirical Brier scores when the predictions for a survival status are updated as time progresses. Such updated predictions can arise, for example, when physicians' probability estimates of survival are updated during daily morning rounds or from joint statistical models of longitudinal biomarkers and survival data. Here, the survival status at time t^* is predicted at time s with $0 \leq s < t^*$ by the probabilistic predictor $\pi_i(s;t^*)$.

Evaluation of Statistical Prediction Rules

Similar to the selection of tuning parameters for statistical prediction rules with a categorical outcome, the Brier score can also be used for model complexity selection for survival models. When the empirical version of the Brier score (with proper weights) is obtained for a range of times and plotted against time, this results in prediction error curves. The tuning parameter of a statistical prediction rule should then be chosen such that the area under this curve, that is, the integrated Brier score, is minimal. However, as with a categorical outcome, the empirical version of the Brier score should not be calculated from the data that the prediction rule was fitted to. One can, for example, set aside a test set, but this has the disadvantage of losing observations for the fitting of the prediction rule. An attractive alternative is provided by the bootstrap procedure, where the drawing of new data sets is imitated by randomly drawing observations from the original data set. The statistical prediction rule is then fitted to each of these bootstrap data sets, and the empirical version of the Brier score is calculated separately for each one based on the observations that are not in the respective bootstrap data set. The final prediction error curve estimate is then obtained from the averaged Brier score over all bootstrap samples. This can then be used not only for selecting tuning parameters for one statistical prediction rule but also to compare the prediction performance of several prediction rules.

Harald Binder and Erika Graf

See also Calibration; Diagnostic Tests; Kaplan-Meier Analysis; Prediction Rules and Modeling; **Receiver Operating Characteristic (ROC) Curve; Survival Analysis**

Further Readings

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review, 78,* 1–3.

Gerds, T. A., & Schumacher, M. (2007). Efron-type measures of prediction error for survival analysis. *Biometrics, 63,* 1283–1287.

Graf, E., Schmoor, C., Sauerbrei, W., & Schumacher, M. (1999). Assessment and comparison of prognostic classification schemes for survival data. *Statistics in Medicine, 18,* 2529–2545.

Hilden, J., Habbema, J. D. F., & Bjerregard, D. (1978). The measurement of performance in probabilistic diagnosis III: Methods based on continuous functions of diagnostic probabilities. *Methods of Information in Medicine, 17,* 238–246.

Schoop, R., Graf, E., & Schumacher, M. (2007). Quantifying the predictive performance of prognostic models for censored survival data with time-dependent covariates. *Biometrics, 64*(2), 603–610.

Yates, J. F. (1982). External correspondence: Decompositions of the mean probability score. *Organizational Behavior and Human Performance, 30,* 132–156.

C

CALIBRATION

Calibration refers to the degree of correspondence between probabilities and observed relative frequencies. Suppose that when a patient is admitted to an intensive care unit (ICU), a physician assesses the probability that the patient will survive until hospital discharge. If data on these probabilities and the resulting outcomes (survival or death) are collected for a large number of patients, the data can be organized by the numerical probabilities. For example, the relative frequency of survival among all the patients for whom a probability of 70% was assessed can be determined. If this relative frequency is 70%, and the relative frequencies for other probability values also match those probabilities, the probabilities are said to be perfectly calibrated. If the probabilities and their associated relative frequencies differ, the probabilities are miscalibrated, with the degree of miscalibration increasing as the differences increase.

The relevance of calibration relates to the use of probabilities in decision making. Medical decisions are typically made under uncertainty, such as the uncertainty about whether a surgical procedure will be successful if performed on a particular patient. The likelihood of success quantifies this uncertainty and should be a key factor in the decision about whether to perform the surgery. Thus, the calibration of the numerical probability assessed for success is relevant. If the relative frequency of success is only 40% among patients for whom a probability of success of 70% had been assessed, a decision based on a probability of 70% could be suboptimal.

Measuring Calibration

Measures of calibration are based on pairs of probability values p_i and the corresponding relative frequencies $r_i = 100(f_i/n_i)$, where n_i is the number of times the probability value p_i is used and f_i is the number of times the event occurs when the probability is p_i. For example, if the probability of survival is assessed to be 70% for 100 of the patients who are admitted to an ICU, and 68 of those patients survive, $r_i = 100(68/100)$, or 68%. If there are m probability values $p_1, \ldots, p_m$, there will be m pairs (p_i, r_i).

Calibration is often studied graphically, through a plot of r_i as a function of p_i. This plot is called a *calibration diagram*, and an example of such a plot is shown in Figure 1. Here the values of p_i (expressed in percentages) are 0, 10, 20,…, 100, and each square represents a pair (p_i, r_i). If the probabilities were perfectly calibrated, the squares would all be on the line from (0, 0) to (100, 100). Of course, a statistical variation in r_i given p_i is likely to cause some deviation from this perfect-calibration line, and such deviations will tend to be larger for small values of n_i (small samples with probability value p_i).

The calibration diagram shown in Figure 1 demonstrates good calibration, although it does reflect a bit of a tendency for r_i to be greater than p_i for lower probability values and to be less than p_i for higher probability values. This tendency is

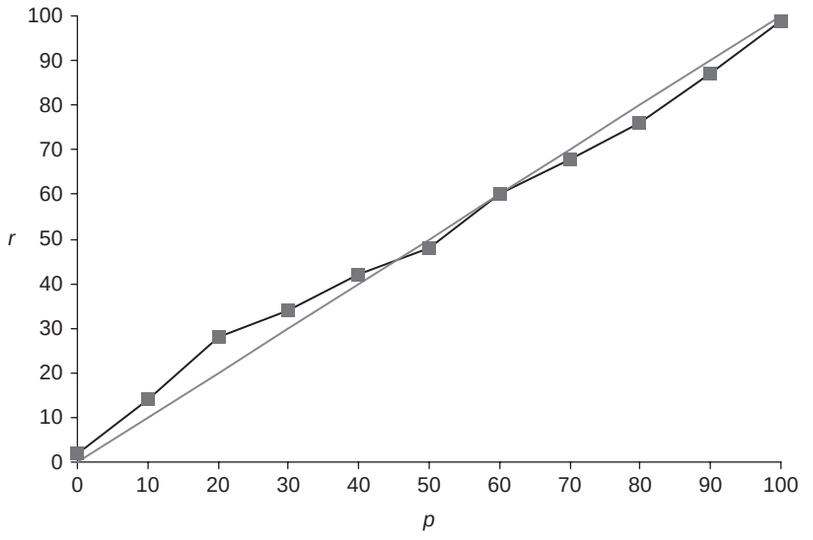


Figure 1 A calibration diagram

often called *overconfidence* because the probabilities are more extreme (lower than the relative frequencies for low probabilities and higher for high probabilities) than justified by the data. Many empirical studies of calibration display overconfidence to varying degrees, whereas other studies demonstrate better calibration.

Calibration diagrams are quite informative, providing at a glance an indication of how far the data points are from the perfect-calibration line, which points are more distant from the line, and whether they are above or below the line. Summary measures of overall calibration are also used, just as summary statistics such as a mean are used in addition to histograms in statistics. The most common summary measure is the weighted average of the squared differences between the probabilities and relative frequencies, with the weights proportional to the sample sizes for the different data points:

$$C = \sum_{i=1}^{m} w_i (p_i - r_i)^2, \text{with } w_i = n_i / \sum_{j=1}^{m} n_j.$$

Here $(p_i - r_i)^2$ is a measure of the calibration of the probability value p_i, and the *calibration score* C is a weighted average of these calibration measures for the m probability values. A lower value of C indicates better calibration, with perfect calibration corresponding to $C = 0$.

A measure such as C is especially useful in comparing the calibration of probabilities from different sources. Probabilities in medical decision making are often subjective in nature, being assessed by experts such as physicians. It can be informative to compare the calibration of probabilities from different physicians or from different groups of physicians. Probabilities can also be generated through models or from past data, in which case comparisons between physicians' subjective probabilities and model-based or data-based probabilities are possible. The methods used to measure calibration can be used regardless of the source of the probabilities.

Calibration and Sharpness: The Evaluation of Probabilities

What characteristics are of interest in the evaluation of probabilities? Calibration is a characteristic that often receives much attention. But it is not the whole story.

It is possible for probabilities to be very well-calibrated but not very informative. For example, suppose that the overall success rate of a surgical procedure at a given hospital is 80% and that the rate has remained quite steady over the past several years. To aid in decisions about whether to perform this procedure on particular patients, a

physician assesses the probability of success for each patient. Information from a patient's records and an examination of the patient should be helpful in assessing the probability for that patient. But if good calibration is the only measure of how good the probabilities are, the physician could just assess a probability of 80% for each patient and be confident of good calibration. Here 80% is the *base rate*, and the expertise of the physician should make it possible to distinguish between patients for whom the surgery is more likely to be successful and patients for whom the surgery is less likely to be successful. A base rate forecast ignores this expertise and is uninformative in the sense of not distinguishing among different patients.

In this example, what would be "perfect" probabilities? A physician would be perfect in predicting the outcomes if the assessed probability was 100% for all patients who then had successful surgery and 0% for all patients with unsuccessful surgery. Of course, this is an ideal that is not likely to be achieved in practice. Nonetheless, it provides a benchmark indicating the most informative probabilities, just as base rate probabilities provide a benchmark indicating relatively uninformative probabilities. To the extent that a set of probabilities can move away from the base rate toward the ideal of perfect forecasts, the probabilities are considered more accurate.

A class of measures called *scoring rules* has been developed to measure the accuracy of probabilities. The most frequently used scoring rule is a quadratic scoring rule, sometimes called the Brier score. For the surgery example, let p be expressed in percentage terms and let the outcome, e, of the surgery for a particular patient be coded as 100 if it is successful and 0 if it is not successful. Then the quadratic score for that patient is a squared-error function: $Q = (p - e)^2$. A lower score is better, with the best possible score being 0 for a perfect probability.

Letting e_{ij} denote the outcome (0 or 100) for the jth patient among the n_i patients for whom the probability, p_i, was assessed, the average quadratic score across all patients is

$$Q = \sum_{i=1}^{m} \sum_{j=1}^{n_i} (p_i - e_{ij})^2.$$

This average score can be decomposed into two terms and written as

$$Q = \sum_{i=1}^{m} w_i r_i (100 - r_i) + \sum_{i=1}^{m} w_i (p_i - r_i)^2 = S + C.$$

Here C is the calibration score defined earlier, and $S = \sum_{i}^{m} w_i r_i (100 - r_i)$ is a measure of the *sharpness* of the relative frequencies, with a lower S indicating greater sharpness. The term $r_i(100 - r_i)$ relates to the sharpness of the relative frequency corresponding to the probability value p_i, and the *sharpness score, S,* is a weighted average of these sharpness measures for the m probability values. The best possible Q is 0, and less-than-perfect sharpness ($S > 0$) or calibration ($C > 0$) lead to a worse score.

To understand the sharpness and calibration terms better, think about the assessment of a probability of successful surgery for a patient as if it were separated into two steps. First, the patient is classified into a "bin" with other patients perceived to have roughly the same likelihood of successful surgery. Then a label is assigned to each bin in the form of a probability number. Suppose that some patients are put into a bin with probability value 80%, which means that each of them is judged to have an 80% probability of successful surgery. If they all undergo the surgery, with success for 74% of them, the calibration measure for the bin is $(80 - 74)^2 = 36$, and the sharpness measure is $74(100 - 74) = 1924$. If we calculated these measures for all m bins and then took weighted averages, we would get S and C, from which we could find $Q = S + C$. The sharpness, S, is related to how discriminatory the bins are and not to the probability values; note that S does not depend on the p_i values. The calibration, C, on the other hand, has to do with how consistent the probability values are with the relative frequencies. In other words, the sharpness has to do with the effectiveness of the separation of patients into bins, and the calibration has to do with the fidelity of the bin labels (the probability numbers) to the data.

A scoring rule such as the quadratic score, then, measures overall accuracy, taking into account both the sharpness and calibration. If just a single bin is used, with a base rate probability assigned to all cases, the calibration should be excellent but the sharpness weak. If the separation into bins is very effective but the labeling of the bins is poor, the sharpness can be excellent while the calibration

is poor. An extreme example of the latter occurs when all patients are given probabilities of 0 or 100 but those with a probability of 0 have successful surgery and those with a probability of 100 have unsuccessful surgery. Sharpness measures the true discriminatory power of the division into bins, but poor calibration can render that power ineffective by causing poor decisions if the probability labels are taken at face value. For the extreme example just given, those assigned probability labels of 100 would most likely go ahead with the surgery, only to see it fail, whereas those assigned probability labels of 0 would avoid the surgery, not knowing that it would be successful if performed.

If probabilities are quite sharp but poorly calibrated, perhaps their calibration can be improved. Training through relevant experience and feedback might help an individual improve calibration. Alternatively, a decision maker can recalibrate probabilities as deemed appropriate. Essentially, this amounts to relabeling the bins. If past data on the probabilities from a particular physician indicate a tendency toward overconfidence, future probabilities might be adjusted, making low probabilities a bit higher and high probabilities a bit lower, in an attempt to calibrate the physician's probabilities. The difficulty is that the decision maker may not be aware of the degree of miscalibration.

A goal to strive for in probability assessment is to make the probabilities as sharp as possible while still maintaining good calibration. The sharpness indicates the true discriminatory power of the probabilities, and the calibration guarantees that this power can be used appropriately by decision makers. In a sense, sharpness is more important than calibration because it is possible to try to improve calibration, as noted above. Improvements in sharpness are more difficult, requiring additional information (e.g., more tests) or greater effort in understanding the implications of the existing information; they cannot be gained by mere relabeling. Nonetheless, miscalibrated probabilities can send misleading messages to decision makers, so striving for good calibration as well as sharpness is desirable.

Robert L. Winkler

See also Brier Scores; Expert Opinion; Judgment; Subjective Probability; Uncertainty in Medical Decisions

Further Readings

Budescu, D. V., & Du, N. (2007). Coherence and consistency of investors' probability judgments. *Management Science, 53,* 1731–1744.

Lichtenstein, S., Fischhoff, B., & Phillips, L. D. (1982). Calibration of probabilities: The state of the art to 1980. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 306–334). Cambridge, UK: Cambridge University Press.

McClelland, A. G. R., & Bolger, F. (1994). The calibration of subjective probabilities: Theories and models 1980–94. In G. Wright & P. Ayton (Eds.), *Subjective probability*. Chichester, UK: Wiley.

Murphy, A. H., & Winkler, R. L. (1977). Reliability of subjective forecasts of precipitation and temperature. *Applied Statistics, 26,* 41–47.

O'Hagan, A., Buck, C. E., Daneshkhah, A., Eiser, J. R., Garthwaite, P. H., Jenkinson, D. J., et al. (2006). *Uncertain judgements: Eliciting experts' probabilities*. Chichester, UK: Wiley.

Sox, H., Blatt, M. A., Higgins, M. C., & Marton, K. I. (2006). *Medical decision making*. Philadelphia: American College of Physicians.

von Winterfeldt, D., & Edwards, W. (1986). *Decision analysis and behavioral research*. Cambridge, UK: Cambridge University Press.

Winkler, R. L. (1996). Scoring rules and the evaluation of probabilities. *Test, 5,* 1–60.

Winkler, R. L., & Poses, R. M. (1993). Evaluating and combining physicians' probabilities of survival in an intensive care unit. *Management Science, 39,* 1526–1543.

Yates, J. F. (1990). *Judgment and decision making*. Englewood Cliffs, NJ: Prentice Hall.

Case Control

Case-control studies are a nonexperimental form of medical research that informs cause-effect relationships. Their main purpose is the identification of risk factors for events of interest. Most famously, case-control studies provided the first evidence of a strong association between cigarette smoking and lung cancer. However, findings from a number of recent case-control studies have been subsequently contradicted or found to overestimate the strength of relationships compared with more robust epidemiological study designs. An example

is the case-control finding that hormone replacement therapy (HRT) had a protective effect against coronary heart disease, following which randomized trial evidence identified a small increased risk associated with HRT.

Case-control studies identify cases as patients who already have a disease or condition of interest, and then attempt to identify characteristics of these patients that differ from those who do not have the condition of interest (the controls). For a defined exposure (e.g., walking alongside golf courses) and a defined outcome (e.g., experience of head injury), Table 1, a 2 × 2 table, represents hypothetical findings and informs analysis of an odds ratio.

The odds are expressed as the ratio of the probability that the event of interest occurs to the probability that it does not. In the example, the probability that a golf course walker experiences a head injury is 10/100 or .1, and the probability that he or she does not suffer such an injury is 90/100 or .9. The odds are therefore 10/90, or .11 (the number of events divided by the number of nonevents). The corresponding odds for non–golf course walkers are 5/150, or .03.

The odds ratio is estimated as the odds in the case group divided by the odds in the control group, that is, .11/.03 or 3.67 in the hypothetical golf course example. This is interpreted as golf course walkers being at more than 5 times the odds of suffering a head injury compared with non–golf course walkers.

Traditional case-control studies only inform estimates of the odds ratio between exposure states; they do not enable the estimation of absolute or relative risk because the full size of the population from which the cases (with and without the exposure(s) of interest) are drawn cannot be estimated in a straight case-control study.

Table 1 Hypothetical 2 × 2 table

	Cases (Head Injury)	*Controls (No Head Injury)*
Exposed (walk by golf course)	10	90
Nonexposed (do not walk by golf course)	5	150

Accounting for Bias

The strength and interpretation of identified relationships is first dependent on a study's ability to match the cases and controls, such that both groups can be defined as random samples from the same underlying population. A second significant issue in the application of case-control studies is the accurate identification of the existence or absence of all potentially relevant factors. The exclusion of factors that are associated with both included exposures and the outcome of interest may introduce a bias in the association estimates due to confounding. A third form of bias is labeled *recall bias* and may occur when the outcome acts as a stimulus to aid the recall of the experience or timing of exposures in cases, which tends to inflate risk estimates in case-control studies.

To illustrate these issues, the study of the safety effects of bicycle helmets is used. The representation of the issues is necessarily brief, and the interested reader is referred to a lively discussion in the journal *Accident Analysis and Prevention*. Case-control studies in this area have generally defined cases as persons experiencing head injuries following a bicycle accident. Control groups have included random samples from a population of bicyclists, as well as patients presenting at an emergency department with nonhead injuries sustained following a bicycle accident. Selecting controls from the full population of bicyclists reflects a random sample from the same underlying population from which the cases were drawn and so avoids selection bias. However, such a control group may be subject to both confounding and recall bias. Confounding may occur if cyclists who wore helmets were generally more careful riders than nonwearers (and therefore less likely to experience a bicycle accident), and so more careful riders would be overrepresented in the control group. If this was the case, then one would want to control for riding care in the analysis. A potential solution could involve the elicitation of risk-taking characteristics from the cases and controls, so as to control for differences in the data analysis. Recall bias may not be perceived as a significant problem but would occur if the controls were less likely to accurately recall their use of a helmet.

The nonrandom selection of controls as individuals presenting with nonhead injuries was the

more common approach. As the research question specifies the effect of helmets on reducing head injuries following bicycle accidents, the population of interest is cyclists who crashed, and this is a reasonable approach if hospital presentation by controls is not related to helmet wearing. The use of hospital-based controls may also reduce the effects of recall bias as all respondents have a similar stimulus (hospital visit) to aid recollection. It also may reduce the impact of differential risk-taking characteristics between controls and cases as a confounding factor as both groups were hurt sufficiently to seek medical care. However, risk differences may remain: For example, cases and controls may have differing distributions of cycle speed at the point of accident.

There are various methods available to control for possible confounding factors (assuming they can be identified and measured). In the bicycle example, analysis can be restricted to accidents occurring at high speed (or low speed) only. Alternatively, cases and controls could be matched with respect to accident speed. Both these approaches require the specification of sufficiently narrow speed categories that represent important differences. A more flexible approach is to use regression analyses to statistically adjust the risk ratio of interest to capture the effect of potential confounders.

Nested Case-Control Studies

An adaptation to the traditional case-control study design is the nested case-control study, which involves applying a case-control study within the confines of an established cohort. Cohort studies follow individuals over time to observe outcome(s) of interest as they occur, with exposure status being defined at the beginning of the study (i.e., prospectively). Advantages of cohort studies (over case-control studies) include the fact that all individuals in the study analysis are automatically derived from the same population (the cohort) and that there is no uncertainty around the time sequence of the exposure preceding the outcome in the establishment of a cause-effect relationship.

A nested case-control study selects cases on the basis of events occurring (either prospectively or retrospectively). A risk set is defined for each case that includes individuals at risk of the event at the time of the observed case (on the cohort time axis) and may include some matching criteria. One or more controls are then randomly selected from the defined risk set for each case.

The advantages of this study design include the natural satisfaction of the requirement that controls are randomly sampled from the same population within which the cases occurred (selection bias) and the fact that data on all individuals in the cohort are more easily obtained (recall bias). The nested approach also enables the estimation of relative risks, as well as odds ratios.

Establishing Causal Relationships

Evidence on the existence of a causal relationship between an exposure and an outcome is required to directly inform public health decisions and the design of clinical interventions. The likelihood of confounding can never be completely eliminated (even in a randomized trial), particularly so in case-control studies, and so the presentation and interpretation of study results should always be accompanied by an open and explicit assessment of the probability that results may be confounded and of the direction and size of any confounding bias.

If an association is adequately demonstrated by a case-control study, the next step is to assess whether the observed association is likely to be causal. A range of criteria have been proposed for testing the existence of a causal relationship:

Detailed Review of Potential Confounders

A detailed and explicit consideration, involving literature reviews, of factors that could be related to the exposure and outcome will increase the credibility of proposed causal relationships.

Temporal Relationship

The exposure will always occur before the outcome in a case-control study, but if the outcome develops too soon after the exposure, then the likelihood of causality is reduced.

Size of Odds Ratio

A higher odds ratio (greater than 1) or a lower odds ratio (less than 1) is, ceteris paribus,

indicative of a greater probability of causation. However, levels of uncertainty should also be considered when interpreting the size of the mean odds ratio.

Etiological Plausibility

Further evidence is provided if the observed odds ratio and hypothesized causal relationship is supported by existing knowledge around the pathway to the outcome.

Repeated Findings

Similar findings of significant differences between cases and controls in alternative populations increase the chances that an association is causal.

Dose-Response Relationship

Further evidence of causality is provided if a consistent trend showing the odds ratio increases or decreases with increasing or decreasing levels of exposure.

Application to Technology Assessment

In the context of intervention evaluation studies, case-control studies have a limited role in the evaluation of therapeutic interventions as such studies generally evaluate homogeneous populations within which interventions have the same expected effect. There may be some scope for case-control studies to inform downstream effects in decision model-based evaluations; for example, post–disease recurrence pathways may be related to factors observed between treatment initiation and point of recurrence. In breast cancer, the likelihood of progression to metastases following locoregional recurrence is influenced by the duration of the prior disease-free interval.

Case-control studies have much greater potential in secondary evaluations of preventive and screening interventions. Such evaluations describe the pathway of full populations with respect to a disease, and often an important component is the description of separate pathways for different risk groups within an aggregate population. For example, a screening program for anal cancer in homosexual men might differentiate between HIV-negative and HIV-positive men.

Advantages and Disadvantages

The main advantage of case-control studies is that they can be undertaken at relatively low cost and within a shorter time frame than other prospective study designs, particularly around outcomes that are rare. However, case-control studies are generally less reliable than either randomized controlled trials or cohort studies, and causal relationships are often difficult to establish. The results of case-control studies are most often used to generate hypotheses that can be tested using more robust study designs.

Jonathan Karnon

See also Attributable Risk; Bias in Scientific Studies; Causal Inference and Diagrams; Causal Inference in Medical Decision Making; Confounding and Effect Modulation; Odds and Odds Ratio, Risk Ratio

Further Readings

Cummings, P., Rivara, F. P., Thompson, D. C., & Thompson, R. S. (2006). Misconceptions regarding case-control studies of bicycle helmets and head injury. *Accident Analysis and Prevention, 38,* 636–643.

Curnow, W. J. (2006). The Cochrane Collaboration and bicycle helmets. *Accident Analysis and Prevention, 37,* 569–573.

Essebag, V., Genest, J., Suissa, S., & Pilote, L. (2003). The nested case-control study in cardiology. *American Heart Journal, 146*(4), 581–590.

Koepsell, T. D., & Weiss, N. S. (2003). *Epidemiologic methods: Studying the occurrence of illness* (pp. 105–108, 247–280, 374–402). New York: Oxford University Press.

Lawlor, D. A., Smith, G. D., & Ebrahim, S. (2004). The hormone replacement-coronary heart disease conundrum: Is this the death of observational epidemiology? *International Journal of Epidemiology, 33,* 464–467.

Rothman, K. J., & Greenland, S. (1998). *Modern epidemiology* (2nd ed., pp. 62, 93–161, 255–259). Philadelphia: Lippincott-Raven.

Thompson, D. C., Rivara, F. P., & Thompson, R. (2004). Helmets for preventing head and facial injuries in bicyclists (Cochrane Review). In *The Cochrane Library* (Issue 2). Chichester, UK: Wiley.

Causal Inference and Diagrams

Causal inference is the science of attributing a particular outcome (or effect) to one or more particular causes. In addition to concluding that there is an association between two variables, causal inference implies that the effect is the direct result of a measurable cause. In medical research, the cause is often an intervention or treatment, and the outcome is often a disease or complication. Outcomes from those receiving the intervention, perhaps a particular drug, are often compared with those of a control group. When the difference in outcomes between the experimental and control groups is attributed to the intervention, causal inference is being made.

Causal inference is made most cleanly in a randomized, blinded study. However, even in a nonrandomized setting, some degree of qualified causal inference may be possible. This depends on the extent of thorough understanding of the relationships involved, careful design, and data collection and analysis. Causal inference relationships can be visualized and clarified using causal diagrams—modern tools that use arrows to visualize the purported relationships between causal variables, outcome variables, and confounding variables in both randomized and nonrandomized studies.

Randomized Studies

In a randomized research study, each subject is randomly assigned to receive one of the interventions to be compared. At randomization, but before receiving intervention, randomized groups are very similar to each other with respect to baseline predictors of outcome, the only systematic difference being the assigned intervention. Unless the process of randomization has been systematically altered, other baseline differences would be due to chance.

Thus, in a properly conducted randomized study, there is no selection bias or treatment assignment bias; neither patients nor doctors choose which intervention an individual will receive. Because a *confounder* is a variable that is associated with both intervention and outcome, and because there is usually no association between treatment assignment and baseline predictors of outcome in a randomized study, confounding does not usually exist. Differences in outcome between randomized groups are correctly interpreted as cause-effect.

Fundamental Problem of Causal Inference

Causal inference is a missing-data problem. It has its basis in individuals, not group averages. Let Y^1 and Y^0 represent an individual's potential (or hypothetical) response on treatment and control, respectively. An individual causal effect is defined as the difference between these two potential outcomes at the same point in time, or $\delta = Y^1 - Y^0$. The average of the individual causal effects, or the average causal effect (ACE), can be written as $E[Y^1 - Y^0] = E[\delta]$, where E is the expectation sign, indicating the average of all subjects. Causal effects may well differ across individuals.

However, individual causal effects are never observable because more than one intervention cannot be independently given to the same individual at the same time. In a parallel-group randomized study, each patient receives only one intervention, either treatment or control, and so the outcome is observed for only one of the potential outcomes for that patient. Causal inference is thus a huge missing data problem, in which half of the data for each individual is unobserved. How, then, can causal inference be made? This is the Fundamental Problem of Causal Inference.

Average Causal Effect

While individual causal effects cannot be observed, the average of the individual causal effects, the ACE, is estimated in a randomized study. If the individual causal effects were observable, it is mathematically true that the average of the individual differences (i.e., causal effects) would equal the difference in average response for treatment and control, ignoring individuals, such that $E[\delta] = E[Y^1 - Y^0] = E[Y^1] - E[Y^0]$. Thus, although no patient receives both treatments, in a randomized study, researchers can estimate the ACE from the difference in mean outcome between the treatment and control groups. This is true because treatment assignment is independent of potential confounders, and as a result, patients in

the two groups are very similar, on average, on factors that might affect outcome. When randomized groups are compared on the outcome, the estimated difference between groups estimates the ACE for *individuals*.

Nonrandomized Studies

Researchers sometimes strive to make causal inference from nonrandomized studies in which patients have received either one or another of two interventions. The major problem with achieving this goal is selection bias since the treatment assignment has not been random and the groups to be compared likely differ on variables (other than the treatment) that cause the outcome of interest. Selection bias results in confounding, or distortion, of the causal effect of interest. In some situations, it may not be possible to conduct a randomized study due to time, resources, or ethics, making the option of causal inference in a nonrandomized study appealing.

Because it is quite difficult to make causal inference in a nonrandomized study, the traditional practice in biomedical research has been to *not* try to make causal inference. Instead of a cause-effect relationship, researchers have typically made inferences about the *association* between an intervention and an outcome in nonrandomized studies. For example, researchers might conclude that patients taking Drug A during surgery are less likely to have a postoperative complication than are patients taking Drug B. They would be less likely to conclude that Drug A *causes* a reduction in outcome compared with Drug B, and rightly so. However, special methods are available to attempt some degree of valid causal inference in nonrandomized studies.

Correlation Does Not Imply Causation

It is critical to remember that *correlation* does not imply causation. Other than a true cause-effect relationship, there are four main reasons why a statistically significant result might be obtained in a nonrandomized study: chance (random error), bias (systematic error), effect-cause relationship, and confounding. To entertain causal inference, each of these four reasons must be considered and ruled out to the best of a researcher's ability.

Random error and bias can lead to spurious findings that do not represent true effects in the population. Random error may occur as a result of measurement error or from the variability inherent in sampling (i.e., Type I error). Significant results due to systematic bias may result from off-target measurements by the observer, the instrument, or the patient. These errors can also occur in randomized studies.

Effect-cause and confounding are based on true effects, but they are not cause-effect. A positive or negative association between an exposure and an outcome might represent a true *effect-cause* relationship, instead of *cause-effect*. For example, researchers might conclude that maintaining deep anesthesia (vs. light) causes poor intraoperative and postoperative outcome for patients, when in truth patients who are already developing complications intraoperatively are the ones who require (or "cause") deeper anesthesia to keep them stable during surgery.

Finally, confounding by one or more variables might explain the association between the exposure and outcome, where a confounder is a variable associated with both. In confounding, a third factor (e.g., smoking) is a cause of the outcome (e.g., cancer) and also of the exposure (e.g., coffee drinking), resulting in an association between coffee drinking and cancer that is real but not causal. Confounding is often due to the unavoidable selection bias or treatment assignment bias in nonrandomized studies. Patients are likely to differ on variables responsible for them being in one treatment group versus the other. Since some of these variables are also likely to be related to the outcome of interest, the treatment effect of interest is confounded, or distorted, by these baseline variables unless addressed in the design or analysis phase.

Adjusting for Confounding in Design Stage

In the design phase of a nonrandomized study, confounding can be tackled by narrowing the inclusion criteria to focus on certain level(s) of a confounder, or by matching. In matching, nonexposed patients may be chosen to be very similar to the exposed patients on important confounding variables (e.g., age, sex, body mass index). Alternatively, case-control designs might match

diseased and nondiseased on important confounders. Matching can be done on individual patients (1:1 or 1:*k*) or by choosing patients so that distributions are similar (i.e., frequency matching). However, it is logistically difficult to match on a large number of confounders. To the extent that all confounders are accounted for, and explanations other than cause-effect have been ruled out, some degree of cause-effect relationship may be inferred when matched groups are compared on outcome.

Adjusting for Confounding in Analysis Stage

In the analysis phase, confounding can be addressed by stratification, multivariable regression models, and propensity score methods. More complex methods, including instrumental variable analysis and structural equation models, may sometimes be used when the above are not adequate or feasible. A common concern is whether all confounding variables are known and observed.

Stratification analysis consists of estimating the relationship of interest between a hypothesized cause and an outcome within levels of a variable believed to confound the relationship. For example, if patients are selected to receive one intervention versus another based on baseline severity, analysis comparing intervention and outcome could be done within each level of baseline severity, and results averaged across levels.

Multivariable regression statistically adjusts for confounding variables by including them in the statistical model used to assess the relationship between the intervention and outcome.

Propensity score analysis is becoming mainstreamed as one of the best ways to remove confounding via selection bias in nonrandomized studies. First, a logistic regression model predicting treatment assignment from available baseline potential confounders is used to assign each patient a score representing the probability that he or she would receive treatment (vs. control). Intervention and control patients are then compared on the outcome(s) of interest after adjusting for the propensity scores through stratification, matching, or weighting.

There is a growing body of literature and practice using special statistical methods where researchers are sometimes able to legitimately make some degree of causal inference in nonrandomized studies. The extent to which causal inference is justified in a nonrandomized study depends on the nature and knowledge of the research question, design of the study, knowledge and availability of all true confounding variables, quality of the available data, and skill with which analytical methods are employed.

Causal Diagrams

A causal diagram is a concise way to explore and explain causal relationships among variables. It was popularized by Judea Pearl as a method of displaying adjustment for a third variable or set of variables (Z) to allow causal inference between a cause (X) and effect (Y). Confounding of a causal relationship of interest and the correct (or incorrect) adjustment for confounding can be visualized in a causal diagram. Causal diagrams depend on subject matter experts to guide the plausibility of the postulated relationships. The most common and basic causal diagram is called a directed acyclic graph, or DAG.

Directed Acyclic Graph

A DAG is a graphical description of causal relationships among variables. Many different statistical models would fit any particular DAG. A DAG consists of vertices or nodes representing variables, edges connecting some of the variables, and arrows indicating the direction of relationships among variables. An edge marked by a single arrow is "directed" and indicates the direction of the causal relationship. For example, $X \rightarrow Y \rightarrow Z$ implies that X causes Y, Y causes Z, and X causes Z only through its effect on Y. Variables that are not directly connected in the DAG are assumed to not be *causally* related. *Acyclic* indicates that the DAG does not allow representation of mutual causation or feedback processes such as $X \rightarrow Y$, $Y \rightarrow X$. Each causal relationship can only go in one direction in a DAG, and, typically, only causal relationships are of interest.

Conditioning on variables that block backdoor paths from causal variables to outcome is the first and most widely used strategy to adjust for confounding. A backdoor path is a connected set of variables going from the cause to the effect of interest through an indirect route. A backdoor

path indirectly connects the exposure and outcome, and includes at least one variable pointing to (i.e., causing) the exposure and one to the outcome.

For example, researchers may want to compare two anesthetic regimens for a particular type of surgery on a postoperative complication outcome by analyzing a patient registry. Patients with higher baseline severity are more likely to receive regimen A than B, and severity is also a strong indicator of outcome apart from its effect on treatment assignment. The causal diagram in Figure 1 displays the relationships between severity, regimen, and outcome, with backdoor path $D \leftarrow C \rightarrow Y$.

Severity (C) confounds the relationship between treatment (D) and outcome (Y) because it is a cause of both D and Y. The confounding effect of C could be removed by conditioning on C in one of several ways. In stratification, the $D \rightarrow Y$ effect of interest would be estimated within levels of C and then averaged. Multivariable regression could account for the $C \rightarrow Y$ relationship when assessing $D \rightarrow$ Y and would simultaneously compare levels of D at the average value of C. Finally, propensity score analysis could be used to match patients who did and did not receive D on the predicted probability that they would receive D, based on severity. Each of these methods would enhance the validity of making a causal inference between D and Y by conditioning on C.

When conditioning strategies are not feasible, instrumental variable (IV) techniques can sometimes be used. An IV has a causal effect on outcome Y only through its effect on the treatment variable D, as in the DAG IV $\rightarrow D \rightarrow Y$. Such variables are rare since variables responsible for treatment assignment are often related to outcome through additional causal paths. A true instrumental variable can be used to estimate the causal effect of D on Y by assessing IV $\rightarrow D$ and IV $\rightarrow Y$ and then using the ratio between these two effects to isolate the relationship between D and Y.

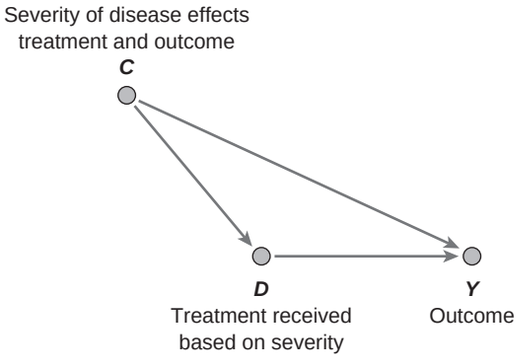


Figure 1 Directed acyclic graph showing confounding of D-Y causal effect by C

Finally, structural equation models can establish an isolated and exhaustive mechanism that relates the causal variable to the outcome and then calculate the causal effect as it propagates through the mechanism.

Using a DAG to Decide Which Variables to Condition On

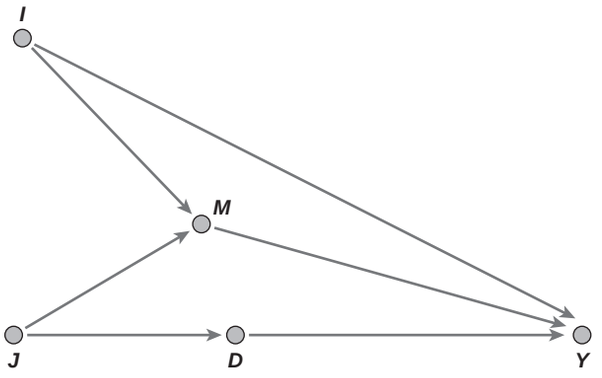
Causal diagrams are useful in deciding which variables need to be adjusted for, to remove confounding in the cause-effect relationship of interest. First, all arrows emanating *away from* the exposure of interest are removed. Then, if there remains a backdoor path connecting the exposure and outcome, one adjusts for variables on the path. Variables not on a backdoor path do not need to be adjusted for.

A *D-separation* criterion, also called *blocking*, was introduced by Pearl as a method to determine if confounding has been removed from a causal pathway in a DAG. A set of variables (or nodes) Z is said to D-separate a set of confounders X from Y if and only if Z blocks every path from a node in X to a node in Y. D-separation requires either causal chains or causal forks or both. Z must not include so-called collider variables (see Figure 2), which can unblock a backdoor path by adjusting for them, thus introducing confounding.

Causal chains ($i \rightarrow m \rightarrow j$) and causal forks ($i \leftarrow m \rightarrow j$) show D-separation because the two extreme variables are marginally dependent but become independent of each other, or blocked, once the researchers condition on (i.e., adjust for) the middle variable, m. In a causal chain, after conditioning on m, the potential confounder, i, has no effect on the probability of outcome, j.

Figure 1 is an example of a causal fork. Adjusting for severity of disease, the middle variable in the fork, removes the confounding on the relationship between the treatment and outcome. The relationship between treatment and outcome can then be estimated free of confounding by m.

Inverted forks ($i \rightarrow m \leftarrow j$), which include colliders, act the opposite way (Figure 2). A collider is



M is a collider. Conditioning on it will create confounding between ***I*** and/or ***J*** with the ***D*-*Y*** relationship by unblocking the backdoor path ***D*←*J*→*M*←*I*→*Y***

Figure 2 Conditioning on a collider variable

variable *m* on a backdoor path that is caused by two or more known or unknown variables. If the extremes, *i* and *j*, are independent, they will *become* dependent once *m* or any of its descendents is conditioned on. This dependence will create confounding if, for example, *i* is also a cause of the outcome, *Y*, and *j* is also a cause of the exposure, *D*. So if a backdoor path between the cause and effect of interest includes a collider variable, confounding will be introduced if researchers adjust for *it alone*.

Consider the potential causal relationship between anesthetic technique *D* and postoperative complication *Y*. Suppose covariables severity of disease, *I*, and surgeon, *J*, are causes of an intraoperative variable *M* (blood loss), which is also a cause of outcome *Y*. Furthermore, suppose the variables are related as in Figure 2, with *J* also causing *D* and *I* also causing *Y*. There would be no confounding if we adjust for all three variables—*I*, *J*, and *M*—since that would completely block the backdoor path between *D* and *Y*. Also, there would be no confounding if we adjust for none of the three variables, since *I* and *J* are independent. However, confounding would be introduced if we *only* adjust for the intraoperative variable, *M*, since that would introduce dependence between surgeon (*J*), a cause of *D*, and severity of disease (*I*), a cause of *Y*. This underappreciated problem due to colliders often surfaces when the variable *M* is observable, such as at baseline measurement of outcome, but there exist unobserved variables (*I* and *J*) causing *M* and also related to the exposure and the outcome.

In analyses attempting causal inference in the nonrandomized setting, a crucial limitation is that all confounding can usually not be accounted for because variables are either unknown or unavailable. In nonrandomized studies, causal inference can only be attempted with this important qualification.

Ed Mascha

See also Bias in Scientific Studies; Causal Inference in Medical Decision Making; Conditional Independence; Confounding and Effect Modulation; Counterfactual Thinking; Propensity Scores; Randomized Clinical Trials

Further Readings

Angrist, J. D., Imbens, G. W., & Rubin, D. B. (1996). Identification of causal effects using instrumental variables. *Journal of the American Statistical Association, 91,* 444–455.

Blackstone, E. (2002). Comparing apples and oranges. *Journal of Thoracic and Cardiovascular Surgery, 123,* 8–15.

Cox, D. R. (1992). Causality: Some statistical aspects. *Journal of the Royal Statistical Society A, 155,* 291–301.

Dawid, A. P. (2000). Causal inference without counterfactuals (with discussion). *Journal of the American Statistical Association, 95*, 407.

Hulley, S., Cummings, S., Browner, W., Grady, D., & Newman, T. (2007). *Designing clinical research*. Philadelphia: Lippincott Williams & Wilkins.

Morgan, S., & Winship, C. (2007). *Counterfactuals and causal inference: Methods and principles for social research*. Cambridge, UK: Cambridge University Press.

Pearl, J. (1995). Causal diagrams for empirical research. *Biometrika, 82,* 669–688.

Pearl, J. (2000). *Causality: Models, reasoning and inference*. Cambridge, UK: Cambridge University Press.

Rosenbaum, P., & Rubin, D. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika, 70,* 41–55.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology, 66,* 688–701.

Causal Inference in Medical Decision Making

One of the most important tasks of decision analysts is to derive causal interpretations, on both the level of decision modeling and the level of statistical analyses of original data sets. Usually, an intervention, action, strategy, or risk factor profile is modeled to have a "causal effect" on one or more model parameters (e.g., probability, rate, or mean) of an outcome such as morbidity, mortality, quality of life, or any other outcome.

This entry introduces the key concepts of causal inference in medical decision making and explains the related concepts such as counterfactuals, causal graphs, and causal models and links them to well-known concepts of confounding. Finally, two examples are used to illustrate causal inference modeling for exposures and treatments.

Background

Decision analyses on risk factor interventions frequently include parameters derived from clinical or epidemiologic studies such as single relative risks or multivariate risk prediction functions (e.g., Framingham risk index for coronary heart disease, cancer risk scores, osteoporosis score). When applied in a decision model, changes in risk factors are then translated to causal effects on the risk of a disease or other outcome in the model. Thus, the causal interpretation of the modeling results strongly depends on the causal interpretation of each modeled risk factor. Therefore, this entry has a strong focus on epidemiologic modeling, which yields the parameters for the decision model.

Study Designs

The gold standard design to evaluate causal effects is the randomized controlled clinical trial. However, most decision models include (at least some) parameters or risk functions derived from epidemiologic (i.e., observational) studies, which have the potential for confounding. It is, therefore, crucial that all model parameters derived from epidemiologic studies be properly adjusted for confounding if one wants to use the results to derive causal interpretations.

Confounding

Definition of Confounding

Time-Independent Confounding

Standard textbook definitions of confounding and methods to control for confounding refer to independent risk factors for the outcome that are associated with the risk factor of interest but are not an intermediate step in the pathway from the risk factor to disease.

Time-Dependent Confounding

The more complicated (but probably not less common) case of time-dependent confounding refers to variables that may vary over time and simultaneously act as confounders (e.g., common cause of both exposure and disease) and intermediate steps (on the causal pathway from exposure to disease). In other words, confounder and exposure of interest mutually affect each other. For example, in a model evaluating the effect of weight loss on the risk of coronary heart disease, physical activity could be a time-dependent confounder because it is an independent risk factor for coronary heart disease, it influences weight, and it can also be influenced by weight.

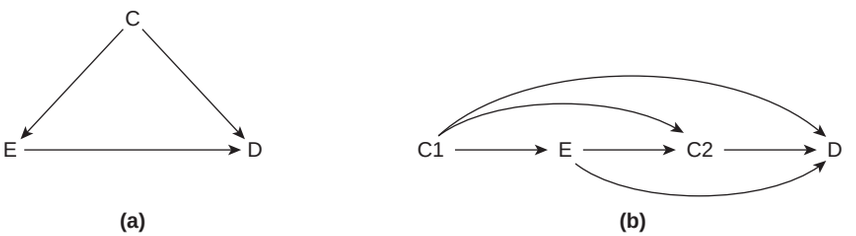


Figure 1 (a) Time-independent confounding and (b) time-dependent confounding

Control for Confounding

Traditional textbook techniques to control for time-independent confounding include restriction, stratification, matching, and multivariate regression analysis. However, these methods have been criticized for being inadequate to control for time-dependent confounding. Other methods such as g-computation, marginal structural models, or structural nested models have been suggested as approaches to this problem.

Relevant Questions

To do a proper causal analysis, one must answer three questions:

1. Which a priori assumptions can be made about the causal relationships between the variables of an epidemiological study?
2. Under these assumptions, are the observed data sufficient to control for confounding?
3. What methods are appropriate to control for confounding?

Causal graphs can guide us in answering these questions.

Causal Graphs

Causal diagrams have a long history of informal application. More recently, formal concepts and rules have been developed for the use and interpretation of causal graphs, for example, in expert systems and operational research. Causal graphs can guide the process of identification of variables that must be measured and considered in the analysis to obtain unbiased (unconfounded) effect estimates. In their milestone paper "Causal Diagrams for Epidemiologic Research," published in 1999 in the journal *Epidemiology*, Sander Greenland, Judea Pearl, and James M. Robins provide an introduction to these developments and their use in epidemiologic research.

Use of Directed Acyclic Graphs in Epidemiology and Medical Decision Making

Directed acyclic graphs (DAGs) are a specific form of causal graph that can be used to understand and explicitly state causal a priori assumptions about the underlying biological mechanisms. DAGs consist of a set of nodes and directed links (arrows) that connect certain pairs of nodes. In medical decision-making research and epidemiology, nodes are used to represent variables, and arrows denote causal relationships. A set of formal and precise graphical rules and assumptions for DAGs has been developed, including a graphical method called d-separation, the causal Markov assumption, and a graphically oriented definition of confounding named the backdoor criterion. These methods allow researchers to determine

- whether they can estimate an unbiased effect from the observed data,
- which variables must be adjusted for in the analysis, and
- which statistical methods can be used to obtain unbiased causal effects.

Specific Applications of Directed Acyclic Graphs

Besides helping with the questions mentioned above, DAGs offer a readily accessible approach to understanding complex statistical issues, including the fallibility of estimating direct effects (i.e., controlling for intermediate steps), the rationale for instrumental variables, and controlling for compliance in randomized clinical trials (when both "intention to treat" and "per protocol" analyses can fail to yield the true causal intervention effect).

Key Lessons Learned From Causal Graphs

There are several lessons to be learned from causal graph theory. In particular, applying the formal rules of DAGs, one can derive four key messages.

> *Key message 1:* Controlling for nonconfounders can induce severe bias in any direction.

The second lesson follows directly from message 1.

> *Key message 2:* The selection of confounders must be based on a priori causal assumptions.

Further messages follow.

> *Key message 3:* Estimating direct effects (i.e., controlling for a known intermediate step variable) can be problematic.

As traditional regression analysis can either control for a variable or not, it cannot appropriately adjust for confounders that are simultaneously affected by the intervention or risk factor of interest (i.e., time-dependent confounding). This leads to the last message.

Key message 4: Traditional adjustment methods (e.g., stratification or multivariate regression analysis) may fail to control for time-dependent confounding.

The following section provides some cases.

Quantitative Models of Causal Inference

Counterfactual Principle

Whereas causal graphs allow for deducting qualitative information about causal effects, medical decision making usually needs quantitative results to inform decisions. One type of quantitative model originating with Neyman and Fisher in the early 20th century is the counterfactual model. In such a model, an association is defined as causal when it is believed that, had the cause been altered, the effect would have changed as well. This definition relies on the so-called counterfactual principle, that is, what would have happened if, contrary to the fact, the risk factor or intervention had been something other than what it actually was.

Classes of Quantitative Causal Models

There are several classes of quantitative models for causal inference that are able to deal with both time-independent and time-dependent confounding. The following model classes are the most recent innovations and increasingly used in medical decision making and epidemiology:

- inverse probability of treatment weighting (marginal structured models),
- g-estimation, and
- parametric g-formula.

In the case of time-dependent confounding, all three of these methods require longitudinal data. This is not—as has been erroneously mentioned—a weakness of these modeling techniques. It is rather quite obvious that disentangling the causality of the feedback loop between the intervention of interest and the time-dependent confounder (which is a cause and effect of the intervention of interest) requires repeated measurements of the same variable. Hence, it is due to the inherent causal nature that observational data with time-dependent confounding can only be solved with longitudinal data.

All these techniques are quite complex and require special programming. The following paragraphs give an overview of how the key approaches differ between these models.

Inverse Probability of Treatment Weighting (Marginal Structured Models)

The imbalance regarding the confounder variable in each of the treatment (or exposure) categories is resolved in the following way: The technique of inverse probability of treatment weighting creates a pseudo population (i.e., counterfactual population); that is, for each subject receiving treatment, another (counterfactual) subject that does not receive the treatment but has the same properties regarding the past variable history is added to the data set. This weighting procedure yields a balanced (unconfounded) data set, and the crude effect estimates derived from this data set represent causal effects. In the presence of time-depending confounding, this step is repeated for each repeated measurement time.

g-Estimation

This approach is based on the assumption of no unmeasured confounding. Under this assumption, the outcome is independent of the exposure, given the past history of covariables. The g-estimation procedure is started with assuming a mathematical model. Then the model parameters are systematically varied in a grid search until the outcome is in fact independent of the exposure in the data set. The final values of the parameters are the ones with a causal interpretation.

Parametric g-Formula

The data set is divided into intervals. For each interval, traditional multivariate regression analysis can be performed (separately or pooled) controlling for the past history of variables but not including future measurements into the regression model. Subsequently, simulation techniques (e.g., Monte Carlo simulation) can be used to simulate the overall effect of one exposure or treatment versus another based on the parametrized regression equations.

Examples

Causal Analysis of Risk Factors: The Causal Effect of Public Health Interventions on the Risk of Coronary Heart Disease

Background

The World Health Organization (WHO) has established a project on comparative risk assessment for coronary heart disease (CHD) that evaluates the

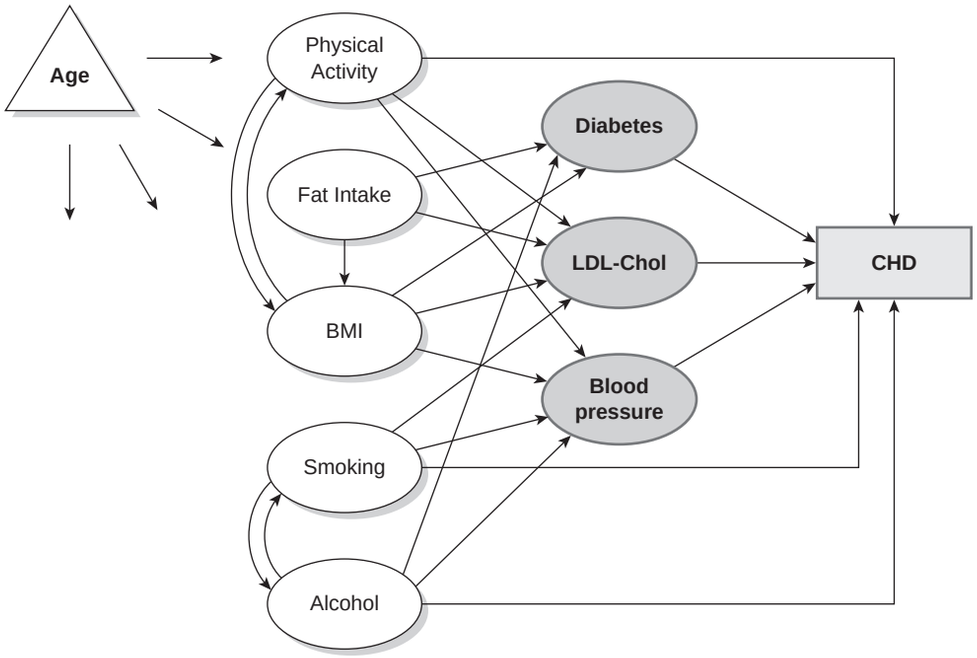


Figure 2 Causal diagram for CHD

overall impact of several public health interventions on the risk of CHD. The causal diagram for CHD was defined by a WHO panel of epidemiologists. This diagram represents the prior knowledge about the causal links among CHD risk factors, potential risk factors, confounders, intermediate variables, and the outcome, CHD.

Given that multiple direct and indirect risk factors are part of the causal web of CHD, such an evaluation not only must consider the direct effect of the risk factors under intervention but should also include their effects mediated through other risk factors. As various risk factors simultaneously act as confounders and as intermediate steps, traditional regression analysis is not an appropriate method to control for confounding, and a causal method must be used.

Methods

The analysis was based on the Framingham Offspring Study longitudinal data ($n = 5{,}124$) with a 20-year follow-up. The parametric g-formula was used to adjust for time-dependent confounding and to estimate the counterfactual CHD risk under each intervention. Pooled logistic regression models were used to predict risk factors and CHD distributions conditional on given risk factor history. The Monte Carlo technique and the bootstrap method were used to estimate relative CHD risks with 95% confidence intervals. Evaluated strategies included interventions on smoking, alcohol consumption, body mass index (BMI), and low-density lipoprotein (LDL), and a combined strategy.

Results

The simulated 12-year risk of CHD under no intervention was about 8% for males and 3% for females. Smoking cessation at baseline in all smokers had a statistically significant relative risk of .8 in males and females ($p < .05$). The relative risk after shifting the LDL distribution to the distribution of the Chinese population was .7 for men and .5 for women (both $p < .05$). Shifting alcohol consumption to moderate alcohol intake or constantly lowering BMI to 22 kg/m^2 did not change CHD risk significantly. The combined intervention on smoking cessation, BMI, and LDL reduced the CHD risk by more than 50% in men and women ($p < .05$).

Conclusions

The parametric g-formula could be applied in a multiple risk factor analysis with time-dependent confounding, where traditional regression analysis fails. It showed that combined interventions have a joint potential of reducing CHD risk by more than 50%.

Causal Analysis of Treatment: The Adherence-Adjusted Effect of Hormone Therapy on Coronary Heart Disease

Background

The Women's Health Initiative (WHI) randomized trial found greater CHD risk in women assigned to estrogen/progestin therapy than in those assigned to a placebo. Observational studies had previously suggested reduced CHD risk in hormone users.

Methods

Miguel A. Hernán and colleagues used the data from the observational Nurses' Health Study. They emulated the design and intention-to-treat (ITT) analysis of the WHI randomized trial. Because the ITT approach causes severe treatment misclassification, the authors also controlled for time-dependent confounding and estimated adherence-adjusted effects by inverse probability weighting. Hazard ratios of CHD were calculated comparing initiators versus noninitiators of estrogen/progestin treatment.

Results

The results showed ITT hazard ratios of CHD similar to those from the WHI. The results from inverse-probability of treatment weighting analysis suggest that continuous hormone therapy causes a net reduction in CHD among women starting therapy within 10 years of menopause, and a net increase among those starting later. However, the authors mentioned that it cannot be excluded that either of these effects could be due to sampling variability.

Conclusions

These findings suggest that the discrepancies between the WHI and Nurses' Health Study ITT estimates could be largely explained by differences in the distribution of time since menopause and length of follow-up. The probability of treatment analysis allowed adjustment for adherence and determination of the CHD risks of hormone therapy versus no hormone therapy under full adherence.

Uwe Siebert

See also Applied Decision Analysis; Bias in Scientific Studies; Causal Inference and Diagrams; Confounding and Effect Modulation

Further Readings

Cole, S. R., & Hernán, M. A. (2002). Fallibility in estimating direct effects. *International Journal of Epidemiology, 31*(1), 163–165.

Cole, S. R., Hernán, M. A., Robins, J. M., Anastos, K., Chmiel, J., Detels, R., et al. (2003). Effect of highly active antiretroviral therapy on time to acquired immunodeficiency syndrome or death using marginal structural models. *American Journal of Epidemiology, 158*(7), 687–694.

Greenland, S., & Brumback, B. (2002). An overview of relations among causal modelling methods. *International Journal of Epidemiology, 31*, 1030–1037.

Greenland, S., Pearl, J., & Robins, J. M. (1999). Causal diagrams for epidemiologic research. *Epidemiology, 10*(1), 37–48.

Hernán, M. A., Hernandez-Diaz, S., Werler, M. M., & Mitchell, A. A. (2002). Causal knowledge as a prerequisite for confounding evaluation: An application to birth defects epidemiology. *American Journal of Epidemiology, 155*(2), 176–184.

Pearl, J. (2000). *Causality*. Cambridge, UK: Cambridge University Press.

Robins, J. M. (1998). Marginal structural models. In *1997 proceedings of the American Statistical Association*. Section on Bayesian Statistical Science (pp. 1–10). Washington, DC: American Statistical Association.

Robins, J. M., Hernán, M. A., & Siebert, U. (2004). Estimations of the effects of multiple interventions. In M. Ezzati, A. D. Lopes, A. Rodgers, & C. J. L. Murray (Eds.), *Comparative quantification of health risks: Global and regional burden of disease attributable to selected major risk factors* (pp. 2191–2230). Geneva: World Health Organization.

Tilling, K., Sterne, J. A. C., & Moyses, S. (2002). Estimating the effect of cardiovascular risk factors on all-cause mortality and incidence of coronary heart disease using g-estimation. *American Journal of Epidemiology, 155*(8), 710–718.

Certainty Effect

Within decision making, the certainty effect is used to describe the impact of certainty on the decision maker. People are drawn to certainty, giving higher preference to options that have high levels

of certainty. An option with high certainty (close to 0% or 100%) is more appealing to people than a complex or ambiguous probability. This causes many decision makers to choose options that go against the expected utility of the problem. A reduction in probability has a greater impact on the decision maker if the initial outcome is certain. For example, a reduction in survivability from 100% to 90% would have a greater impact than a reduction in survivability from 70% to 60%.

The underlying reason for the certainty effect falls on a person's preference for certain or absolute values. People will bear psychological effects from feelings both of certainty and of uncertainty. They prefer certainty, rather than complexity and ambiguity. Most decision makers cannot clearly define the difference between two probabilities, especially if they are ambiguous. Rather than consider exact probabilities, people often lump outcomes into categories such as "likely" and "unlikely." This makes comparison between two "likely" probabilities difficult. For example, if a healthcare provider explains two courses of treatment to a patient, he or she may present some probability of full recovery. If both options presented a midrange probability, it would be difficult for the patient to decipher the true difference between them. Consider the case where the first course of treatment presents a 70% chance of full recovery, whereas the second presents a 60% chance of full recovery. Most people would be unable to differentiate between these two probabilities but would rather refer to them as "good chances," "likely," or "better than average." If one course of treatment had extreme certainty (close to 100% in this example), the decision maker would put a higher weight on the certain treatment. This is due to the fact that decision makers tend to eliminate uncertainty altogether by overweighting the certain outcomes.

Consider the following case, originally presented by Amos Tversky and Daniel Kahneman. Treatment A leads to a 20% chance of imminent death and an 80% chance of normal life, with a longevity of 30 years. Treatment B leads to a 100% chance of normal life with a longevity of 18 years. According to expected utility theory, rational decision makers would choose Treatment A as it provides a higher utility in terms of lifespan (24 years compared with 18 years). However, the majority of decision makers choose Treatment B. This is a prime example of the certainty effect in practice. Decision makers, be they physicians or patients, have a high preference for certain outcomes, regardless of the comparative utilities associated with them.

Decision makers are confident when handling extreme probabilities (near 0 or 1.0). When the probabilities are not as certain, however, the weighting of alternatives becomes disproportionate. Decreasing a risk from 5% to 0% should have the same utility as decreasing that risk from 20% to 15%. However, decision makers greatly prefer the first.

An experiment introduced by Richard Zeckhauser illustrates the certainty effect phenomenon. Respondents in the experiment were asked to imagine that they were compelled to play Russian roulette. They were given the opportunity to purchase the removal of one bullet from the loaded gun by choosing one option. Option 1 allowed them to reduce the number of bullets from four to three. Option 2 allowed them to reduce the number of bullets from one to zero. The respondents were asked to how much they would be willing to pay for each option. The result was that a majority of respondents would pay much more for the second option. This is the option that reduced their chances of being shot to 0.

On examination of both options, it is clear that the utility of each option is equal. Option 1 has a probability of being shot of 67%, which is reduced to 50% on removal of one bullet. Option 2 has a probability of being shot of 17%, which is reduced to 0% on removal of one bullet. Both options experienced a reduction of probability (or risk) of 17%. From the perspective of utility, both options are the same. However, people strongly preferred the option that led to certainty: Option 2.

The certainty effect is noticeable in situations that have positive prospects as well as those with negative prospects. In the positive domain, decision makers address scenarios in which there is a probability of a gain. Examples could include winning money in a lottery or increasing life expectancy. The key component of the certainty effect, one's overweighting of certainty, favors risk aversion in the positive domain. The decision maker would prefer a sure gain over a larger gain that is merely probable. In the negative domain, decision makers consider effects when presented with a loss scenario. This could include loss of life, increased illness, or side effects. The overweighting of certainty favors risk seeking in the domain of losses.

In the negative domain, the same effect leads to a risk-seeking preference for a loss that is merely probable over a smaller loss that is certain.

The certainty effect is a demonstration of how humans do not make rational decisions. This is not to say that they make incorrect decisions but rather that they have a stated preference toward things that are absolute. The certainty effect should be considered when evaluating how people make decisions.

Lesley Strawderman and Han Zhang

See also Allais Paradox; Certainty Equivalent; Expected Utility Theory; Prospect Theory

Further Readings

Cohen, M., & Jaffray, J.-Y. (1988). Certainty effect versus probability distortion: An experimental analysis of decision making under risk. *Journal of Experimental Psychology: Human Perception and Performance, 14,* 554–560.

Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica, 47,* 263–292.

Tversky, A., & Kahneman, D. (1981). The framing of decisions and the psychology of choice. *Science, 211,* 453–458.

Tversky, A., & Kahneman, D. (1986). Rational choice and the framing of decisions. *Journal of Business, 59,* 251–278.

Certainty Equivalent

When examining the potential outcomes a given event may hold, a person is likely to approximate the probability of each possible result. Taken into consideration by the individual is the degree to which any of these outcomes may be certain. That is, although the benefit derived from an event with a lesser or unknown likelihood of occurring may be much greater, people often tend to opt for the less advantageous, although more certain, outcome. An influential variable, however, is to what degree the individual finds the certain outcome to be of value. Thus, certainty equivalents are the amount of utility, or usefulness, that a person will consider to forgo an offered gamble. When the person becomes indifferent between the choice of a certain event and a probabilistic one, the value of the certain event is called the *certainty equivalent*.

Certainty equivalents are used most frequently in an outward sense within the realm of economic ventures, though individuals may subconsciously use the framework for any scenario in which a gamble presents itself. The utility terms will therefore vary with the application as what is considered beneficial is highly circumstantial. Within medical decision making, however, certainty equivalents could include financial aspects relating to choices in care, various measures of quality of life for the self and for others, or potential recovery periods. The difference between the expected value of indefinite outcomes and the certainty equivalent is referred to as the risk premium.

Finding Certainty Equivalents

A number of mathematical methods exist for finding the certainty equivalent based on the utility function being presented. However, in practice, a person's certainty equivalent can be found more pragmatically by asking a series of questions. Each question should ask the person to choose one of two options. The first option presents a gamble, whereas the second option presents a certain outcome. If the person chooses the gamble, a second question is posed. This time, the first option remains the same, but the conditions of the gamble are altered. The question is presented to the individual so that the perceived benefit or the probability of such has been increased. This line of questioning continues until the person either chooses the second option (a given payout) or says he or she cannot decide. At this point, the value of the payout becomes the certainty equivalent.

Responses to Risk

Enhancements to actual gain or likelihood ratios will produce varying responses, many of which are dependent on the individual's bearing. While one person may be inclined to accept a gamble for a larger disbursement based on a lower probability, another may require a large probability for even the lowest of disbursements. While these variances are indeed environmentally produced, it has been suggested that inherent personality differences affect an individual's willingness to entertain the idea of a gamble over a more certain outcome.

Most people are considered risk-averse. That is, they avoid risk whenever necessary, rather opting for a certain outcome. For a risk-averse person, the

certainty equivalent is generally lower than the expected value of the gamble. This condition is present in individuals who desire to behave in a way so as to reduce potential uncertainties. That is, people of this nature are likely to respond to expected outcomes, however smaller the advantage of a less assured event may be. The risk premium for a risk-averse person would be positive. He or she would require an extra incentive to take the gamble, usually found in the form of an increase in the probability of a given event occurring rather than an augmentation of the prospective payout.

A risk-seeking person, however, would have a certainty equivalent that is higher than the expected value of the gamble, indicating his or her preference for scenarios in which the outcome is less certain. The risk premium for a risk-seeking person would be negative. He or she would require an extra incentive to not take the gamble. It should be noted, however, that people who possess this characteristic are more likely to respond to rewards than to punishments. For example, given the same outcome for two given events, one certain and one probabilistic, an individual who is risk-seeking is unlikely to be deterred by the potential loss or detriment caused by an unsuccessful gamble. That is, the payout is of more importance than the probability; decreasing the likelihood of success in the gambling scenario is unlikely to dissuade that choice. Rather, a more efficient discouragement would come in the form of increasing the payout of the certain event or decreasing the payout of the gamble.

A risk-neutral person would have a certainty equivalent equal to the expected value of the gamble, leading to a risk premium of zero. Whereas people averse to risk require alterations in probabilities, and people attracted to risk require alterations in perceived benefits, those who are risk-neutral may respond to either of these variants. Individuals exhibiting indifference toward two more or less certain outcomes will be equally influenced by alterations to probability as well as to benefit. As opposed to other scenarios, within medical decision making it should also be considered that individuals are apt to take into account not only the expected benefit but the prospective harm that may result as well.

Application to Healthcare

As with other concepts within expected utility theory, frameworks of certainty equivalents can be applied to healthcare. Particularly when a patient is capable of receiving treatment through multiple options, the treating physician as well as the individual and his or her family are likely to consider certainty equivalents. The utility at hand becomes the treatment outcomes (recovery period, additional life expectancy), rather than the traditional financial outcomes. Therefore, within healthcare, although financial considerations are certainly taken into account when determining courses of treatment, the degrees of likelihood and their associated risks and benefits are more influential. Within more conventional gambling scenarios, the perceived benefit is often what drives the decision to participate or forgo the opportunity. When considering medical decision making, however, patients are likely to weigh equally the potential risks, such as the amount of pain expected or the risks associated with a given course of treatment.

For example, consider a patient who is presented with two treatment options. Treatment A is a lottery. Treatment A gives the patient a 50% chance of living an additional 10 years and a 50% chance of living an additional 5 years. Treatment B, however, is certain, giving the patient a 100% chance of living X years. According to expected utility theory, these two treatment options would have equal utility when $X = 7.5$ years. If the patient is risk-averse, their certainty equivalent would be lower, possibly $X = 6$ years. This means that he or she would choose Treatment B only if $X \geq 6$. A risk-seeking patient, however, would have a higher certainty equivalent, possibly $X = 8$ years. In this case, the patient would choose Treatment B only if $X \geq 8$ years. Otherwise, he or she would opt for Treatment A, preferring to take a gamble. The above example assumes, however, that the most important consideration when determining courses of medical treatment is the additional life expectancy gained. A more holistic approach realizes that multiple considerations are often influential within medical decision making, such as the expected quality of life associated with varying treatments.

Lesley Strawderman and Lacey Schaefer

See also Certainty Effect; Expected Utility Theory; Risk Aversion

Further Readings

Benferhat, S., & Smaoui, S. (2007). Hybrid possibilistic networks. *International Journal of Approximate Reasoning, 44*(3), 224–243.

Feeny, D. H., & Torrance, G. W. (1989). Incorporating utility-based quality-of-life assessment measures in clinical trials. *Medical Care, 27*(3), 190–204.

Hennessy, D. A., & Lapan, H. E. (2006). On the nature of certainty equivalent functionals. *Journal of Mathematical Economics, 43*(1), 1–10.

Hsee, C. K., & Weber, E. U. (1997). A fundamental prediction error: Self-others discrepancies in risk preference. *Journal of Experimental Psychology, 126*(1), 45–53.

Quiggin, J., & Chambers, R. G. (2006). Supermodularity and risk aversion. *Mathematical Social Sciences, 52*(1), 1–14.

Chained Gamble

Chaining (also called indirect linking) as used in the expression *chained gamble* or *chained lottery* is best conceived of as a strategy of adjusting preference measurement used within preference elicitation techniques such as the standard gamble and time tradeoff methodologies. The approach of chaining gambles (or chaining lotteries) has been offered as a solution to the problem of within-technique inconsistency found with real-world use and testing of the standard gamble as a preference elicitation methodology in economic and medical decision making. The goal here is to understand why such chaining is proposed as an attempt to solve problems of lack of internal consistency (presence of internal inconsistency) in preference elicitation methodologies. This entry illustrates chained gambles for the standard gamble.

Detecting a Lack of Internal Consistency

Internal consistency can be examined in the following way. One technique is based on a more direct value preference elicitation. This more basic technique is used to generate values based on a "basic reference exercise," which is in fact the basic technique used to generate (elicit from patients) values across a set of outcomes. Such a basic reference exercise may invoke a simple elicitation exercise where the individual is asked to rank order outcomes on a scale from most desirable to least desirable. The second strategy, indirect value elicitation through a chained exercise, generates values based on the use of the standard gamble technique. Adam Oliver, who has examined the internal consistency of a variety of techniques, including standard gambles, argues that if the first and the second strategies yield results that do not significantly or systematically differ from one another, then one might be able to say that the strategies are each internally consistent in that they yield the same ordering of preferences. If individuals distinguish between and among states on the basis of the basic simple direct rank ordering of preferences strategy but are unable to distinguish between their preferences for outcomes in the preference elicitation procedure, then there are potential problems.

Once an inconsistency is found in the use of a preference elicitation methodology, the search is on for what is causing this inconsistency. Initial considerations may fall on issues related to the patients as respondents whose preferences are being elicited and who may lack experience with the use of the technique. To eliminate the inexperienced respondent as the potential source of the problem, an attempt is then made to seek out and to study more experienced respondents, for example, more experienced professionals more familiar with the use of such techniques, to see if the experienced professionals also have problems with internal consistency of results using the techniques. If the same problems (or roughly the same problems) are found with both groups, then the next question that comes up is whether the problem is with the technique being used in preference elicitation itself.

If the same type of inconsistency is found in the elicitation of preferences from both inexperienced respondents and experienced professionals, then the technique itself may be causing the inconsistency. Chaining is a technique aimed at amending the inconsistencies found in the use of the standard gamble as a preference elicitation technique.

It should be noted, though, that any time one has to introduce a technique (a methodology or a procedure) into an arena of decision making of any sort, one needs to recognize that the arena being studied is of such a level of complexity that simply straightforward asking and answering of questions cannot always be employed to achieve the desired result, that is, the understanding of what the individual's (the patient's) preferences are across a set of outcomes.

Cross-Technique Inconsistencies

Inconsistencies between techniques and among techniques may exist, but the question that needs addressing is inconsistency within techniques. If an inconsistency between or among techniques is found, one still needs a counting procedure for determining answers to the following three questions: (1) What is to count as an inconsistency? (2) When does an inconsistency exist within a technique? (3) What is to be used as the gold standard for a consistent technique? Here, there must be an agreement on some basic preference ordering tool (instrument) that then serves as the basis for deciding which technique is "better" and on what grounds. In the example of chaining, *better* is defined in terms of more degrees of agreement with the results of the basic ordering tool.

Within-Technique Inconsistency and Standard Gamble

Using a basic ordering tool, within-technique inconsistency has been found with the standard gamble technique. Before elucidating the type of within-technique inconsistency that has been demonstrated with the standard gamble, it is important to understand the following definitions and concepts that exist within contemporary standard gamble discussions.

To understand chained gambles, one needs to understand the following basic assumptions about the division of healthcare states: extreme states, moderate states, and minor states. Extreme, moderate, and minor health states may be described in terms of severity of state, in terms of permanence (degree of irreversibility) of state, or in terms of severity and permanence of state.

Extreme States

Typically, in the decision sciences, there are two extremes of health states considered. At the negative extreme, there is death (considered in many frameworks as "immediate death"); at the other, positive extreme, there is "full or perfect health." Yet there are questions whether the state "immediate death" is itself the extreme end of the negative range of ill health states. More extreme states than death as considered by reasonable patients may include (a) end-staged neurodegenerative disease processes or severe cerebrovascular accidents (strokes) causing loss of memory, loss of thinking capacity, and progressive motor loss, and (b) end-stage cardiopulmonary disease (heart and lung failure) where there is a tremendous work of breathing and inability to carry out any exertion in one's daily life.

Moderate States

In neurology, any "less severe" state of loss of memory, loss of thinking capacity, loss of motor abilities, or less severe sensory loss may be considered a more moderate state of impaired health. In cardiology, states of increasingly severe chest discomfort or increasing limitations on one's abilities to exert oneself in walking can be described by some individuals as moderate states of ill health.

Minor States

Minor states are impaired states of a much lower intensity or severity or shortened temporal course than moderate states. In neurology, a minor degree loss of motor strength, 4.9 on a scale of 5.0, or in cardiovascular disease, 5 minutes of mild chest discomfort per week, may be considered as minor health states. However, as one walks patients down from extreme to moderate to minor heath states and then on down to full or perfect health, some patients may find it hard to distinguish between minor states of impaired health and states of full or perfect health.

Interestingly, within-technique inconsistency has been found as a problem within standard gambles when minor states are being considered by the patient whose preferences are being elicited.

When individuals are asked their willingness to trade chance of survival for improvements in health status in a standard gamble, oftentimes they are unwilling to trade *chances of survival* for *improvements in health status*. In a basic simple direct rating exercise given prior to a preference elicitation of an individual, the individual reports that he or she is able to distinguish between states, but then when approached with a standard gamble, he or she reports that he or she is unwilling to trade. For example, an individual who is able to order full or perfect health as "better than" (more desirable than) a state of minor adverse or poor health in the basic reference exercise above is unwilling to trade chances of survival for improvements in health status in a standard gamble.

In the case of medical decision making, the approach of standard gambles has been found to be inconsistent in the following way: When minor or temporary states of health, most notably negative (adverse) states of ill health (poor health), are being evaluated, it is difficult for patients to evaluate the preferences for minor states of poor health as different from states of full or perfect health. Thus, there is an inability to truly assess minor states of poor health through the standard gamble elicitation methodology.

Chained Gambles

One way that has been proposed to improve on the standard gamble as a preference elicitation technique is to use chained gambles. Chaining links minor or temporary health states to death through intermediate states that then all become the links of a chain. Here, instead of valuing a minor or temporary health state against immediate death, one values the minor state with a moderate (intermediate) state and then the moderate (intermediate) state with immediate death.

For example, in neurology, if a slight hand tremor is the adverse outcome being valued, the treatment failure outcome in the chained comparison could be hand paralysis. The hand paralysis could then be "chained" in to form a further gamble where the paralysis of a hand is valued against a treatment that offers a chance of full or perfect health or immediate death. Another example can be found in the area of vascular surgery and peripheral vascular disease. If an individual is considering a state of intermittent claudication (cramplike discomfort felt in the lower legs and thighs often due to blockages in the supply of blood to the lower legs), intermittent claudication could be valued as the intermediate state in the chain against the loss of the ability to walk.

Here, minor and temporary adverse health states are valued relative to moderate and severe health states that are then valued against full or perfect health and immediate death. This use of chaining then assumes that through the use of such intermediate states, preferences will be preserved (more matching of preferences with the standard gamble when compared with preferences elicited by the basic reference exercise) and will be detected and picked up in a standard gamble that had previously been considered as "sufficiently insensitive" to pick up key nuances in patient preferences. Oliver phrases the goal of chaining as the achievement of a consistent methodology where "direct value preference elicitation through a basic reference exercise" and "indirect value elicitation through a chained exercise" generate values that do not significantly or systematically differ from one another.

Future Use and Research

There is much need for preference elicitation strategies beyond disease states requiring surgery of terminal medical conditions. Possible uses of chaining are with patients with chronic illnesses, such as rheumatoid arthritis. Here, many patients are in chronic states of compromised health and need help with comparing management or treatment strategy states of consideration of "degree of increase in bodily versus mental functioning" on one therapy versus another therapy, without consideration of immediate death or full (or perfect) health because immediate death or full perfect health are not reasonable short- or medium-term outcomes in these patients with chronic diseases.

More research needs to be done in states of function through all their degrees of severity and states of permanence (degrees of irreversibility) as chronic diseases will continue to progress in these individuals over time throughout their lives. Challenges also exist in the area of developing new techniques for preference elicitation that aim to further reduce or eliminate internal inconsistency as a problem within preference elicitation methodologies.

Dennis J. Mazur

See also Expected Utility Theory; Utility Assessment Techniques

Further Readings

Baker, R., & Robinson, A. (2004). Responses to standard gambles: Are preferences "well constructed"? *Health Economics, 13,* 37–48.

Jones-Lee, M. W., Loomes, G., & Philips, P. R. (1995). Valuing the prevention of non-fatal road injuries: Contingent valuation vs. standard gambles. *Oxford Economic Papers, 47,* 676–695.

Llewellyn-Thomas, H., Sutherland, H. J., Tibshirani, R., Ciampi, A., Till, J. E., & Boyd, N. F. (1982). The measurement of patients' values in medicine. *Medical Decision Making, 2,* 449–462.

McNamee, P., Glendinning, S., Shenfine, J., Steen, N., Griffin, S. M., & Bond, J. (2004). Chained time

trade-off and standard gamble methods: Applications in oesophageal cancer. *European Journal of Health Economics, 5,* 81–86.

Oliver, A. (2003). The internal consistency of the standard gamble: Tests after adjusting for prospect theory. *Journal of Health Economics, 22,* 659–674.

Oliver, A. (2004). Testing the internal consistency of the standard gamble in "success" and "failure" frames. *Social Science & Medicine*, *58,* 2219–2229.

Oliver, A. (2005). Testing the internal consistency of the lottery equivalents method using health outcomes. *Health Economics, 14,* 149–159.

Rutten-van Mölken, M. P., Bakker, C. H., van Doorslaer, E. K., & van der Linden, S. (1995). Methodological issues of patient utility measurement: Experience from two clinical trials. *Medical Care, 33,* 922–937.

Torrance, G. W. (2006). Utility measurement in healthcare: The things I never got to. *Pharmaco-Economics, 24,* 1069–1078.

Witney, A. G., Treharne, G. J., Tavakoli, M., Lyons, A. C., Vincent, K., Scott, D. L., et al. (2006). The relationship of medical, demographic and psychosocial factors to direct and indirect health utility instruments in rheumatoid arthritis. *Rheumatology (Oxford), 45,* 975–981.

Chaos Theory

A major breakthrough of the 20th century, which has been facilitated by computer science, has been the recognition that simple rules do not always lead to stable order but in many circumstances instead lead to an apparent disorder characterized by marked instability and unpredictable variation for reasons intrinsic to the rules themselves. The phenomenon of rules causing emerging disorder, counterintuitive to many people, is the environment currently being explored as *self-organization, fractals* (a fragmented geometric shape that can be split into parts, each of which is a reduced-size copy of the whole, a property called self-similarity), *nonlinear dynamical systems,* and *chaos.*

Chaos theory, also called nonlinear systems theory, provides new insights into processes previously thought to be unpredictable and random. It also provides a new set of tools that can be used to analyze physiological and clinical data such as the electric signals coming from the heart or from the brain.

Chaos theory was born originally as a branch of mathematical physics in the 20th century thanks to the work of Edward Lorenz in meteorology. Chaos theory is concerned with finding rational explanations for such phenomena as unexpected changes in weather and deals with events and processes that cannot be modeled or predicted using conventional mathematical laws and theorems, such as those of probability theory. The theory basically assumes that small, localized perturbations in one part of a complex system can have profound consequences throughout the system. Thus, for nonlinear systems, proportionality simply does not hold. Small changes can have dramatic and unanticipated consequences. The fascinating example often used to describe this concept, which is known as the butterfly effect, is that the beating of a butterfly's wings in China can lead to a hurricane in Brazil, given a critical combination of air pressure changes.

The key word is *critical,* and many of the efforts of scientists working on chaos theory are concerned with attempts to model circumstances based on specific conditional conjunction. Unpredictable events in medicine, such as ventricular arrhythmias and sudden cardiac death in athletes, the course of certain cancers, and the fluctuations in frequency of some diseases, may be attributable to chaos theory.

Nonlinear Dynamics in Human Physiology

Chaos theory can be considered a paradigm of the so-called nonlinear dynamics. The issue of nonlinearity of medical data has very rarely been raised in the literature. Clearly, epidemiologists and statisticians devoted to the medical field are quite happy with linear techniques since they have been trained from the beginning with them; physicians and other health professionals, due to their proverbial poor mathematical competence, are also happy, provided that statisticians and regulatory agencies do not think differently.

What does a linear function signify? If one considers a Cartesian chart in which axis x represents the money a person gets and axis y measures the degree of happiness that person obtains as a result, then the more money a person has, the happier he or she is. In this scenario, one can easily predict the value of one variable by the value of the other, with a simple (linear) equation. However, this scenario, as with many others in real life, is actually more an exception than a rule. In real life, the relations are generally more complex. In fact, as many people can witness, an increase in earning can sometimes

produce fears of losing money or uncertainties on how to invest this money, and this can reduce the feeling of happiness. This complex (nonlinear) relation does not permit one to understand, at first glance, from data gathered experimentally, the relationship between money and happiness.

Therefore, persisting in the linear approach is not without danger: If, for instance, for two given variables a correlation coefficient of .018 is calculated under the linear hypothesis and a p value of .80 is added, a relationship between the two is ruled out. Revisiting the relationship between these two variables through the nonlinear approach could change the situation dramatically since fuzzy and smooth interactions may determine significant effects through a complex multifactorial interplay.

Mathematical analyses of physiological rhythms, such as those of Jerry Gollub, show that nonlinear equations are necessary to describe physiological systems. The physiological variation of blood glucose, for example, has traditionally been considered to be linear. Recently, a chaotic component has been described both in diabetic patients and in normal subjects. This chaotic dynamic has been found to be common in other physiologic systems. Table 1 summarizes some of the best examples of nonlinear dynamics in human physiology. It has, for instance, been shown that the interbeat interval of the human heart is chaotic and that a regular heart beat is a sign of disease and a strong predictor of imminent cardiac arrest.

The work of Ary L. Goldberger has pointed out how traditional statistics can be misleading in evaluating heart time series in health and disease. In fact, there are circumstances in which two data sets belonging to two subjects can have nearly identical mean values and variances and, therefore, escape statistical distinction based on conventional comparisons. However, the raw time series can reveal dramatic differences in the temporal structure of the original data, wherein one time series is from a healthy individual and the other from a patient during episodes of severe obstructive sleep apnea. The time series from the healthy subject reveals a complex pattern of nonstationary fluctuations. In contrast, the heart rate data set from the subjects with sleep apnea shows a much more predictable pattern with a characteristic timescale defined by prominent, low-frequency oscillations at about .03 Hz. Both the complex behavior in the healthy case and the sustained oscillations in the pathologic one suggest the presence of nonlinear mechanisms.

Table 1 Examples of nonlinear dynamics in human physiology

Processes With Chaotic Behavior	*Processes With Complex Fractal Fluctuations*
Shape of EEG waves	Heart frequency
Insulin blood levels	Respiration
Cellular cycles	Systemic arterial pressure
Muscle action potential	Gait control
Esophagus motility	White blood cells number
Bowel motility	Liver regeneration patterns
	Uterine pressure

Source: Glass, L., & Mackey, M. C. (1988). *From clocks to chaos: The rhythms of life*. Princeton, NJ: Princeton University Press.

Other researchers such as Bruce McEwen and John Wingfield have introduced the concept of allostasis—maintaining stability through change—as a fundamental process through which organisms actively adjust to both predictable and unpredictable events. *Allostatic load* refers to the cumulative cost to the body of allostasis, with *allostatic overload* being a state in which serious pathophysiology can occur. In this regard, chaos theory seems to fit quite well with biological adaptation mechanisms.

The importance of chaotic dynamics and related nonlinear phenomena in medical sciences has been only recently appreciated. It is now quite clear, as noted by David Ruelle, that chaos is not mindless disorder—it is a subtle form of order—and that approximate results of treatment can be predicted.

Chaotic dynamics are characterized most of the time by what is called a strange attractor. This roughly means that during the chaotic evolution, the variables characterizing the state of the system remain in a restricted range of values. This leads to the possibility of characterizing the system evolution in terms of probabilities.

Applications to Medical Settings

One promising application of dynamic analysis involves strategies to restore complex biological variability, including fractal fluctuations (i.e., harmonic changes to self-similar heart rhythms), to cardiopulmonary systems. Initial results using artificial ventilation in experimental animals and

clinical settings suggest the possibility of improving physiologic function with "noisy" versus "metronomic" parameter settings. The use of dynamic assays to uncover basic and clinical information encoded in time series also promises to provide new, readily implemented diagnostic tests for prevalent conditions such as sleep-disordered breathing. The extent to which dynamic measures and complexity-informed models and interventions will enhance diagnostic capabilities and therapeutic options in chronic obstructive lung disease is an intriguing area for future study.

Another paradigmatic area of interest and application is represented by electroencephalography (EEG). The 19 channels in the EEG represent a dynamic system characterized by typical asynchronous parallelism. The nonlinear implicit function that defines the ensemble of electric signals series as a whole represents a meta-pattern that translates into space (hypersurface) what the interactions among all the channels create in time.

The behavior of every channel can be considered as the synthesis of the influence of the other channels at previous but not identical times and in different quantities, and of its own activity at that moment. At the same time, the activity of every channel at a certain moment in time is going to influence the behavior of the others at different times and to different extents. Therefore, every multivariate sequence of signals coming from the same natural source is a complex asynchronous dynamic system, highly nonlinear, in which each channel's behavior is understandable only in relation to all the others.

The neurophysiologic community has had the perception that in the EEG signals there is embedded much more information on brain function than is currently extracted in a routine clinical context, moving from the obvious consideration that the sources of EEG signals (cortical postsynaptic currents at dendritic tree level) are the same ones attacked by the factors producing symptoms of chronic degenerative diseases such as dementia. The main problem, then, is the signal (relevant information)-to-noise (nonrelevant information) ratio, in which at the present moment the latter is largely overwhelming the former. As an example, when considering the EEG fluctuations at the 19 recording electrodes, it is like the fluctuation of 19 stock exchange securities in time (minutes, hours, days, etc.) due to the purchases/sales ratios as carried out by millions of invisible investors, following a logic that is unknown to the analyzer but that is based on the intrinsic mechanism regulating the market. In this context, the "analyzer" ignores all the following variables:

1. why at each time the value of a given security (EEG signal) is going up or down;
2. how many investors (neurons, synapses, synchronous firing) are active on that security at a given time; and
3. when new investors, eventually organized, suddenly enter the market that is regulating that security and significantly alter the trend of the previous fluctuations (i.e., the subject's condition is altered because of an external or internal event).

The only two variables that the analyzers know for sure are the following:

1. The chaotic stock market entirely depends on the interplay of a large number of investors (brain, neurons, synapses).
2. Within the dynamics (variability) of the stock securities are embedded the investors' styles and abilities.

A 2007 article by Massimo Buscema and colleagues presents the results obtained with the innovative use of special types of artificial neural networks (ANNs) assembled in a novel methodology named IFAST (Implicit Function as Squashing Time) capable of compressing the temporal sequence of EEG data into spatial invariants (patterns of structures that remain stable across time). The principal aim of the study was testing the hypothesis that automatic classification of mild cognitive impairment (MCI) and Alzheimer's disease (AD) subjects can be reasonably corrected when the spatial content (the inherent structure) of the EEG voltage is properly extracted by ANNs.

Resting eyes-closed EEG data were recorded in 180 AD patients and in 115 MCI subjects. The spatial content of the EEG voltage was extracted by the IFAST stepwise procedure using ANNs. The data input for the classification operated by ANNs were not the EEG data but the connections weights of a nonlinear auto-associative ANN trained to reproduce the recorded EEG tracks. These weights represented a good model of the peculiar spatial features of the EEG patterns at the scalp surface. The classification based on these parameters was binary

(MCI vs. AD) and was performed by a supervised ANN. Half of the EEG database was used for the ANN training, and the remaining half was used for the automatic classification phase (testing).

The results confirmed the working hypothesis that a correct automatic classification of MCI and AD subjects can be obtained by extracting the spatial information content of the resting EEG voltage by ANNs and represents the basis for research aimed at integrating the spatial and temporal information content of the EEG. The best results in distinguishing between AD and MCI reached up to 92.33%. The comparative result obtained with the best method so far described in the literature, based on blind source separation and Wavelet preprocessing, was 80.43% ($p < .001$).

Future Outlook

The advancement of knowledge and progress in understanding the nature of bodily rhythms and processes have shown that complexity and nonlinearity are ubiquitous in living organisms. These rhythms arise from stochastic (involving or containing a random variable or variables), nonlinear biological mechanisms interacting with fluctuating environments.

There are many unanswered questions about the dynamics of these rhythmic processes: For example, how do the rhythms interact with each other and the external environment? Can researchers decode the fluctuations in physiological rhythms to better diagnose human disease? Mathematical and physical techniques combined with physiological and medical studies are addressing these questions and are transforming our understanding of the rhythms of life.

Enzo Grossi

See also Complexity

Further Readings

Buscema, M., Rossini, P., Babiloni, C., & Grossi, E. (2007). The IFAST model, a novel parallel nonlinear EEG analysis technique, distinguishes mild cognitive impairment and Alzheimer's disease patients with high degree of accuracy. *Artificial Intelligence in Medicine, 40*(2), 127–141.

Firth, W. J. (1991). Chaos, predicting the unpredictable. *British Medical Journal, 303,* 1565–1568.

Glass, L., & Mackey, M. C. (1988). *From clocks to chaos: The rhythms of life.* Princeton, NJ: Princeton University Press.

Goldberger, A. L., Amaral, L. A. N., Hausdorff, J. M., Ivanov, P. C., Peng, C. K., & Stanley, H. E. (2002). Fractal dynamics in physiology: Alterations with disease and aging. *Proceedings of the National Academy of Sciences of the United States of America, 99,* 2466–2472.

Goldberger, A. L., & Giles, F. (2006). Filley lecture: Complex systems. *Proceedings of the American Thoracic Society, 3*(6), 467–471.

Gollub, J. P., & Cross, M. C. (2000). Nonlinear dynamics: Chaos in space and time. *Nature, 404,* 710–711.

Kroll, M. H. (1999). Biological variation of glucose and insulin includes a deterministic chaotic component. *Biosystems, 50,* 189–201.

Lorenz, E. N. (1963). Deterministic non periodic flow. *Journal of the Atmospheric Sciences, 20,* 130–141.

McEwen, B. S., & Wingfield, J. C. (2003). The concept of allostasis in biology and biomedicine. *Hormonal Behavior, 43*(1), 2–15.

Ruelle, D. (1994). Where can one hope to profitably apply the ideas of chaos? *Physics Today, 47,* 24–30.

Singer, D. H., Martin, G. J., Magid, N., Weiss, J. S., Schaad, J. W., Kehoe, R., et al. (1988). Low heart rate variability and sudden cardiac death. *Journal of Electrocardiology, 21,* S46–S55.

Choice Theories

Choice theories can be classified in a number of ways. *Normative* theories seek to clarify how decisions should be made; *descriptive* theories try to understand how they are made in the real world. Theories may also concentrate on decisions made by individuals, groups, or societies. Normative theories tend to emphasize rational decision making and provide the underpinnings for economic evaluations, decision analysis, and technology assessment. Variations, including shared decision making, often focus on who should be making decisions but retain the assumptions of rationality. In contrast, descriptive models often emphasize psychological factors, including heuristics and biases. At the policy-making level, however, the recognition of the difficulties in constructing social welfare functions has led to intermediate models with both normative and descriptive elements, including bounded rationality, incrementalism, and mixed scanning.

Normative Theories

Rational Decision Making

Rational choice theory assumes that individuals act to maximize their own utility. A rational individual must therefore

1. determine the range of possible actions that might be taken,
2. determine the possible outcomes that might result from each of these actions,
3. affix a probability to each possible outcome (these must sum to 1.0),
4. affix values to the costs and consequences of each possible outcome, and
5. do the math.

The rational choice will be the one that produces the "best" outcome, as measured in terms of costs and consequences.

Rational decision making is highly data-intensive. It requires a decision maker to collect extensive information about all potential choices, outcomes, costs, and consequences. He or she must be able to order his or her preferences for different outcomes, and these preferences must satisfy the requirements of being complete (i.e., all potential outcomes are assigned preferences) and transitive (i.e., if someone prefers A to B, and B to C, he or she must prefer A to C). In the real world, these assumptions are often unrealistic.

Economists have adopted the theory of revealed preferences to omit some of these steps. Rather than attempt to measure preferences directly, this approach assumes that if someone has chosen a particular outcome, he or she must, by definition, prefer it to the alternatives. Associated with Paul Samuelson, this approach has been highly influential in the study of consumer behavior. It is also tautological and does not leave much room for improving choices (e.g., through providing additional information).

Rational Choice in Medical Decision Making

Decision Analysis

Medical decision making relies heavily on rational choice theory. One common way of analyzing treatment choices, decision analysis, employs the same structure. Constructing a decision tree requires specifying the possible actions ("choice nodes"), specifying the possible outcomes of each action ("chance nodes"), attaching probabilities to each outcome (which must sum to 1.0), and then affixing costs and consequences to each outcome. The tree is then "folded back" by computing the expected value at each node by multiplying the probability by the costs and by the consequences.

For example, in their five-part primer, *Medical Decision Analysis*, Allan Detsky and colleagues work through the example of how to model the choice of management strategies for patients presenting with clinical features that suggest giant cell arteritis (GCA). In this simplified model, the only treatment considered is treating with steroids, which can involve side effects. The rational model they employ thus involves a choice between three possible actions at the choice node—treating, not treating, and testing and treating only if the test result is positive. The possible outcomes can be simplified to four possibilities, depending on whether or not there was an adverse outcome as a result of the disease (in that case, blindness), and whether or not the person had side effects as a result of the treatment. Note that some of these outcomes cannot occur on some branches—for example, someone who did not receive treatment could not experience any outcomes involving side effects. The next step for the decision maker is to determine how likely each of these possible outcomes would be at each choice node (e.g., how likely would an untreated individual with those symptoms be to experience blindness if the person was not treated). Next, the decision maker would affix costs and utilities to each possible outcome. For example, these papers assigned a value of 1.0 to the state with no disease and no side effects, and a value of .5 to the state of having the disease without treatment (or side effects) but ending up with blindness. Sensitivity analysis can be used to modify these values (e.g., change the probability of adverse outcomes or the value attached to particular outcomes) and see how much they affect the resulting choices.

One way to simplify decision trees is to see whether any alternatives are "dominated" by others. Dominated choices are clearly inferior. In a condition of strong dominance, other alternatives are both less costly and of greater benefit. Rational decision makers can accordingly "prune" their decision trees to eliminate all dominated alternatives.

There is an extensive literature relating to how best to model these decisions (including the use of

Markov models) and how to compute costs and consequences. Sensitivity analysis can allow systematic variation in the values assigned to probabilities and outcomes. The underlying assumptions, however, are of rational decision makers maximizing their expected utilities.

Economic Analysis

Economic analysis refers to a family of related methods for weighing costs against consequences. All involve costing the potential outcomes; they vary only in how they assess consequences.

1. *Cost minimization* assumes that the outcomes are identical. In that case, it is not necessary to value them. The decision can be based solely on costs, and a rational decision maker will select the lowest-cost alternative.
2. *Cost benefit* assumes that consequences can also be valued in monetary terms. In that case, the rational decision maker will determine return on investment and select the alternative that produces the highest ratio of consequences to costs.
3. *Cost-effectiveness* assumes that consequences can be valued in a single, albeit nonmonetary, measurement of outcome. Again, the rational decision maker will select the alternative producing the highest ratio of consequences to costs.
4. *Cost utility analysis* is a variant of cost-effectiveness, which computes the "utility" attached to each outcome (on a scale of 0 to 1).

Again, there are many details about how to conduct these analyses, including how to value costs and consequences occurring in the future (e.g., discounting) and how to incorporate different ways of valuing risk. The underlying model, however, continues to assume that rational individuals will act to maximize their expected return, however defined and measured.

Technology Assessment

Technology assessment shares the underlying premise that rational individuals will seek to maximize outcomes for the given inputs. It can be considered a subset of economic models, and presents similar variation in which costs to include (and whose costs), and how to measure consequences. Modern technology assessment is heavily influenced by such organizations as the Cochrane Collaboration and places considerable emphasis on ensuring that data are of high quality. Accordingly, there is often considerable dispute as to where to gather the data and what counts as evidence. Nonetheless, the underlying model remains rational choice.

Decision Makers

Multiple Decision Makers

An additional complexity occurs if there are multiple decision makers. In that case, the decision makers must be able to determine preference orderings that apply to the society. These are referred to as *social welfare functions;* Kenneth Arrow won a Nobel prize for demonstrating the General Possibility Theorem, which proves that, under many circumstances, it is not possible to construct a transitive preference ordering for a society, even given that all members of that society have individual preference orderings satisfying this requirement. For that reason, choice theories for multiple decision makers differ from those for individuals.

Shared Decision Making

If individuals are considered consumers, then the person paying for a particular service should be sovereign. Professionals may provide expert advice but would not determine the preference ordering. If there are externalities, however, such that one person's choice affects the outcomes for others, it is less simple to decide whose preferences should count. If costs are pooled (e.g., through insurance or public financing), then presumably those paying would have some say in the matter. If the consumer is misinformed (e.g., wants a clinical intervention where professionals do not believe that the benefits outweigh the risks), again, there may be disputes about whose preferences should matter. Note that these models do not necessarily require that there be a social welfare function but do require some methods for dispute resolution. Raisa Deber and colleagues have suggested distinguishing between *problem-solving* tasks (defined as preference-independent, where expertise is required) and *decision-making* tasks (which involve deciding based on personal preferences). A substantial literature has attempted to examine shared decision

making in medicine; note that it assumes that patient preferences should be decisive. These models can be seen as subsets of rational models.

Descriptive Models

Heuristics and Biases

In contrast, another set of models seeks to understand how choices are actually made. These models draw heavily on psychology. One key literature examines the simplifying assumptions (often termed *heuristics and biases*) often made by individual decision makers. Even here, the strong dominance of rational decision making persists; the theory of cognitive dissonance stresses that people tend to be adept at justifying choices they have made, even when this requires distorting facts to convince themselves that they are being rational.

Other descriptive models examine how decisions are made within groups and how various pressures exist to influence the choices made.

Modification of the Model

Bounded Rationality

Another set of modifications recognize that it may not be rational to collect full information. Whereas a rational decision maker seeks to maximize, what Herbert Simon terms *administrative man* seeks to "satisfice" and select the first alternative that is "good enough." Some of this literature incorporates the information about cognitive limitations.

Incrementalism and Mixed Scanning

Policy analysts have suggested that most policies begin with what is. Charles Lindblom suggested that policy making is usually incremental, consisting of a series of small adjustments to existing policy resulting from a series of "successive limited comparisons." In this model, rather than beginning by setting goals and objectives, policy makers will perform limited analyses of immediate issues and implement a series of small decisions. This model has the advantage of being low-risk; smaller decisions have fewer short-term consequences and are less likely to elicit opposition. Continuous feedback allows the policies to be modified as required. The weakness is that the big picture is rarely examined. The incremental model is both normative and descriptive. It describes how policy usually occurs. However, some authors also consider it a desirable normative model, particularly under circumstances where bounded rationality is likely to be appropriate, where the risks of error are high, or where interest groups are highly invested.

Amitai Etzioni has suggested an intermediate stance he termed *mixed scanning*. In this model, incrementalism is the default mode, but policymakers are scanning for areas where more in-depth, rational decision making could be beneficial. His analogy is the job of a sentry, who scans the landscape to see where there is movement that calls for further investigation.

Social psychologists have noted the importance of the system within which decisions are made. The patient safety movement, for example, has noted that improving clinical performance is less a matter of removing bad apples than it is a matter of ensuring that systems are set up to encourage optimal performance. These models thus draw on the descriptive material and seek to set up models within which optimal choices are more likely to be made.

Raisa Deber

See also Applied Decision Analysis; Bounded Rationality and Emotions; Cognitive Psychology and Processes; Decision Psychology; Heuristics; Treatment Choices

Further Readings

Berwick, D. M. (1989). Continuous improvement as an ideal in health care. *New England Journal of Medicine, 320*(1), 53–56.

Deber, R., & Goel, V. (1990). Using explicit decision rules to manage issues of justice, risk, and ethics in decision analysis: When is it not rational to maximize expected utility? *Medical Decision Making, 10*(3), 181–194.

Deber, R., Kraetschmer, N., & Irvine, J. (1996). What role do patients wish to play in treatment decision making? *Archives of Internal Medicine, 156*(13), 1414–1420.

Detsky, A. S., Naglie, G., Krahn, M. D., Naimark, D., & Redelmeier, D. A. (1997). Primer on medical decision analysis: Part 1. Getting started. *Medical Decision Making, 17*(2), 123–125.

Detsky, A. S., Naglie, G., Krahn, M. D., Redelmeier, D. A., & Naimark, D. (1997). Primer on medical decision analysis: Part 2. Building a tree. *Medical Decision Making, 17*(2), 126–135.

Gold, M. R., Siegel, J. E., Russell, L. B., & Weinstein, M. C. (Eds.). (1996). *Cost-effectiveness in health and medicine*. New York: Oxford University Press.

Hastie, R., & Dawes, R. M. (2001). *Rational choice in an uncertain world*. Thousand Oaks, CA: Sage.

Krahn, M. D., Naglie, G., Naimark, D., Redelmeier, D. A., & Detsky, A. S. (1997). Primer on medical decision analysis: Part 4. Analyzing the model and interpreting the results. *Medical Decision Making, 17*(2), 142–151.

Naglie, G., Krahn, M. D., Naimark, D., Redelmeier, D. A., & Detsky, A. S. (1997). Primer on medical decision analysis: Part 3. Estimating probabilities and utilities. *Medical Decision Making, 17*(2), 136–141.

Naimark, D., Krahn, M. D., Naglie, G., Redelmeier, D. A., & Detsky, A. S. (1997). Primer on medical decision analysis: Part 5. Working with Markov processes. *Medical Decision Making, 17*(2), 152–159.

Classification and Regression Tree (CART) Analysis

See Recursive Partitioning

Clinical Algorithms and Practice Guidelines

Clinical algorithms and practice guidelines may be viewed as a targeted effort to provide the best clinical advice about specific management conditions. They are most useful if clinicians incorporate them as additional tools to specifically improve patient outcomes while offering holistic clinical care to patients.

Clinical Algorithms

Definition

Algorithms are branching-logic pathways that permit the application of carefully defined criteria to the task of identifying or classifying different types of the same entity. Clinical algorithms are often represented as schematic models or flow diagrams of the clinical decision pathway described in a guideline.

Clinical findings, diagnostic test characteristics, and treatment options are abbreviated into their basic components. Algorithmic flow diagrams are then constructed as branching logical pathways with decision points represented as yes/no nodes. Such a flowchart sequence is useful in identifying or classifying entities based on carefully devised criteria (Figure 1). Application of clinical algorithms is most defensible when the evidence supports choices in the decision tree. Although very useful for clinical decision making, algorithms cannot account for all patient-related variables. Therefore, algorithms are not intended as a substitute for the clinician's best judgment.

Proposed Standards

The Society for Medical Decision Making Committee on Standardization of Clinical Algorithms has proposed certain standards for construction of clinical algorithms. Their technical note has specific recommendations on the types and shapes of algorithm boxes (clinical state box—*rounded rectangle*; decision box—*hexagon*; action box—*rectangle*; and link box—*small oval*), titles, abbreviations, annotations and their format, and schemes for arrows, numbering, and paging.

Classification

Simple Classification Algorithms

Simple classification algorithms serve only as diagnostic aids and do not advocate any clinical intervention. They contain question nodes (algorithmic boxes) leading to yes or no exit arrows.

Management Algorithms

Management algorithms encompass both diagnostic and treatment modalities. They employ decision-relevant yes/no question nodes. Each question node in turn leads to an instruction node, denoted by a single exit arrow. Instruction nodes advocate for specific interventions. Thus, patients get classified into distinct clinical subgroups that would benefit from specifically targeted management strategies.

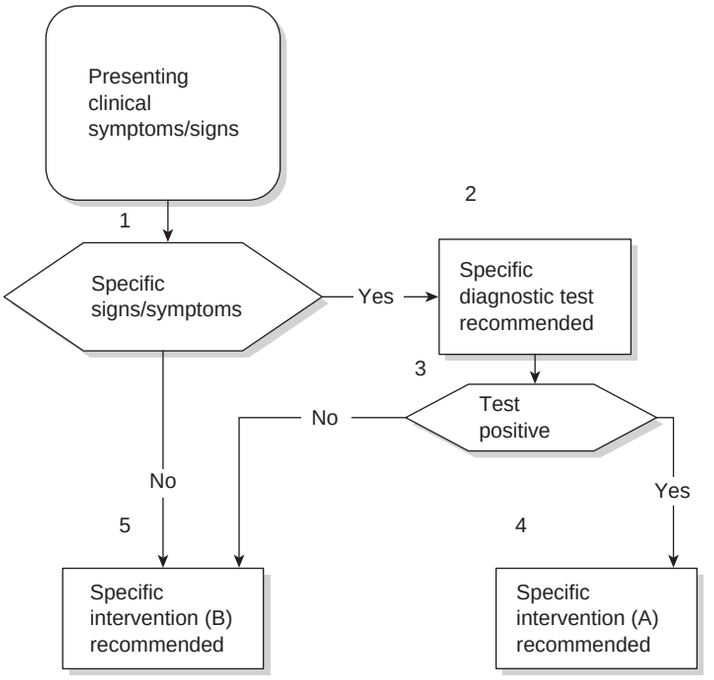


Figure 1 Schematic model of a management clinical algorithm

Note: Management algorithm for a patient with a hypothetical set of presenting symptoms/signs.

Outcome studies are essential to providing support for each management strategy.

Validity and Flexibility

It is often argued that algorithms are not always backed by empirical data, are infrequently linked to the literature, and are not adequately flexible when dealing with clinical uncertainties. To circumvent these inadequacies, two important modifications have evolved.

First, in an attempt to enhance the validity, *annotated management algorithms* have been devised. Here, each node concerned with specific findings, characteristics, or interventions is annotated with the intent of summarizing the guideline's detailed textual material. The textual material is in turn replete with citations. Thus, in this fact-based approach, the algorithm links the recommendations of the guideline to systematic literature reviews or, when appropriate, to expert consensus. The Agency for Health Care Policy and Research guideline development process exemplifies this approach.

Second, *counseling and decision nodes* are specifically implanted in the algorithm where therapeutic decisions are expected to be constrained due to a gap in current knowledge. This is particularly relevant when patient preferences vary with respect to two (or more) different therapeutic options (e.g., medical vs. surgical management). At each decision node, the expected outcome associated with each option is indicated to the extent possible. Thus, counseling and decision nodes facilitate therapeutic deliberations between physician(s) and patient.

Methodological Considerations

Four methodological issues are commonly cited as influencing the development of annotated management algorithms.

1. *Selection of a descriptor variable:* The best descriptor variables are easily observed and have great discriminatory power to categorize patients into different subtypes suited for separate management strategies. Discriminatory power of the variable

is gauged in terms of its sensitivity, specificity, and predictive value. So, ideally, these data should result from controlled studies. Often, evidence is sparse and expert panels develop a consensus about the best discriminatory variables to be used, based on certain stated rationales. Nevertheless, future research should aim to address such gaps in knowledge.

Variables can be optimally selected using sound statistical techniques. Regression analyses are the most common statistical methods used for estimating the discriminatory power of a variable. They quantify the impact of a given variable on outcome probabilities. Recursive partitioning, on the other hand, is an algorithmic strategy that identifies homogeneous, meaningful patient subtypes for guideline/algorithm development. Here, the patient population is reclassified into ever smaller subgroups based on Boolean combinations of variables. An example cited in the literature is as follows: age > 65 years and hematocrit < 30 and systolic blood pressure > 160 mmHg.

2. *Incomplete literature and consequent uncertainties:* Annotated algorithms depicting the lowest common denominator on which there is reasonable consensus may be used to clarify algorithms when uncertainty exists. In addition, the expert panel may identify and explicitly delegate possible management strategies into one of the four prototype categories defined in the literature: (1) necessary care (practice standard, recommendation), (2) appropriate, but not necessary care (option), (3) equivocal care (not recommended), and (4) inappropriate care (recommendation against use).

3. *A dilemma may be encountered when attempting to integrate qualitative descriptor variables* (e.g., increased postvoid residual urine) *with quantitative information* (e.g., specific cutoff volume for increased postvoid residual urine): Often, such an approach may not be backed by literature or panel consensus. Such dilemmas should be highlighted in the algorithm. Additionally, the annotation should include a tally of panelists' quantitative recommendations and any ensuing discussion of factors that permit the calculation of such variables.

4. *Optimal representation of health outcomes:* Health outcomes must be precisely defined and properly annotated, and they should be relevant to the algorithm. Therapeutic side effects and their estimated occurrence risk should also be reported.

Technical Suggestion

Algorithms should be logically and succinctly laid out with carefully selected nodes representing the lowest common denominator. Nodes should apply to a significant proportion of patients. Otherwise they should be incorporated as annotations or combined with other nodes to reduce excessive and unnecessary detail.

Practice Guidelines

Definition

Clinical practice guidelines (CPGs) attempt to transform evidence into practice to improve patient outcomes. The approach of evidence-based CPGs is to define clinical questions, review current evidence, and determine grades of recommendation. Patient questions are also addressed. While CPGs reflect a broad statement of good practice with little operational detail, *protocols* are the result of their local adaptation.

Contents of High-Quality Clinical Practice Guidelines

This is best exemplified by the National Guideline Clearinghouse (NGC) Guideline Summary Sheet, available from the NGC's Web site. The Web site also provides links to NGC's Brief and Complete Guideline Summary, Guideline Comparison, Guideline Synthesis, and Classification Scheme.

The NGC Complete Guideline Summary describes the guideline's title, scope (includes disease/conditions, intended users, and target population), methodology, recommendations (major recommendations and clinical algorithms), evidence supporting the recommendation(s) and benefits and risks of its implementation, contraindications, qualifying statements, implementation strategy, and Institute of Medicine (IOM) national healthcare quality report categories. In addition, identifying information and availability of the guideline is provided, including details about bibliographic sources, adaptation from any previous guidelines, date of release, guideline developers,

committees and endorsers involved, funding source(s), financial disclosure, guideline status and availability, and patient resources.

Characteristics of High-Quality Clinical Practice Guidelines

The IOM expert committee on guidelines has identified validity as the most important attribute. Validity is based on strength of scientific evidence underlying the recommendations and their impact on health and cost outcomes. Reproducibility, reliability, clinical applicability, clinical flexibility, cost-effectiveness, and clarity are other key aspects. CPGs should be a multidisciplinary process, with documentation of participants, assumptions, and methods, and be subjected to *scheduled reviews*.

Strength of Recommendation Taxonomy

Strength of recommendation is graded based on evidence into A (consistent and good-quality patient-oriented evidence), B (inconsistent or limited quality patient-oriented evidence), and C (consensus, usual practice, opinion, disease-oriented evidence, or case series for studies of diagnosis, treatment, prevention, or screening). The quality of a study measuring patient outcome(s) is similarly graded into levels 1 (*good quality evidence*), 2 (*limited quality evidence*), and 3 (*other evidence*).

Grades of Recommendation

It is controversial whether the level of evidence and grade of recommendation (GOR) should be standardized across CPGs in different areas. One such GOR proposed by the Joint Committee of Development of Clinical Practice Guidelines for the Treatment of Stroke is as follows: A (*strongly recommended*), B (*recommended*), C1 (*acceptable although evidence is insufficient*), C2 (*not recommended because evidence is insufficient*), and D (*not recommended*).

Role and Utility of Clinical Practice Guidelines

CPGs play an important role in enhancing clinicians' knowledge by keeping them abreast of the latest developments in medicine. This is intended to change their attitude about standard of care and shift their practice pattern, leading to improved patient outcomes. Healthcare policy makers can use CPGs to assign resources to the most needed areas. CPGs also guide plan administrators and insurers to arrive at reimbursement decisions for patients. In addition, public and patient education, research priorities, and medicolegal issues are influenced by CPGs.

Perspective of Clinicians and Patients

Most clinicians agree that CPGs are helpful educational tools intended to improve quality of care. Nevertheless, CPGs have been variously described as anti-intellectual, impractical, limiting clinical autonomy and discretion, cost-cutting, standardizing practice around the average, and causing increased litigation. Apart from negative attitudes and resistance to change, other barriers to CPGs include administrative and financial obstacles as well as limited time and resources for education and implementation. Sometimes, patients' choices may also be in conflict with the guidelines.

Successful Implementation

Successful implementation involves organizational commitment and raising awareness among intended users through dissemination of information (conferences, meetings, and publications), alongside education and preparation of staff. Other useful strategies include use of local clinical leadership; inclusion of CPGs within the contracting process; support of practitioners, including information giving and feedback; reminders and incentives; audit and feedback of results; and patient/client-mediated interventions.

Chenni Sriram and Geoffrey L. Rosenthal

See also Decision Making in Advanced Disease; Decision Modes; Decision Tree: Introduction; Diagnostic Process, Making a Diagnosis; Recursive Partitioning

Further Readings

American College of Physicians. (2009). *Algorithms* [Electronic version]. Retrieved October 7, 2008, from http://www.acponline.org/clinical_information/guidelines/process/algorithms

Duff, L. A., Kitson, A. L., Seers, K., & Humphris, D. (1996). Clinical guidelines: An introduction to their

development and implementation. *Journal of Advanced Nursing, 23,* 887–895.
Ebell, M. H., Siwek, J., Weiss, B. D., Woolf, S. H., Susman, J., Ewigman, B. D., et al. (2004). Strength of Recommendation Taxonomy (SORT): A patient-centered approach to grading evidence in the medical literature. *American Family Physician, 69,* 548–556.
Greer, A. L., Goodwin, J. S., Freeman, J. L., & Wu, Z. H. (2002). *International Journal of Technology Assessment in Health Care, 18,* 747–761.
Hadorn, D. C. (1995). Use of algorithms in clinical guideline development. In *Clinical practice guideline development: Methodology perspectives* (AHCPR Pub. No. 95-0009, 93-104). Rockville, MD: Agency for Health Care Policy and Research.
Lohr, K. N., Eleazer, K., & Mauskopf, J. (1998). Health policy issues and applications for evidence-based medicine and clinical practice guidelines. *Health Policy, 46,* 1–19.
Nakayama, T. (2007). What are "clinical practice guidelines"? *Journal of Neurology, 254*(Suppl. 5), 2–7.
Natsch, S., & van der Meer, J. W. M. (2003). The role of clinical guidelines, policies and stewardship. *Journal of Hospital Infection, 53,* 172–176.
Society for Medical Decision Making, Committee on Standardization of Clinical Algorithms. (1992). Proposal for clinical algorithm standards. *Medical Decision Making, 12,* 149–154.
Welsby, P. D. (2002). Evidence-based medicine, guidelines, personality types, relatives and absolutes. *Journal of Evaluation in Clinical Practice, 8,* 163–166.

Cognitive Psychology and Processes

Cognitive psychology is the study of the thinking mind. It emerged as a field of psychology in the 1980s and includes perception, attention, memory, decision making, problem solving, reasoning, and language among its areas of study. Using theory and empirical study, cognitive psychology aims to understand the cognitive processes used and what influences their use. In medical decision making, for decisions relevant to individuals, systems, and society, understanding the cognitive processes that people typically use, and why, would help (a) developers of decision support interventions target the interventions most effectively and (b) identify outcomes appropriate for judging the effectiveness of particular decision-making strategies.

Decision making in cognitive psychology focuses on how people make choices. The field is distinct from problem solving, which is characterized by situations where a goal is clearly established and where reaching the goal is decomposed into subgoals that, in turn, help clarify which actions need to be taken and when. In the medical world, making a diagnosis, for example, typically requires problem-solving processes. Decision making is also distinct from reasoning, which is characterized as the processes by which people move from what they already know to further knowledge. Although historically, decision making, problem solving, and reasoning were studied independently within cognitive psychology, it is recognized that in complex decisions both reasoning and problem-solving processes can be required to make a choice.

Decision making requires the integration of information with values. The information in a medical decision is often about a health state and the options for addressing it. Values are the qualities that underlie worth or desirability. A decision maker's values determine the particular subset of information that is most germane to his or her decision. Although both information and values are part of most medical decisions, the particular cognitive processing required can vary significantly from one decision to another.

Levels of Decisions

Four levels of decisions have been described—the higher the level, the greater the energy required and the more complex the decision processes.

Level 1—simple, familiar decisions: They are made quickly and largely automatically (unconsciously). An example occurs when people prone to headaches automatically reach for a particular painkiller in response to early headache signs.

Level 2—decisions that use static mappings when evaluating options: An example occurs when people choose particularly invasive treatments only because they believe that the more a treatment makes one suffer, the more likely it is to be successful.

Level 3—decisions that belong to a class of decision that is familiar to the decision maker, although the

particular instance is not and can include options that have both pros and cons: An example occurs when people choose a family doctor after losing their doctor for the third time and therefore know what is important to them in the decision, but they need to learn about the new choices.

Level 4—decisions in unfamiliar situations when the choices are also not familiar: These decisions often require problem-solving (and possibly reasoning) processes to learn about the situation and the options. An example is a person, newly diagnosed with a relatively unfamiliar medical condition, needing to choose a treatment.

Cognitive Processes

Making decisions beyond lower-level decisions is typically protracted in time, requiring many types of cognitive processes. Ola Svenson is an early pioneer in describing decision processes, and he still provides one of the most comprehensive descriptions of those processes. He suggests that the process goal of decision making is to select one option that is superior enough over the other options that it can protect the decision maker from experiencing cognitive dissonance (discomfort from having values that conflict with the decision) and regret later. He describes three phases of processing: the initiation phase, differentiating the options, and after the decision.

Initiation Phase

Decision-making processes begin with the decision maker establishing the goal(s) of the decision and identifying options and attributes of the options that are important. Salient aspects of the situation tell the decision maker where to start. This phase structures the decision in the decision maker's mind. Therefore, the early-identified salient aspects can have important implications for the processing that follows. For example, a diagnosis of cancer generating fear of death can trigger the automatic elimination of a do-nothing-for-now option. Early screening of options is not unusual in situations where there are many options.

The initiation phase can include singling out one option. Sometimes it is a reference option, against which other options can later be compared. When there are many options to consider, the singled-out option tends to be a preliminary preferred choice. Such a strategy limits energy demands that can become huge very quickly in situations where there are multiple options.

In addition to possible screening or selection of a preliminary preferred option, this early stage can involve information search. Exactly what information is searched for and retained can follow, to some extent, the salient attributes mentioned above.

Differentiating the Options

The major cognitive processing involved in decision making focuses on differentiating the options, one from the other. Svenson has identified three types of differentiating processes:

1. *Holistic differentiation* is quick, automatic (not within conscious control) processing.
2. *Structural differentiation* involves changes to the way the decision problem is represented in the mind of the decision maker. The structure can be altered by changing
   a. how attractive a particular aspect of an option is judged to be (e.g., shifting from a judgment that saving 10 of 100 people is not significant to considering that it is significant),
   b. the importance given to a specific attribute (e.g., shifting from being very concerned about possible incontinence to not being concerned about it),
   c. the facts about an option (e.g., shifting from believing that most men diagnosed with prostate cancer die of the disease to learning that most men do not die of their prostate cancer), and
   d. the particular set of attributes used to describe the options (e.g., shifting from interest in only treatments' effects on survival to their impact on quality of life).
3. *Process differentiation* involves using information about the options to arrive at a decision, following decision rules. Some rules involve combining all available information in a process of weighing pros against cons, while other rules involve using only some of the information, such as judging an attribute against a threshold.

For complex decisions, differentiation can be intermingled with information searches. In new situations, the decision maker may also need to discover which values are relevant to the decision, sometimes needing to figure out what their values are and, when values are in conflict with one another, their relative weightings. Because these processes are extended in time, research at different time points can suggest that values shift from one time to the next. Evidence suggests, however, that the processes eventually stabilize.

After the Decision

After the decision is made, the decision maker continues cognitive processing of the decision. Postdecision processes can include both the structural and process differentiation described above. Both implementation of the decision and outcomes of the decision can also be followed by yet further differentiation, though the specifics of what is processed may be altered. The postdecision processes manage the emotional consequences of having made the decision, potential cognitive dissonance, or regret.

Factors That Complicate Cognitive Decision Processes

Several factors about decision situations complicate both the actual processes used and our ability to learn about what is being done.

Uncertainty

Situations with information missing about a potential outcome are often distinguished from situations where an outcome is known but has a less-than-certain chance of occurring. People find it hard to act in the first type of situation; bad news is better than no news. In medical decision making, when making a decision for an individual, the two types of situations are not very different; knowing that, of a particular group of people, some will experience an outcome but others will not does not clarify what will happen to the individual. Discomfort with uncertainty can lead some patients, for example, to decide that they know what will happen to them.

Structure of the Environment

People are sensitive to the structure of the environment when judging a situation; thus, changing one aspect of the environment can change responses. Framing effects, where responses shift according to how a situation is described, is one example of such a change; an example of a framing effect in medical decisions is the response shift seen when an outcome is described as numbers of lives saved rather than numbers of lives lost. People's sensitivity to the environment means that asking questions in one way can produce different results compared with asking the apparently same question in a different way. It has been suggested, for example, that issues around compatibility between inputs (how the problem is described, an environmental structure) and outputs (the responses requested, another environmental structure) contribute to a broad range of what have been identified as nonnormative "biases" in human decision making.

Stress

Stress describes people's responses to what they judge to be threats. While mild stress can actually improve cognitive performance, high stress is generally seen as detrimental. It can increase distraction, making it harder to focus attention that can, for example, reduce the numbers of options or the numbers of attributes of each option being considered. It can also compromise the organization of the information in the decision maker's mind.

Intuition

Intuition has been defined as thinking processes that occur when the input is mostly knowledge acquired automatically (without conscious control) and the output is a feeling that can then be used as the basis for judgments and decisions. In some types of situations, intuitive decisions are more accurate than deliberate (with conscious control) decisions, but in other types of situations, deliberate decisions are more accurate. Intuition seems favored when people have prior experience with the relevant options and when the automatically acquired knowledge matches the demands of the decision. Intuitive attitudes are more likely to reflect the entire corpus of information acquired

about the options, whereas attitudes related to deliberate learning are more likely to reflect only part of that information.

Heuristics

Heuristics are general rules of thumb that people use in cognitive processing to reduce mental energy demands. While the general thinking has been that using heuristics reduces the accuracy of processing, evidence now suggests that in some situations heuristics can actually improve the accuracy of decisions. Heuristics include simple rules about how to search for more information, when to stop the search, and estimating the likelihood of an event. For example, the representative heuristic can lead a teenager to ignore warnings about smoking because the typical image is that people with lung cancer are old.

Why Understand Cognitive Decision-Making Processes?

Understanding the cognitive processing used can naturally provide several types of guidance in the field of medical decision making. It can help identify the specific challenges that make a particular decision difficult, which, in turn, clarifies how to make decision support interventions most effective. Understanding the cognitive processing also reveals important complexities in human behavior that should be considered when creating interventions. For example, sensitivity to environmental structure implies that how information is presented (not just what is presented) can make a big difference in whether an intervention is helpful or not.

Understanding the particular processes people use and why they use them can also help guide selection of outcomes that indicate good quality decisions. For example, people naturally aiming their decision processes to protect them from experiencing cognitive dissonance and postdecisional regret suggests that measures of value concordance and of regret are important quality indicators.

Deb Feldman-Stewart

See also Decision Making and Affect; Decision Rules; Regret

Further Readings

Gigerenzer, G., Todd, P. M., & ABC Research Group. (1999). *Simple heuristics that make us smart*. New York: Oxford University Press.

Plessner, H., & Czenna, S. (2008). The benefits of intuition. In H. Plessner, C. Betsch, & T. Betsch (Eds.), *Intuition in judgment and decision making*. New York: Lawrence Erlbaum.

Selart, M. (1997). Aspects of compatibility and the construction of preference. In R. Raynard, W. R. Crozier, & O. Svenson (Eds.), *Decision making: Cognitive models and explanations*. London: Routledge.

Svenson, O. (2001). Values and affect in human decision making: A differentiation and consolidation theory perspective. In S. L. Schneider & J. Shanteau (Eds.), *Emerging perspectives on judgment and decision making research*. Cambridge, UK: Cambridge University Press.

COINCIDENCE

A coincidence is a random co-occurrence of two or more events that are perceived to be meaningfully associated with each other, even though there is no meaningful causal relationship linking them. A collision between an ambulance carrying an injured bullfighter and a cattle truck would constitute a coincidence, while internal bleeding following ingestion of broken glass would not. The need to distinguish true associations from coincidences is critical to good medical decision making, yet the human mind is ill equipped to make this distinction.

Co-occurrences of events can be perceived as meaningful when they happen along a number of dimensions, such as time (e.g., when a patient develops symptoms shortly after taking a drug), space (e.g., when multiple cases of a rare disease occur in the same town), or heredity (e.g., when several members of a family tree are found to have the same disorder).

While co-occurrences often indicate the existence of a direct causal relationship or a common underlying factor, many are simply the result of chance, and their constituent events should be considered independent of each other. However, determining whether a co-occurrence reflects meaningful

or random covariance is often difficult. In fact, research shows that the human mind is limited in its ability to distinguish meaningful associations from coincidences. People (even those with medical degrees) tend to commit predictable errors when trying to distinguish random chance events from meaningful causal processes. As a result, we often overreact to coincidences and underreact to co-occurrences that deserve our attention.

Some events are, by their very nature, especially likely to capture our attention and generate an emotional response. Accordingly, they are more likely to be initially encoded in memory and are later more accessible for recall. As a result, we tend to notice their co-occurrences much more and infer more from these than from co-occurrences of other events. For example, we are overly influenced by the probability of each event occurring on its own. The more unlikely each of the events is thought to be, the more surprising we find their individual occurrences, and this makes their *co*-occurrence seem all the more surprising and meaningful. Taking vitamins and experiencing mild stomachaches are both relatively common events, so their co-occurrence is likely to go unnoticed. In contrast, taking a new experimental drug and experiencing acute abdominal pains are both relatively uncommon events, so their co-occurrence is likely to raise suspicion of a causal link. Another closely related factor is the number of events co-occurring: The greater the number of events that co-occur, the more we tend to find this co-occurrence meaningful. A physician is more likely to suspect the presence of a disease when his or her patient shows five unusual symptoms than when the patient shows two unusual symptoms.

While these two factors can provide rational bases for judging the meaningfulness of co-occurrences (though not always), others are much less justifiable. For example, co-occurrences are perceived to be more indicative of a causal relationship when they are experienced firsthand than when they are experienced by others. This helps explain why patients and their loved ones are more likely to see, in the co-occurrence of symptoms, the threat of a serious medical condition, where the physician sees harmless coincidence.

The need to distinguish meaningful co-occurrences from simple coincidences regularly arises across a variety of medical decision-making contexts. Physicians and other medical professionals are often confronted with the difficult task of recognizing when co-occurrences are meaningful or coincidental: Does the simultaneous occurrence of certain symptoms imply the presence of a disease, or did it happen by chance? Does the apparent relationship between administration of a new medical drug and improved health signal effectiveness, a placebo effect, or a meaningless coincidence? Should a physician be concerned when his or her patient reports experiencing unpleasant symptoms following a medical procedure, or is this mere happenstance? Are multiple outbreaks of a rare disease within a small geographic area the sign of a growing epidemic or just random clustering?

Separating coincidence from causality is a problem that also confronts patients and nonmedical professionals: Are feelings of nausea following a dining experience the first signs of serious food poisoning, which calls for a trip to the emergency room, or are they unrelated? Are the higher rates of surgical death associated with a particular hospital the result of malpractice or bad luck? Even when medical professionals are able to recognize coincidences, they must confront the objections of patients and loved ones who are quick to see meaningful associations in the co-occurrence of significant events (e.g., two family members dying from a rare disease) and resistant to the possibility that these could happen by chance alone.

A number of real-life examples illustrate the importance of distinguishing causation from coincidence. One striking case is the controversy that erupted in a number of Western countries, when many parents were convinced, by anecdotal evidence, that vaccination for measles, mumps, and rubella (MMR) caused autism. A number of studies were carried out in response to the resulting public outrage, with the majority of them finding no association between MMR vaccination and the occurrence of autism. As it turns out, children tend to be diagnosed with autism around the time they turn one, which also happens to be when they are administered the MMR vaccine. As result, a number of children who would have been diagnosed with autism, even without the vaccine, received this diagnosis shortly after receiving the MMR vaccine, leading many parents to perceive a direct link between the two.

Because of biases in human probabilistic reasoning, medical professionals and their patients

are subject to misunderstanding coincidental occurrences as causally related. For this reason, teaching medical professionals to be aware of these biases is a prerequisite for good medical decision making and effective communication with patients.

Christopher Y. Olivola and Daniel M. Oppenheimer

See also Biases in Human Prediction; Causal Inference in Medical Decision Making; Judgment; Probability Errors

Further Readings

Gilovich, T. (1993). *How we know what isn't so: The fallibility of reason in everyday life.* New York: The Free Press.

Griffiths, T. L., & Tenenbaum, J. B. (2007). From mere coincidences to meaningful discoveries. *Cognition, 103,* 180–226.

Hastie, R., & Dawes, R. M. (2001). *Rational choice in an uncertain world: The psychology of judgment and decision making.* Thousand Oaks, CA: Sage.

Nickerson, R. S. (2004). *Cognition and chance: The psychology of probabilistic reasoning.* Mahwah, NJ: Lawrence Erlbaum.

Complexity

Complexity science is the study of systems characterized by nonlinear dynamics and emergent properties. In contrast, simple mechanical systems are describable by sets of linear equations lending themselves to conventional scientific methods. But living systems and systems composed of living things display and adapt to changes in unpredictable ways. Complexity science studies systems as complete wholes instead of components or subsystems to effectuate better decisions that are more realistic for clinical and policy decisions.

Complexity is a term that describes how system components interact with one another. For instance, conventional medical science may be less able to predict which patients will experience unanticipated side effects from a new drug. Likewise, healthcare delivery and its many elements are not likely to respond predictably to policy or reimbursement changes.

As knowledge of health and illness progressed through the 20th century, Cartesian notions of a mechanical and predictable universe were inadequate to describe some natural phenomena. Researchers believed that multifaceted and complex findings associated with the health sciences might benefit from more advanced or comprehensive frameworks than the mechanical ones typically employed. To this end, a science of complexity was sought to improve predictability and quality of medical decision making with the use of specialized scientific methods.

Healthcare interventions draw on accumulated knowledge and wisdom concerning disease processes, formalized as science, to prevent, ameliorate, or cure conditions. Given that diseases, treatment options, and patients are often complex and unpredictable systems, perhaps clinical decision making could benefit through a deeper understanding of the system dynamics impinging on individual patients to a greater or lesser degree. For instance, complex patient ecologies include in addition to physiology, the cultural, local community, social, psychological, genetic, emotional, and relational domains, all of which can augment or impede treatment.

A science of complexity is attractive because of a potential to describe and predict systems phenomena more congruently with what is known about actual living system attributes and behaviors. For instance, the inputs, processes, and outputs associated with living systems are often described as *nonlinear* since system inputs yield unpredictable outputs. Furthermore, system behaviors may be *deterministic, stochastic,* or *random responses* to environmental challenges and changes; and all types of system responses may appear similar. Also, describing the "essence" of a given system through conventional repeated sampling methods of system outputs may never converge on fixed system *parameters*. Furthermore, living whole systems are *logically irreducible* to description and prediction by simple "reductionist" methods. Slicing a system conceptually or literally for study has limits. At some threshold, the emergent and nonlinear properties of a whole system cease to function normally, and the subject of inquiry is lost.

Important too is the tendency for conventional scientific tools that tend to favor group responses over individual or *idiographic* ones. Emergent or unexpected clinical or policy system behaviors are

likely to be dismissed as measurement errors under conventional research methods.

Complex systems share some characteristics: (a) many nonlinearly interacting components; (b) system ordering initiated from the bottom up; (c) emergent structures to meet environmental demands; (d) self-organization; and (e) nonstationary characteristics. Paul Plsek challenges us when he says that "the real power [in understanding systems] lies in the way the parts come together and are interconnected to fulfill some purpose" (p. 309). Plsek further charts the domain of command and control, chaos, and complexity in Figure 1.

History

The scientific method and discourse since Descartes (1596–1650) progressed under two main assumptions: (1) System components could be analyzed as independent entities and then (2) added linearly together to describe and predict the behavior of the entire system. Thus was the Cartesian mechanistic worldview a radical and welcome departure from the previous *scholastic* forms of inquiry. Cartesian mechanics carried scientific inquiry for nearly three centuries and persists in various forms to this day.

Pierre Simon de Laplace formalized the continuity of a Cartesian "clockwork" universe over time by suggesting that the current system state is a consequence of its state in the moment immediately preceding the current one. Thus, by comprehending all nature's laws, a system's past may be described all the way back to its *initial state* and its future predicted. Over a century later, Henri Poincaré disagreed. He said that subtle factors in complex systems are amplified over time, leading to difficulty in perceiving earlier system state conditions and making long-range prediction of the future impossible. Importantly, both Laplace and Poincaré described deterministic models—but the Laplacian universe was defined as stable and predictable through time and a thorough understanding of individual system components. The universe described by Poincaré was more uncertain and best understood from its most recent states. But Poincaré did assert that complex-appearing system outputs might be produced by simple deterministic mechanisms but that some systems may be so sensitive to initial and slight perturbations that long-term prediction of a system is nearly impossible. While Poincaré's ideas led the way for modern chaos theory, Werner Heisenberg was showing that at least subatomic systems were unpredictable and best describable in probabilistic—or *stochastic*—terms.

Cartesian mechanics (Descartes's *The World*, 1633) became inadequate to the task of explaining scientific observations as the 20th century began. In 1928, Ludwig von Bertalanffy proposed a *general systems theory* to correct Cartesian assumptions of reductionism and linear additivity. By 1951, he had extended general systems theory specifically to biology. In 1956, Kenneth Boulding classically defined general systems theory as "a level of theoretical model-building which lies somewhere between the highly generalized construction of pure mathematics and specific theories of specialized disciplines" (p. 197).

Meanwhile, two other forms of systems science were emerging. Economists adopted system dynamics, which studied information amplified through systems, circular causality, and the self-regulating and self-influencing of systems. System dynamics found that systems can be self-destructive or self-sustaining. Engineers found useful the system properties of information feedback and transmission between system components for describing, predicting, and modifying system behavior—the field of cybernetics.

Ultimately, complexity science emerged from systems theory with computer-based weather simulations by Edward Lorenz in 1961. Lorenz found that complex systems are highly sensitive to initial conditions, making nearly impossible accurate predictions for complex systems, depending on system stability. Thus, according to Kerry Emanuel, Lorenz had driven "the last nail into the casket of the Cartesian universe." Complexity science might offer a more realistic model of systems as stochastic and ever changing, rather than mechanical and stable from the beginning of time. Decision making about systems required a new set of parameters and tools to estimate them.

Complex Systems: Description and Measurement

Ruben McDaniel and others remind us that the terms *complexity* and *chaos* do not refer to the same phenomena. Three types of systems output illustrate the differences: (1) noncomplex deterministic (simple

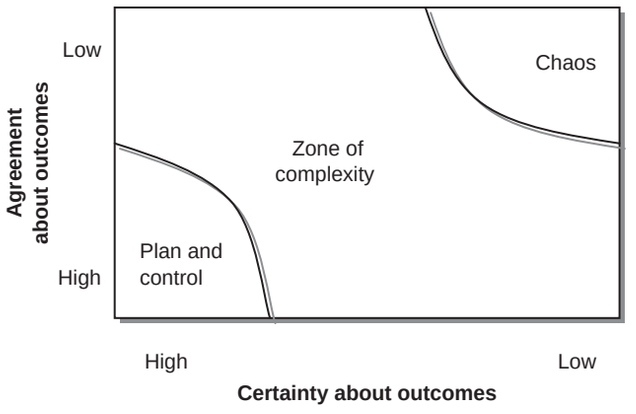


Figure 1 Depiction of the decision domains of plan and control, chaos, and complexity

Source: Stacey, Ralph D., *Strategic Management and Organizational Dynamics*, 2nd Edition, © 1996, p. 47. Adapted by permission of Pearson Education, Inc., Upper Saddle River, NJ.

mechanical), (2) complex deterministic (a chaos or chaotic system), and (3) complex random (from a complex system). Automobile engines are mechanical, noncomplex deterministic systems. All future behaviors of a specific automobile are predictable, given enough data.

Chaotic systems produce complex output that can be generated by a simple deterministic process. The formal definition and notion of a chaotic system is that a chaotic system is very sensitive to initial conditions. Chaotic processes describe electrical heart activity and nerve impulse transmission. The weather is a classic example of a chaotic system.

Chaotic systems, though deterministic, cannot be predicted over the long run. This is because of their sensitivity to initial conditions. Chaotic system behavior "drifts" over the course of time. The best predictor of tomorrow's weather is today's weather, and not the weather of 2 weeks ago. Equations that could predict system behavior yesterday become increasingly unstable or poor predictors into the future. Thus, each chaotic system's behavior is a consequence of, and best predicted by, its behavior in the preceding moment—a property called *dynamical* (see below).

Complex systems produce deceptively simple looking output from either complex or random processes. The problem is that chaotic (complex output/simple generator) and complex systems (simple or complex output/unknown processes) are difficult to distinguish by simple observation. Biological and genetic evolution is an example of an ongoing random process in a complex system.

Not unexpectedly, measurement of complex systems may require some unconventional descriptors and tools. *Fractals* can describe dynamic processes or output outside familiar fixed Gaussian parameters that are assumed to converge to fixed or "true" values in conventional research. For instance, increasing the numbers of fractal samples causes fractal parameters to approach zero or positive infinity as their asymptotic limits, rather than an estimate of central tendency with dispersion of measurement errors. *Fractal dynamics* are used to describe bacterial growth in an inoculated petri dish, recapillarization after muscle trauma, or other biological space-filling potentials. The *fractal dimension* is a ratio of the number of branches produced by a living system compared with the resolution of measurement. For instance, how fast does respirable anthrax grow in lung tissue, and how much of the lung will be damaged irreparably in how much time? On a more constructive note, how much reperfusion of damaged heart muscle may occur if a heart attack victim is administered a certain drug within a certain time?

System bifurcation is graphic evidence that the behavior of a complex system is undergoing a major shift from extreme environmental pressures. *Dissipative structures* are distinctive physical changes observed in a system as it moves into a new equilibrium state after taking on too much information or energy for a system to maintain its current state. Boiling water is such an example, as liquid water becomes steam. Figure 2 is a graphical depiction of system bifurcation for the relation $x_{n+1} = rx_n(1 - x_n)$.

System state or type distinctions become crucial when attempting to predict system behavior. Clinicians may have to infer the kind of system they are dealing with based only on immediate observation. System behaviors determined only by previous states are *dynamical systems*. Dynamical systems output is not random, though it may appear to be. However, it is nearly impossible to identify all dynamical influences of chaotic system output. Five dynamical determinants constitute a natural upper limit for modeling chaotic systems. Some dynamical systems are more *stable* than others, that is, they generate self-similar outputs over

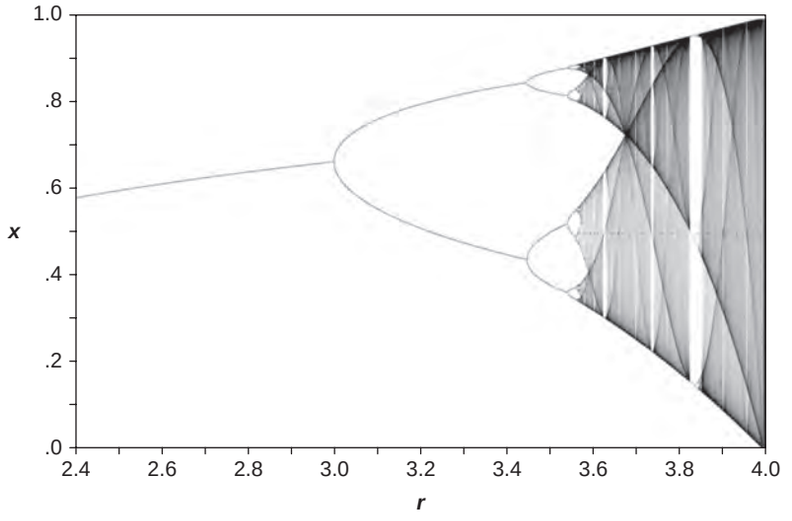


Figure 2 Graphic depiction of a system in bifurcation from environmental challenges

Note: The patterns of angled lines to the right represent dissipative structure as the system changes in outputs. A homogeneous pattern without lines would depict a system not undergoing bifurcation.

longer times. Dynamical stability is measured with low *Lyapunov* exponents, lambda (λ) in the dynamical function $f(x) = x^{\lambda}$. High Lyapunov exponents indicate less stable systems, and shorter time horizons for decision making. Variance in time horizons and individual patient outcomes are not captured in conventional reductionistic science.

Phase spaces and *attractors* are also complex system concepts. Phase space sets are plots of output variables as a function of either time of observation or its immediate prior state. Observing system output over a *phase space set* yields clues to the nature of the system processes. While random and chaotic data appear similar when plotted in a phase space set over time, a chaotic deterministic pattern may emerge when each output datum is plotted as a function of the previous one (Figure 3).

Random output homogeneously fills the phase space. Chaotic system output also may appear to fill phase space randomly. However, when each event is plotted as a function of the event immediately preceding it, the "noise" reveals a serial determinacy *if* the system is truly chaotic and not just random. Poincaré originally formalized the *strange attractor*, a type of dynamical system output. New information or vectors entering into an attractor system tend to settle into a small range of values, giving the appearance of attraction to a "center of gravity" (see Figure 4). If the center is described as a fraction and not an integer, the attractor is called *strange*. If the attractor parameter (called a *fractal dimension*) is not an integer, the attractor is called *strange*. Normally one, two, three, or more dimensions are conceivable for locating an observation in space and time; it is indeed *strange* to conceive of 1.2619, or some other noninteger number of dimensions.

Strange attractors were developed by Lorenz to describe the fluid dynamics in weather patterns when two air masses of different temperatures converge. Initially, there is turbulence, followed by a new equilibrium state. Small changes in initial conditions can make for large differences in the new state—a phenomenon called the *butterfly effect*. Biological applications of strange attractors include ecological breeding patterns, dynamics of neuron action potentials, and heart rate variability. Fractal dimensions can describe the extent of physiological damage from pulmonary hypertension.

Medical Decision Making

Complexity science may play an important role in medical decision making in the future by adding new descriptors, predictors, and parameters of complex system behavior. At the clinician level, patient care decisions based on clinical judgment,

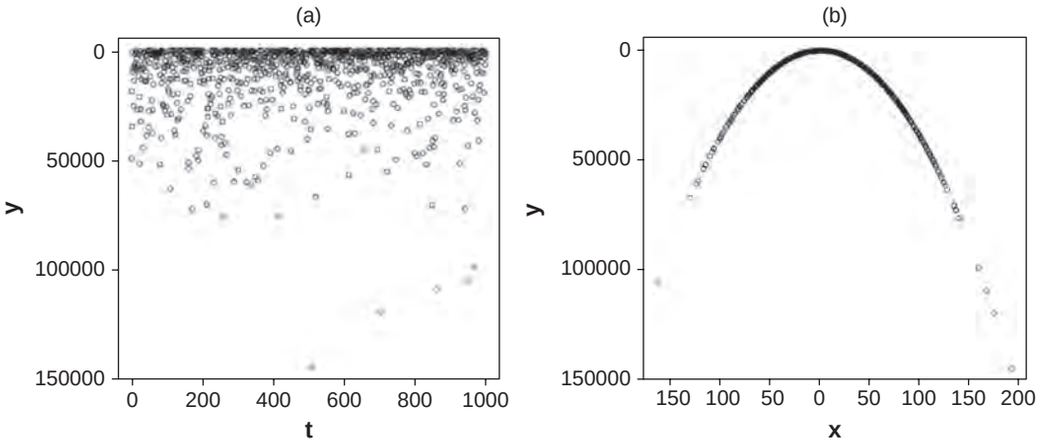


Figure 3 An example of chaos output plotted as (a) a function of time and (b) a function of preceding state

Note: The first plot (a) is the distribution of data as a function of time; the second plot (b) shows the same data plotted as a function of its immediately preceding value.

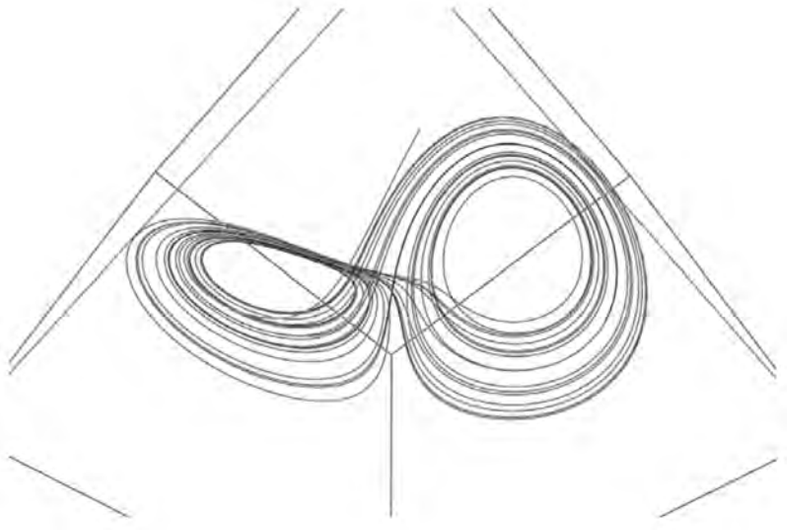


Figure 4 Phase space of the strange attractor

Note: A three-dimensional plot of data converging on a smaller area as they enter a three-dimensional system.

tacit knowledge, and current evidence may be improved by acknowledging limits of conventional clinical trial research but also by recognizing the remarkable successes of conventional research methods. If complexity science improves medical decision making, then system quality, cost, efficiency, efficacy, and safety should measurably improve in the future as more benefit from reduced uncertainty and fewer risks. Ideally, uncertainty would be reduced in the healthcare system and in clinical practice. Eventually, knowledge added by complexity science may reach some practical limits, to be replaced by other decision-informing tools.

At the policy level, complexity science perhaps also has potential to reduce system uncertainty and improve efficiencies. Even the simple exercise of thinking systemically can be helpful. The notion of time horizons is useful; so is appreciating the fact that systems adapt and that attempting system

change often precipitates reorganizing and unanticipated responses. Recognizing the signs of a distressed system expressed as "dissipative structures" may give health and financial decision makers more notice, more control, and the ability to effectively anticipate, adapt to, and manage accelerating change.

J. Michael Menke and Grant H. Skrepnek

See also Chaos Theory; Deterministic Analysis; Managing Variability and Uncertainty; Markov Models; Markov Models, Applications to Medical Decision Making; Markov Models, Cycles; Uncertainty in Medical Decisions

Further Readings

Boulding, K. E. (1956). General systems theory: The skeleton of science. *Management Science, 2,* 197–208.

Emanuel, K. (2008). Retrospective: Edward N. Lorenz (1917–2008). *Science, 320*(5879), 1025.

Liebovitch, L. S. (1998). *Fractals and chaos simplified for the life sciences*. New York: Oxford University Press.

McDaniel, R. R., Jr., & Driebe, D. J. (2001). Complexity science and health care management. In M. D. Fottler, G. T. Savage, & J. D. Blair (Eds.), *Advances in health care management* (pp. 11–36). Oxford, UK: Elsevier Science.

Plsek, P. (2001). Redesigning health care with insights from the science of complex adaptive systems. In *Crossing the quality chasm: A new health system for the 21st century* (pp. 309–317). Washington, DC: Institute of Medicine.

Rickles, D., Hawe, P., & Shiell, A. (2007). A simple guide to chaos and complexity. *Journal of Epidemiology and Community Health, 61*(11), 933–937.

Stacey, R. D., Griffin, D., & Shaw, P. (2000). *Complexity and management: Fad or radical challenge to systems thinking?* (Vol. 1). New York: Routledge.

Complications or Adverse Effects of Treatment

Complications, adverse effects, and adverse outcomes are the downside of disease and treatment. If they occur, they lower the quality of care as experienced from the patients' perspective while increasing cost. The result is worse care at a higher price, the opposite of what we all strive for; better care at a lower price. Thus, complications have lately attracted much attention, and much effort is spent trying to prevent them. By studying them and the mechanisms that underlie them, measures of prevention or reduction may be identified and implemented, thereby improving the quality and cost-effectiveness of care.

Definitions

There are various definitions available on the concepts of "complications" and "adverse effect" in the medical context. Although they may differ on details, common denominators in most definitions are the focus on four elements.

Harm

This element describes the fact that the patient experienced some event or condition that had a negative effect on the patient's health and maybe even resulted in (severe) harm to the patient or in death. It is not clear in all studies how unfavorable an event, outcome, or experience must be with respect to health-related harm to be considered a complication or an adverse effect. If a patient experiences some pain after an operation, does not sleep well, or must stay in the hospital one or two days longer than expected, or if there is a small hematoma, or some wound reddening that disappears in a few days without treatment, most people will agree that this is not really a complication but more an inherent and acceptable consequence of the intervention that was deemed necessary. On the other hand, if a wound abscess or the size of a hematoma and its pressure on the skin necessitate an operation, or if a patient is still in severe pain 6 weeks after discharge and still needs morphine for that, most people will agree that a complication has arisen.

Quality of Care

This element refers to the extent to which the care delivered caused, or contributed to, the harm that was experienced.

Unintentional Harm

This element refers to the fact that the harm was indeed unintentional and not an intentional sacrifice (as, for instance, is quite common in oncological

surgery to achieve a radical cancer resection), either as a calculated risk, deemed acceptable in light of an intervention's greater or more likely expected benefit, or as an unpleasant and unexpected surprise to both patient and doctor.

Harm Caused by Substandard Performance

This element addresses the question of the extent to which the harm was (in part) caused by professional substandard performance or even by an obvious error or mistake on the part of a person, group, organization, or other entity. If this is considered to be the case, it will easily lead to the additional questions of whether those substandard elements in the delivery of care *could* have been prevented or *should* have been prevented. From a legal perspective, a positive answer to both the "could" and "should" questions would suggest that some form of compensation might be justified, by a liability procedure or otherwise, depending on the extent of the harm and the extent of the causality.

The worldwide interest in these concepts and in patient safety in general was strongly increased by the Harvard Medical Practice Study (HMPS) and by the report *To Err Is Human* that followed it. In the HMPS, an adverse event was defined as

> an injury that was caused by medical management (rather than the underlying disease) and that prolonged the hospitalization, produced a disability at the time of discharge, or both. (Brennan et al., 2004, p. 145)

The fact that this definition takes element 4, the causality criterion, on board opens it up to the accusation of subjectivity. For many outcomes, the cause may not be entirely clear, or it may be attributed to several risk factors that may even reinforce one another in their causality. The subjective and oversimplified "yes/no somebody's fault" assumption does not often lead to an appropriate representation of causality. The HMPS defined the quality element (negligence) as "care that fell below the standard expected of physicians in their community."

In the current era in which guidelines abound, the "expected standard" is generally reasonably clear. However, at the time of the HMPS, even this may have been less unambiguous.

Complication Registry: An Example

In the Netherlands, around the turn of the century, a nationwide initiative to standardize the prospective registration and analysis of complications took a different approach. Here the issue of causality was intentionally left out of the definition, assuming that at the time a complication or adverse effect is noticed or registered, there will often be insufficient insight into the causality or preventability of the harm inflicted. The Dutch definition of *complication* does include an unambiguous harm threshold, thereby providing a clear-cut criterion for when something is serious enough to be considered a complication. It states,

> An unintended and unwanted event or state occurring during or following medical care, that is so harmful to a patient's health that (adjustment of) treatment is required or that permanent damage results. The adverse outcome may be noted during treatment or in a predefined period after discharge or transfer to another department. The intended results of treatment, the likelihood of the adverse outcome occurring, and the presence or absence of a medical error causing it, are irrelevant in identifying an adverse outcome. (Marang-van de Mheen, van Hanegem, & Kievit, 2005, p. 378)

Thus the assessment of causality is postponed to a later date, when more information is available. This has the advantage of providing the opportunity to standardize or improve the way causality is analyzed, thus providing a clearer answer on epidemiological questions about attributable risk. As a consequence, both judgment subjectivity and interobserver variation should be lower.

As the Dutch complication registry is not confined by a limited set of specified "complications-to-register" at the exclusion of all others, the result is that an essentially unlimited number of different complications could meet the definition. Registering them inevitably may require that free text be used. For analysis purposes, however, free text is useless and must be recoded into a meaningful set of dimensions. Within the Dutch system, a so-called Master Classification has been created that characterizes complications on three main dimensions, and in addition on a severity scale.

The first dimension defines the nature of the complication, answering the "What?" question. Its subcategories have been adapted from the ICD9/10, to better fit the adverse outcome registration purpose, and include types such as bleeding, dysfunction, and infection/inflammation.

The second dimension answers the "Where" question and specifies location, both by organ systems and organs, and by body topography (chest, abdomen, etc.).

In the third dimension, contextual information and (potential) determinants are recorded, while the fourth dimension specifies the harm inflicted on a patient (varying, for the surgical specialties, from full recovery without reoperation, to requiring reoperation, leaving permanent damage, and resulting in death).

The main purpose of the Master Classification is not so much the recoding itself as the facilitation of later data analysis. In combination with a minimum data set, in which elementary patient and context characteristics are recorded, the three-dimensional Master Classification provides maximum analytic flexibility. Sampling and analyzing complications can vary from broad categories such as "all bleeding complications" to very specific subsets such as "all infections in hip replacements leading to death in male patients over 70."

The Dutch approach may not be unique. What is good about it is that the definition and the coding system used, in combination with a minimum data set and the database structure, make it possible to address a wide range of questions without having to return to patient records. That is a crucial characteristic of any online system that aims at monitoring (and improving) the safety of healthcare by analyzing and reducing complications and adverse outcomes.

Causality and Complication Rates

Whether the issue of causality was or was not tackled adequately by the HMPS is still, after all these years, a matter of debate. The treatment of severe, sometimes life-threatening disease may require the weighing of potential benefits and potential harms. Choices will have to be made, and it is not unusual that the higher the goal is, the graver the risks are.

However, unintentional harm does occur to patients on quite a large scale. That in itself is sufficient reason to strive for a reduction of its occurrence or impact. To what extent this harm results from some event totally beyond anyone's control, is the consequence of a calculated risk, or is the consequence of below-standard care may not always be immediately clear, even to insiders.

In the field of patient safety, of complications and adverse effects, simple notions of one-cause-with-one-effect rarely hold. Instead, multiple causes, some within the grasp of doctor or patient, and some totally beyond their control, may combine or even reinforce one another and bring about a single but multicausal complication. Likewise, a single underlying cause may contribute to more than one complication, some less severe and some more so (see Figure 1).

The epidemiologically correct way to deal with the causality of complications would be to calculate the relative risk or odds ratio per determinant, for instance using logistic regression. This, however, does not immediately solve the problem as it requires adequate identification of all relevant covariates. The problem is thus transformed into adequately identifying all potential causal elements, obtaining relevant data of sufficient quality, and correctly analyzing those data using state-of-the-art statistical methods, such as logistic regression or multilevel analysis. Subsequently, relative risks can be used to calculate the attributable risk of one or more particular determinants. An important advantage of this statistically more refined approach, over an assumed simple one-to-one cause-effect relationship, is that such an attributable risk will provide a realistic notion of the health gains that can be expected when this shortcoming is eliminated or reduced, where an inappropriately simple causal relationship will overestimate the health gains of interventions targeting the assumed causes.

Given the definition and other issues, it is not surprising that for many comparable treatments or procedures, there are large differences in published complication rates. The fact that many studies neither provide specific information on the definition used nor have common standards in methodology may explain part of the variation in reported complication rates. In particular, how the definition deals with the issue of causality has been found to

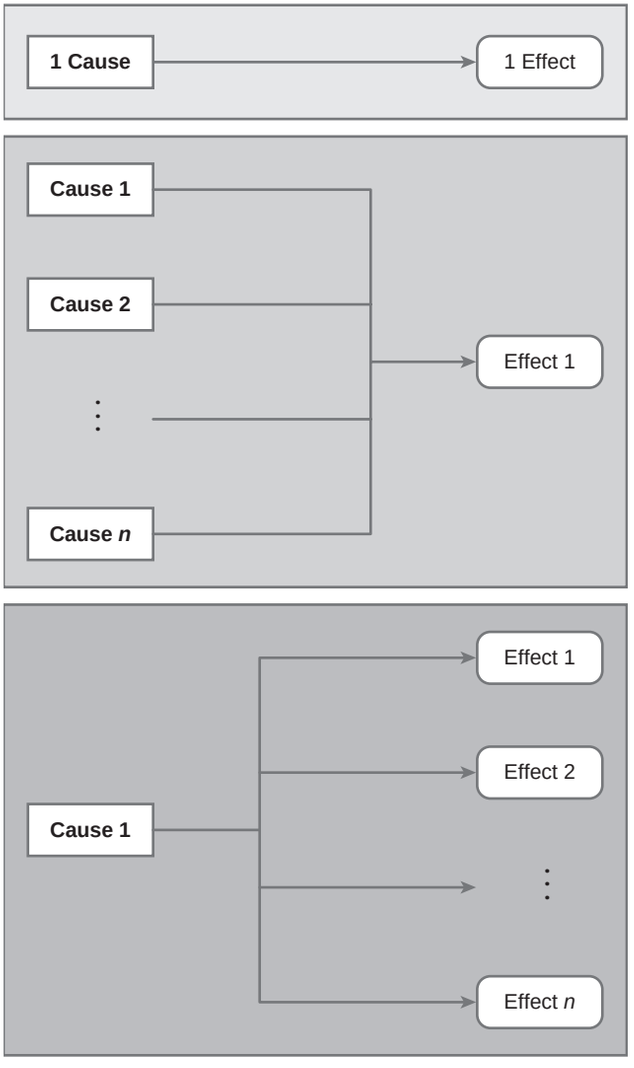


Figure 1 Causal relationships

Note: With respect to complications and adverse effects, one-to-one causality rarely holds. Instead, multiple causes may work together to bring about a single harm, or one cause may lead to several types of harm.

be an important determinant in the differences between complication rates.

Other methodological issues that are related to the reported incidence of complications or adverse outcomes are the type of data collection (prospective or retrospective) and the number of reviewers used. Most important are patient characteristics, which, apart from age, sex, and disease diagnosis, include more subtle determinants such as disease spectrum, comorbidity, the context in which patients are seen, and the type of admissions.

Relevance to Healthcare Providers, Patients, and Medical Decision Making

Complications and adverse effects are, because they compromise quality and increase cost, relevant to many interested parties, not in the least to patients and healthcare providers. For healthcare providers, they are relevant because adequate decision making in medicine requires a weighing of potential harms and benefits. Doctors, when they must make choices on risky interventions for severe diseases, must have a keen insight into the risks of any treatment they consider and into the determinants that define this risk, which may differ between different subgroups of patients.

For patients, information about complications and adverse effects is even more relevant because it is they who bear the consequences. Therefore, patients have a right to know what their risks are and what the consequences are if such a risk materializes and the harm really occurs. It goes without saying that such risk information should not be limited to the risk of mortality but should include morbidity, both less and more severe. Only on the basis of appropriate information can expected benefits and expected harms be adequately weighed.

For the field of medical decision making, complications and adverse outcomes are relevant in more than one way. First, there is the classical threshold approach to medical decision making, which holds that costs (C) and benefits (B) of treatment are weighed in light of an uncertain disease diagnosis. Treatment is the preferred policy if the chance of the disease being present exceeds the treatment threshold, defined by

$$\frac{C}{C+B}.$$

Thus, the higher the cost is (i.e., the risk of adverse effects), the more certain one should be that the disease is indeed present for the benefits to outweigh the harm. Thus, the treatment threshold will be closer to 1. Likewise, the more effective a treatment is, the lower the diagnostic certainty and thus the treatment threshold will be. Or, under the same treatment threshold, the higher the potential risk of harm may be that is deemed acceptable.

Second, modern medical decision making lays great emphasis on shared decision making.

Essential in shared decision making is that patient and caregiver communicate openly about risks and benefits of various choices and decide on the way to go taking into account not only objective evidence, but in addition priorities and preferences of the patient.

Third, a lot of research in the field of medical decision making is on risk perception and on the way that patients, doctors, and others transform an objective chance into a subjective notion of risk. Such research will improve insight into how patients (and doctors) perceive the risks of adverse effects, and weigh them into a final healthcare choice.

J. Kievit and P. J. Marang-van de Mheen

See also Causal Inference in Medical Decision Making; Complexity; Risk Communication

Further Readings

Brennan, T. A., Leape, L. L., Laird, N. M., Hebert, L., Localio, A. R., Lawthers, A. G., et al. (2004). Incidence of adverse events and negligence in hospitalized patients: Results of the Harvard Medical Practice Study I. 1991. *Quality and Safety in Health Care, 13*(2), 145–152.

Edwards, A., & Elwyn, G. (1999). How should effectiveness of risk communication to aid patients' decisions be judged? A review of the literature. *Medical Decision Making, 19*(4), 428–434.

Hayward, R. A., & Hofer, T. P. (2001). Estimating hospital deaths due to medical errors: Preventability is in the eye of the reviewer. *Journal of the American Medical Association, 286*(4), 415–420.

Marang-van de Mheen, P. J., Hollander, E. J., & Kievit, J. (2007). Effects of study methodology on adverse outcome occurrence and mortality. *International Journal for Quality in Health Care, 19*(6), 399–406.

Marang-van de Mheen, P. J., van Hanegem, N., & Kievit, J. (2005). Effectiveness of routine reporting to identify minor and serious adverse outcomes in surgical patients. *Quality and Safety in Health Care, 14*(5), 378–382.

Whitney, S. N., Holmes-Rovner, M., Brody, H., Schneider, C., McCullough, L. B., Volk, R. J., et al. (2008). Beyond shared decision making: An expanded typology of medical decisions. *Medical Decision Making, 28*(5), 699–705.

Computational Limitations

There are two aspects to computational limitations in decision making. On the one hand, there is the idea that the human brain is computational and that optimal decisions require lengthy computations but that the human computational capacity is limited, and therefore human decision performance is less than optimal and humans must use alternative strategies (heuristics, etc.) to make decisions.

The second aspect is that computers are limited as well from recommending optimal decisions because the algorithms required, by necessity, take too much time. So computers too must use alternative approaches.

The primary dialectic in decision making pits the *rational-man model,* where decisions are made in accordance with the goal of maximizing utility, against the *natural-man model,* where decisions are made in a way that has been evolutionarily designed to best fit our environment. One engine of this dialectic is the issue of how much computational power is available to the decision maker. Computer scientists have attempted to model both sides in their machines, with results that have important implications for decision makers and those trying to help them. On the other side, psychologists have tried to apply computational models to observed and experimentally induced behavior.

As steam engines provided the motivating analogy in 19th-century science beyond mechanical engineering, computation provides the current leading analogy for many fields, including theories of the mind. Colloquially and even scientifically, authors discuss the brain as if it were a von Neumann computer: We separate thinking memory from storage memory, and we ask what operations our thinking self can perform with how many memory "cells." Research results need to be clear whether they mean computation in its strict sense or in its analogous sense. This entry first addresses machine-based computational limitations and then explores these difficulties in human cognition.

Machines

The field of artificial intelligence is the primary field where computational models of decision

making get computer scientists' full attention and where their models get tested. Traditional areas of artificial intelligence—game playing, visual understanding, natural-language processing, expert advice giving, and robotics—each requires processing data taken from the environment to result in a conclusion or action that embodies knowledge and understanding. The nature of "computation" differs in each case. In speech recognition, the current leading methods involve numerical calculation of probabilities. In game-playing, the methods call for deriving and considering many alternative game configurations. For expert systems—beginning with the program MYCIN, whose goal was supporting the management of fever in a hospitalized patient—the computer explores pathways through rules. Thus, the computations involve a mix of quantitative and "symbolic" processing.

In computer science, *computational limitations* refer to two primary resources: time and space. *Space* refers to how much computer memory (generally onboard RAM) is needed to solve a problem. *Time* refers not only to the amount of time needed to solve a problem, in seconds or minutes, but also to the algorithmic complexity of problems. Problems whose solution time doubles if the amount of data input into the solver doubles have linear complexity; problems whose solution quadruples have quadratic complexity. For example, inverting a matrix that may represent transition probabilities in a Markov model of chronic disease has cubic complexity, and sorting a list has between linear and quadratic complexity. These algorithms are said to be *polynomial* (*P*) in the size of their data inputs. On the other hand, problems whose solution time doubles even if only one more piece of information is added have *exponential complexity*, and their solution takes the longest amount of time (in the worst case). For instance, enumerating by brute force all possible potential strategies for treatment in a specific clinical problem, where order matters, such as the question of which tests should be done in which order (diagnosing of immunological disease being a classic case, with the multitude of tests available), leads to an exponential number of pathways. If a single new test, for instance, becomes available, then every single pathway would have to consider using that test or not, thereby doubling the number of possibilities to be considered. If these strategies are represented as decision trees, the number of terminal nodes would double.

There is a complexity class between polynomial and exponential called *nonpolynomial* (NP). Most of the interesting problems in decision making have been shown to be in this class, or *NP complete*. For instance, the problem of diagnosis is NP complete, meaning that a general solution could potentially take an unlimited amount of time (for all intents and purposes) of coming up with the best list of diagnoses for a particular patient. A central mystery of computer science is whether an algorithm can be found that would make NP complete algorithms polynomial: Is $P = NP$? If yes, then, with the right programming, the process of diagnosis would *not* take an "unlimited amount of time." However, most computer scientists believe this equation *not* to be the case, that is that $P \neq NP$, and that these time-saving algorithms do not exist. Their belief stems from the fact that it has been shown that, if one NP complete problem can be shown to be solved in polynomial time, then all other NP complete problems can be solved in polynomial time as well. However, so many problems are NP complete and so many people have been looking for solutions for 30 years that it appears unlikely that a solution will be found.

If it is true that $P \neq NP$, then the most important problems for which we want computers to supplement human thought, processing, and decision making will not be able to provide the correct answers in the time in which we need them to do so.

The result in computer science has been the reliance on heuristics that, when used, give good-enough results. The support for this use of reliance was provided by Herbert Simon, who called this primary heuristic *satisficing*.

Heuristic methods were the hallmark of early expert systems that provided decision advice. These were mostly rule-based systems with basically ad hoc methods of adjudicating conflicting rules, such as certainty factors or other measures. Metarules, in particular, were carefully crafted to look "reasonable." Thus, in the case of conflicting rules, a heuristic to be used might be the more "specific" rule (i.e., one where there were more "left-hand side" [antecedent] conditions that were met) over the less specific rule. For instance, a rule that pertained to a specific white blood cell count would be chosen over a rule that simply cited "WBC >

15,000." If costs were represented, the metarule would counsel using the less costly rule (e.g., get a complete blood count rather than biopsy).

Other heuristic systems included *blackboard* systems, where rules or (later) *agents,* with both data-gathering and action-taking capabilities, shared a common data space. The agents reacted relatively autonomously to data available on the common blackboard. Coordination among the agents relied, again, on metarules: More specific agents "won" over less specific ones.

In logic-based systems, the inference systems were based on the soundness and consistency of logical derivation: *modus ponens* (if all men are mortal and Socrates is a man, then Socrates is mortal) and *modus tollens* (if all men are mortal and Thor is not mortal, then Thor is not a man) in predicate logic, or binding and resolution in logic programming, such as is used in Prolog and XSB. However, to deal with uncertainty and apparently conflicting rules, *modal* logics were created that could reason about rules, much as metarules in expert systems were needed to adjudicate among conflicting rules. Some modal logics, for instance, made the implicit assumptions that, unless exceptions were explicitly stated, no exceptions were assumed to be present. While this assumption is reasonable, it may fail in environments where the knowledge base is assembled by experts who do not realize that the system has no idea of what an exception might be or who may be inconsistent in pointing out what exceptions indeed arise.

These disparate efforts converge on a common conclusion: If rational (computer-based) systems want to act rationally in the world, then their metacognition cannot follow straightforward utility-maximization procedures. Satisficing and heuristics will play major roles.

Humans

Researchers in cognitive science address decision making as computational in a number of ways. At the conscious level, they point to the language of decision making: We "weigh" evidence and "balance" pros and cons. At the preconscious level, we make choices based on some sort of psychological version of conscious weighing and balancing—but with the limitations imposed by time and by cognitive boundaries.

A classic limit due to "space" is embodied in the truisms of the *magic number 7*: that our short-term memory (note the computer-like label) can accommodate 7 ± 2 chunks. Since 1956, this limit has become part of social lore. We can remember phone numbers, along with satellite access and secret codes (a potential total of 21 digits), by chunking them into three entities and then recalling each number as 3 chunks (area code, "exchange," number) or 7 digits of access (usually divided into a unit of 3 digits and then 4 remaining digits). Chess masters apparently remember board arrangements because they chunk patterns of many pieces into one pattern, much as expert diagnosticians recall many details about a patient because many findings may be chunked into syndromes that explain the findings or make the findings memorable specifically because they are exceptions to the syndromic rule. For instance, a patient with no crackles in the lung fields but with fever, cough, diminished air entry, infiltrate on an X-ray, and sputum culture positive for pneumococci has pneumococcal pneumonia, notable for the absence of crackles, much like Sherlock Holmes's dog was notable for not barking.

Cognitive scientists have gone further, to opine that the computational limits forced evolution to mold in humans heuristics that are successful precisely because of these limits. Gerd Gigerenzer and Reinhard Selten title their compendium on the subject *Bounded Rationality* as a direct response to Simon and match specific information environments to specific heuristics. Thus, in a noisy but stable information environment, people use the Imitate Others heuristic. Where rational-man theorists see deficiencies in people's abilities to act totally rationally, as defined by the rules of maximizing expected utility, these experimentalists see strength and power in people's abilities to do as well as they can in the limits nature set them.

Thus, the psychologists see people using heuristics in much the way that computer scientists learned to rely on them: for metacognition. In Gary Klein's famous example, firemen, when faced with a new and clearly dangerous situation, rather than calculate all the possibilities and choose the optimal path, use the heuristic of moving to the last safe place ("Take the Last") and consider the options from there. These major heuristics, according to Girgerenzer, fall under the

general class of Fast and Frugal—to arrive at a conclusion quickly and with the use of minimal resources is itself a goal.

Other cognitive researchers have gone deeper, delving into the structure of memory that undergirds much decision making. Fuzzy Trace theory points out that memory, beyond short-term and long-term, contains two further types: gist and verbatim. The fuzzy-processing preference is to operate at the least precise level of representation that can be used to accomplish a judgment or decision. This preference is clearly related to the "frugal" aspect of Girgerenzer's conception.

Behavioral-economist researchers, such as Amos Tversky and economics Nobel prize winner Daniel Kahnemann, discovered many biases, discussed elsewhere in this encyclopedia, by comparing human behavior with behavior that maximizing expected utility would dictate. In this sense, the heuristics are *biases* to be corrected. From the cognitive psychologist's perspective, each "bias" reflects a mental mechanism built on a particular strength of the human brain. Thus, people's powerful abilities in pattern matching become the *representativeness* bias; efficient memory retrieval becomes the *availability* bias; the abilities to discern salience and signals become the *anchoring and adjusting biases*.

Synthesis

The demand for decision support delivered by computers in practice forces developers to confront this dialectic. On the one hand, the computer is expected to be correct, evidence-based, and rational. On the other hand, it participates in a real work environment with all the limitations of its users and the time pressures of their jobs. An ideal synthesis would have the knowledge infrastructure of the decision support based on the rational-man model but with a user interface built on principles of bounded rationality and heuristic actions. Current research and practice work toward these ideals on several fronts, although no solution is currently offered. The rational-man-based decision support is reserved for policy recommendations, while frontline decision making depends on heuristics-based decision support that generally does not take human cognitive thinking or relations with the human-computer interface into account. Hopefully, we shall see proper syntheses in the future.

Harold Lehmann

See also Bounded Rationality and Emotions; Clinical Algorithms and Practice Guidelines; Cognitive Psychology and Processes; Computer-Assisted Decision Making; Expected Utility Theory; Fuzzy-Trace Theory; Heuristics

Further Readings

Cooper, G. F. (1990). The computational complexity of probabilistic inference using Bayesian belief networks. *Artificial Intelligence, 42,* 393–405.

Garey, M. R., & Johnson, D. S. (1979). *Computers and intractability: A guide to the theory of NP-completeness*. New York: Freeman.

Gigerenzer, G., & Selten, R. (Eds.). (2001). *Bounded rationality: The adaptive toolbox*. Cambridge: MIT Press.

Klein, G. (1998). *Sources of power: How people make decisions*. Cambridge: MIT Press.

Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review, 63,* 81–97.

Reyna, V. F., & Adam, M. B. (2003). Fuzzy-trace theory, risk communication, and product labeling in sexually transmitted diseases. *Risk Analysis, 23*(2), 325–342.

Simon, H. A. (1985). *The sciences of the artificial* (2nd ed.). Cambridge: MIT Press.

Computer-Assisted Decision Making

Computer-based decision support software can assist in arriving at decisions regarding diagnoses and diagnostic workup, therapy choices, and prognoses. Generally, such software systems function by interpreting data about patients using biomedical knowledge that has been encoded into the software. The results of these interpretations are often decision alternatives that are pertinent to the patient under consideration and are presented to the users of the software to assist them in their decision making. The users of the software may be clinicians or patients.

Approaches

A decision support software system has several conceptual components:

1. An inferencing approach typically embodied in an algorithm that enables the system to interpret patient data based on the knowledge available to the system. Examples of such approaches include Bayesian inferencing and production rule evaluation.
2. A knowledge base that comprises the biomedical knowledge available to the system. The knowledge is encoded in a form that corresponds to the inferencing approach being used. For example, a knowledge base for trauma diagnosis might consist of a Bayesian network relating patient symptoms and findings to internal organ injuries.
3. Optional interfaces to other computer systems to obtain data about a patient.
4. A user interface to interact with the user, such as to obtain data, and to present the results from the inferencing.

Based on the mode in which the decision support is invoked, the system may be characterized as one providing solicited advice or one providing unsolicited advice. In the former case, a clinician may seek recommendations from a decision support system to assist with making a differential diagnosis in a patient with an unusual presentation. Such systems usually contain a large knowledge base that spans a domain of clinical interest such as internal medicine or infectious diseases. Unsolicited advice is rendered by systems (a) in response to clinical events that are being monitored, such as the reporting of a critically low value for a serum potassium test, or (b) as a critique of a proposed physician intervention such as prescribing a medication to which the patient is hypersensitive. Decision support systems that offer unsolicited advice, to be able to function, must be integrated with sources of patient data such as an electronic medical record (EMR) system or a computer-based provider order entry (CPOE) system. Systems that offer solicited advice may be integrated with sources of patient data or may be freestanding.

An important aspect of decision support systems for clinical use is how it integrates into the clinical workflow. In other words, the successful use of these systems depends on when and where the system's advice is presented to the clinicians. Thus, various kinds of tools have been created to present advice at particular points in the clinical workflow. For example, reminder systems are used often to advise clinicians in the ambulatory setting about preventive care actions that are applicable to a patient. Electrocardiography (ECG) machines incorporate features to analyze the ECG and print or display the resulting interpretation of the findings with the ECG trace. Rule-based systems critique physician orders and prescriptions in the CPOE application to prevent orders that might have the potential to harm the patient or those that might be ineffective. Such systems also might suggest additional orders called corollary orders: For example, an order for a nephrotoxic medication might lead to a corollary order for performing kidney function tests. Abnormal laboratory test results are highlighted on the screen to draw the attention of the clinician to those values. Furthermore, links to didactic informational resources, also known as infoButtons, can be shown next to the results. These information resources can be used by the clinicians to help interpret the test result and decide on an appropriate action. Treatment planning systems for surgery or radiation therapy are used in a laboratory setting initially to plan the treatment. The outputs of these systems are presented to the clinician during the treatment in the operating room.

Applications and Examples

One of the well-known, early examples of software for computer-assisted medical decision making is MYCIN, developed by Edward Shortliffe at Stanford University. MYCIN provided, on solicitation by physicians, antimicrobial therapy recommendations for patients with bacterial infections. MYCIN was capable of explaining how it arrived at its recommendations.

Internist-1, another early system, assists in diagnostic decision making. It is capable of making multiple diagnoses from patient symptoms and findings in the domain of internal medicine using a very large knowledge base. Internist-1 is available commercially as the Quick Medical Reference

(QMR) system. Over the years, many other computer-based decision support systems have been created to assist with making diagnoses. Among these are the DXplain system for diagnoses in internal medicine, and a Bayesian network-based system for diagnosing the cause of abdominal pain, created by F. T. de Dombal and colleagues.

Computer-assisted decision aids have been used for complex tasks such as radiation therapy treatment planning and surgical planning. In the former case, computer-based tools are used for designing a treatment plan that optimizes the delivery of radiation to a tumor. In the latter case, computer software is used with three-dimensional images of the patient to plan and simulate the surgical procedure.

A separate class of decision-making systems has been investigated to support the need for planning, coordinating, and executing care over extended time periods. The intended use of such systems is often to implement decision support based on clinical practice guidelines. The systems support decision making around the diagnosis, evaluation, and long-term management of a patient. Examples of these systems include the Guideline Interchange Format, ProForma, EON, and Asbru.

Effectiveness and Usage

In clinical studies, clinical decision support (CDS) systems have been shown largely to affect the performance of practitioners in a desirable manner. For example, reminder systems have increased the frequency with which preventive care actions are carried out; diagnostic decision support systems have been shown to help make the correct diagnosis; and CDS embedded in CPOE systems has reduced the ordering of drugs that might cause harm to the patient. Systems that provide unsolicited advice are more likely to affect practitioner performance than are systems that require the practitioner to seek advice. Few studies on computer-assisted decision-making systems have measured the impact of such systems on patient outcomes. Among these studies, relatively few have demonstrated an improvement in patient outcome.

In spite of the beneficial impact of computer-assisted decision-making tools on practitioner performance, these tools are not being used widely yet. One of the challenges in the adoption of CDS systems is the lack of specificity of the decision support, especially for the systems that offer unsolicited advice. These systems must have access to codified patient data. If such data are lacking or are imprecise for a patient, advice is delivered to the practitioner that may not apply to that patient. For example, if there is a record of an allergy to a particular medication, but the severity is not documented for a patient, a decision support system might advise the physician to not order the medication, even though the sensitivity is very mild in this patient and the clinical benefit potentially is large. Another major barrier to the widespread usage of CDS systems is the availability of knowledge bases to cover the different domains of healthcare and of clinical practice. The creation and maintenance of knowledge bases requires much effort from subject matter experts and knowledge engineers. Furthermore, such knowledge bases must be usable in a variety of different host CDS systems and many different practice environments. Financial incentives can also help increase the adoption of computer-assisted decision-making tools. The increasing use of pay-for-performance measures, where providers are reimbursed by payers based on their performance in a range of quality measures, might lead to increases in adoption of tools for decision making.

The use of standards for representing the knowledge and providing patient data to the CDS system will reduce technical barriers for implementing and using CDS systems. The Clinical Decision Support Technical Committee at Health Level Seven (HL7), an organization with international participation that creates standards for healthcare data interchange, is leading the effort for developing knowledge representation standards. HL7 sponsors the Arden Syntax standard for representing rules that are used in a number of commercially available clinical information systems.

Aziz A. Boxwala

See also Bayesian Networks; Clinical Algorithms and Practice Guidelines; Decision Rules; Expert Systems

Further Readings

Boxwala, A. A., Peleg, M., Tu, S., Ogunyemi, O., Zeng, Q. T., Wang, D., et al. (2004). GLIF3: A representation

format for sharable computer-interpretable clinical practice guidelines. *Journal of Biomedical Informatics, 37*(3), 147–161.

de Clercq, P. A., Blom, J. A., Korsten, H. H., & Hasman, A. (2004). Approaches for creating computer-interpretable guidelines that facilitate decision support. *Artificial Intelligence in Medicine, 31*(1), 1–27.

Garg, A. X., Adhikari, N. K., McDonald, H., Rosas-Arellano, M. P., Devereaux, P. J., Beyene, J., et al. (2005). Effects of computerized clinical decision support systems on practitioner performance and patient outcomes: A systematic review. *Journal of the American Medical Association, 293*(10), 1223–1238.

Greenes, R. A. (Ed.). (2007). *Clinical decision support: The road ahead*. New York: Academic Press.

Hripcsak, G. (1994). Writing Arden Syntax medical logic modules. *Computers in Biology and Medicine, 24*(5), 331–363.

Kawamoto, K., Houlihan, C. A., Balas, E. A., & Lobach, D. F. (2005). Improving clinical practice using clinical decision support systems: A systematic review of trials to identify features critical to success. *British Medical Journal, 330*(7494), 765.

Kuperman, G. J., Teich, J. M., Gandhi, T. K., & Bates, D. W. (2001). Patient safety and computerized medication ordering at Brigham and Women's Hospital. *Joint Commission Journal on Quality Improvement, 27*(10), 509–521.

Maviglia, S. M., Yoon, C. S., Bates, D. W., & Kuperman, G. (2006). KnowledgeLink: Impact of context-sensitive information retrieval on clinicians' information needs. *Journal of the American Medical Informatics Association, 13*(1), 67–73.

Osheroff, J. A., Teich, J. M., Middleton, B. F., Steen, E. B., Wright A., & Detmer, D. E. (2006). *A roadmap for national action on clinical decision support*. Washington, DC: American Medical Informatics Association.

Purdy, J. A. (2007). From new frontiers to new standards of practice: Advances in radiotherapy planning and delivery. *Frontiers of Radiation Therapy and Oncology, 40,* 18–39.

Conditional Independence

The concept of conditional independence plays an important role in medical decision making. Conditional independence itself concerns whether information about one variable provides incremental information about another variable. The concept is also important in articulating assumptions needed to reason about causality.

Independence and Conditional Independence

The concepts of independence and conditional independence concern whether information about one variable also contains information about another variable. Two variables are said to be independent if information about one gives no information about the other. For example, one might expect that whether an individual is left-handed or right-handed gives no information about the likelihood of developing pneumonia; it would then be said that being left-handed or right-handed is independent of the development of pneumonia. More formally, if $P(Y = y|Z = z)$ is the probability that $Y = y$ given $Z = z$ and if $P(Y = y)$ is the overall probability that $Y = y$, then the variables Y and Z are said to be independent if $P(Y = y|Z = z) = P(Y = y)$; in other words, Y and Z are independent if the information that $Z = z$ gives no information about the distribution of Y; equivalently, Y and Z are independent if $P(Z = z|Y = y) = P(Z = z)$. When two variables X and Y are not independent, they are said to be correlated or to be statistically associated. Independence is also often referred to as "marginal independence" or "unconditional independence" to distinguish it from conditional independence.

The concept of conditional independence is a natural extension of the concept of independence. Conditional independence is similar to independence, except that it involves conditioning on a third variable (or set of variables). Thus suppose that one is interested in the relationship between X and Y within the strata of some third variable C. The two variables, X and Y, are said to be conditionally independent given C if information about X gives no information about Y once one knows the value of C. For example, a positive clinical breast exam is predictive of the presence of breast cancer; that is to say, a positive clinical breast exam and the presence of breast cancer are not independent; they are statistically associated. Suppose, however, that in addition to the results of a clinical breast exam, information is also available on further evaluation procedures such as mammogram and biopsy results. In this case, once

one has information on these further evaluation procedures, the results from the clinical breast exam give no additional information about the likelihood of breast cancer beyond the mammogram and biopsy results; that is to say the presence of breast cancer is conditionally independent of the clinical breast exam results given the results from the mammogram and biopsy. More formally, if $P(Y = y|Z = z, C = c)$ is the probability that $Y = y$ given that $Z = z$ and $C = c$ and if $P(Y = y|C = c)$ is the probability that $Y = y$ given that $C = c$, then the variables Y and Z are said to be conditionally independent given C if $P(Y = y|Z = z, C = c) = P(Y = y|C = c)$. When two variables Y and Z are not conditionally independent given C, then they are said to be associated conditionally on C or to be conditionally associated given C. The notation $Y \amalg Z|C$ is sometimes used to denote that Y and Z are conditionally independent given C; the notation $Y \amalg Z$ is used to denote that Y and Z are unconditionally independent. A. P. Dawid's article "Conditional Independence in Statistical Theory" gives an overview of some of the technical statistical properties concerning conditional independence. The focus here will be the relevance of the idea of conditional independence in medical decision making.

Conditional Independence in Causal Reasoning

For medical decision making, the idea of conditional independence is perhaps most important because of its relation to confounding and the estimation of causal effects. Suppose that a researcher is trying to compare two drugs, Drug A and Drug B, in their effects on depression. Suppose that it has been demonstrated in randomized trials that both drugs result in higher recovery rates than a placebo but that it is unclear whether the recovery rate for Drug A or for Drug B is higher. Suppose that observational data are available to compare Drugs A and B but that no randomized trial has been conducted to make such a comparison. Let X_i be the variable that indicates which treatment individual i in fact received, so that $X_i = 1$ denotes individual i's receiving Drug A and $X_i = 0$ denotes individual i's receiving Drug B. Let Y_i denote whether or not individual i is clinically depressed 1 year after the initiation of drug therapy. For each individual, it might be of interest whether the individual's depression status would be different under Drug A compared with Drug B. Let $Y_i(1)$ denote individual i's depression status 1 year after the initiation of drug therapy had the individual, possibly contrary to fact, been given Drug A. Let $Y_i(0)$ denote individual i's depression status had the individual, possibly contrary to fact, been given Drug B. The variables $Y_i(1)$ and $Y_i(0)$ are sometimes referred to as counterfactual outcomes or potential outcomes. For any given individual, one only gets to observe one of $Y_i(1)$ or $Y_i(0)$. For individuals who in fact received Drug A, one observes $Y_i(1)$; for individuals who in fact received Drug B, one observes $Y_i(0)$. Because only one of the potential outcomes is observed, it is not possible to calculate the causal effect, $Y_i(1) - Y_i(0)$, for individual i since one of $Y_i(1)$ or $Y_i(0)$ is always unknown.

Although it is not possible to estimate individual causal effects, one can in some contexts, under assumptions articulated below about conditional independence, estimate average causal effects for a particular study population. In what follows, the index i is generally suppressed, and the variables are treated as random, assuming that the subjects in the study are randomly sampled from some study population. One might thus be interested in comparing the average depression rate for the population if the whole study population had been given Drug A, denoted by $E[Y(1)]$, with the average depression rate for the population if the whole study population had been given Drug B, denoted by $E[Y(0)]$. Although it is not possible to observe $Y(1)$ or $Y(0)$ for each individual, one might consider comparing the observed depression rates for the group that in fact received Drug A, denoted by $E[Y|X = 1]$, and the observed depression rates for the group that in fact received Drug B, denoted by $E[Y|X = 0]$. The problem with such an approach is that the group that received Drug A and the group that received Drug B might not be comparable. For example, the group that received Drug A might have had more severe depression or might have consisted of older subjects or might have had worse diets. To attempt to make the groups comparable, control may be made for as many confounding variables as possible, variables that affect both the treatment and the outcome, denoted by C. It is then hoped that within strata of the confounding variables C the group receiving Drug A is comparable with the group receiving Drug B.

More formally, to estimate the average causal effect, $E[Y(1)] - E[Y(0)]$, by control for confounding, it is necessary that the counterfactual variables $Y(1)$ and $Y(0)$ be conditionally independent of the treatment received, X, given the confounding variables C. This conditional independence assumption can be written as $P(Y(1)|X = 1, C = c) = P(Y(1)|X = 0, C = c)$ and $P(Y(0)|X = 1, C = c) = P(Y(0)|X = 0, C = c)$; in other words, within strata of the confounding variables C, what happened to the group that received Drug A is representative of what would have happened to the group that received Drug B if they had in fact received Drug A; and similarly, within strata of the confounding variables C, what happened to the group that received Drug B is representative of what would have happened to the group that received Drug A if they had in fact received Drug B. If this holds, then average causal effects can be estimated using the following formula:

$$E[Y(1)] - E[Y(0)] = \sum_c \{E[Y|X = 1, C = c] - E[Y|X = 0, C = c]\}P(C = c).$$

Average causal effects can be estimated because, within strata of the confounding variables C, the groups that received Drug A and Drug B are comparable. The assumption that the counterfactual variables $Y(1)$ and $Y(0)$ are conditionally independent of the treatment received, X, given the confounding variables C is sometimes referred to as the assumption of "no-unmeasured-confounding" or as "exchangeability" or as "ignorable treatment assignment" or as "selection on observables" or sometimes as simply the "conditional independence" assumption. The assumption plays an important role in causal inference. In practice, data are collected on a sufficiently rich set of variables C so that the assumption that the groups are comparable within strata of C is at least approximately satisfied. Different techniques are available to make adjustment for the covariates C; adjustment can be made by stratification, regression, or propensity score modeling.

In the context of medical decision making, conditional independence is also important for a number of other problems. In many studies subjects drop out of a study before an outcome can be observed. It is not always clear that those subjects that remain in the study are comparable to those that drop out of the study. To make such problems tractable, a certain conditional independence assumption is sometimes made, namely that censoring status is conditionally independent of the potential outcomes $Y(1)$ and $Y(0)$ given the covariates C. In other words, it is assumed that within strata of the covariates C, the groups dropping out of the study are comparable with those who do not drop out; the set C contains all variables that affect both the dropout and the outcome. Conditional independence is also important in the analysis of surrogate outcomes in which some intermediate outcome is taken as a surrogate for the final outcome, which may be more difficult or expensive to collect data on than the surrogate. For example, the Prentice criteria for a valid surrogate outcome consist of the following three conditions: (1) the surrogate outcome, S, must be correlated with the true outcome, Y (i.e., S and Y must not be independent); (2) the surrogate outcome, S, must be affected by the exposure, X; and (3) the exposure, X, and the outcome, Y, should be conditionally independent given the surrogate, S. The third criterion captures the notion that all information about Y contained in the exposure A is in fact also available in the surrogate outcome, S.

Graphical Representation

More recently, graphical models and causal diagrams have been used to reason about independence and conditional independence relations. The technical details concerning such reasoning are beyond the scope of this entry. These diagrams make it clear that statistical association (lack of independence) can arise in a number of ways. The variables X and Y may be associated if X causes Y or if Y causes X. Even if neither X nor Y causes the other, the variables X and Y may be associated if they have some common cause C. In this case, if C contains all the common causes of X and Y, then X and Y will not be marginally independent but will be conditionally independent given C. Finally, if X and Y are independent but if they have some common effect C, then it will in general be the case that X and Y are conditionally associated given the common effect, C, that is, they will not be conditionally independent given C. Another interesting property relating conditional independence to these diagrams can be stated as follows: If the set C contains all variables that are common causes of X and Y and contains no common effects of X and Y, then if neither X nor Y

causes the other, then X and Y must be conditionally independent given C; thus, if X and Y are found not to be conditionally independent given such a set C, then one could conclude that either X has an effect on Y or Y has an effect on X.

These causal diagrams and their relation to conditional independence can also be helpful in understanding different forms of selection bias. Suppose that the occurrence of pneumonia and the level of sugar intake are such that sugar intake has no effect on pneumonia and that sugar intake and pneumonia are completely independent in the population. Sugar intake is however a risk factor for diabetes. Suppose now that all the subjects in a study are taken from a particular hospital. Hospitalization is then a common effect of both pneumonia and of diabetes, which is in turn an effect of sugar intake. By restricting the study to those subjects who are hospitalized one is implicitly conditioning on a common effect of pneumonia and sugar intake/diabetes, namely hospitalization. Thus, in the study it will appear that sugar intake and pneumonia are statistically associated because of the conditioning on the common effect, hospitalization, even though sugar intake has no effect on pneumonia; although sugar intake and pneumonia are marginally independent, they are not conditionally independent given hospitalization. This is an instance of what is often now called Berkson's bias. It is one of several types of selection bias that can be viewed as resulting from conditioning on a common effect and thereby inducing conditional association. See Hernán, Hernández-Diaz, and Robins (2004) for a discussion as to how the ideas of independence and conditional independence, causal diagrams, and the conditioning on a common effect can be used to understand better other forms of selection bias.

Tyler J. VanderWeele

See also Axioms; Causal Inference and Diagrams; Causal Inference in Medical Decision Making; Conditional Probability; Confounding and Effect Modulation; Ordinary Least Squares Regression; Propensity Scores

Further Readings

Berkson, J. (1946). Limitations of the application of fourfold table analysis to hospital data. *Biometrics Bulletin, 2,* 47–53.

Dawid, A. P. (1979). Conditional independence in statistical theory. *Journal of the Royal Statistical Society, Series B, 41,* 1–31.

Greenland, S. (2003). Quantifying biases in causal models: Classical confounding vs collider-stratification bias. *Epidemiology, 14,* 300–306.

Greenland, S., Pearl, J., & Robins, J. M. (1999). Causal diagrams for epidemiologic research. *Epidemiology, 10,* 37–48.

Hernán, M. A., Hernández-Diaz, S., & Robins, J. M. (2004). A structural approach to selection bias. *Epidemiology, 15,* 615–625.

Pearl, J. (1995). Causal diagrams for empirical research. *Biometrika, 82,* 669–688.

Pearl, J. (2000). *Causality: Models, reasoning, and inference*. Cambridge, UK: Cambridge University Press.

Prentice, R. L. (1989). Surrogate endpoints in clinical trials: Definition and operational criteria. *Statistics in Medicine, 8,* 431–440.

VanderWeele, T. J., & Robins, J. M. (2007). Directed acyclic graphs, sufficient causes and the properties of conditioning on a common effect. *American Journal of Epidemiology, 166,* 1096–1104.

Conditional Probability

The probability that event E occurs, given that event F has occurred, is called the conditional probability of event E given event F. In probability notation, it is denoted with $p(E|F)$. Conditional probabilities express the probability of an event (or outcome) under the condition that another event (or outcome) has occurred. You could also think of it as the probability within a particular subset, that is, in the subset of patients with event F in their history, the proportion that develop event E.

The conditional probability $p(E|F)$ is the ratio of the *joint probability* of events E and F, which is denoted as $p(E, F)$ or as $p(E \text{ and } F)$, and the *marginal probability* of event F, denoted with $p(F)$:

$$p(E|F) = \frac{p(E, F)}{p(F)}.$$

If the conditional information ("given event F") makes no difference to the probability of event E, then the two events E and F are said to be *conditionally independent*. For example, if F made no

difference to the estimate of the probability of E, that is, if $p(E|F+) = p(E|F-) = p(E)$, then E and F are said to be conditionally independent.

M. G. Myriam Hunink

See also Conditional Independence

Further Readings

Hunink, M. G. M., Glasziou, P. P., Siegel, J. E., Weeks, J. C., Pliskin, J. S., Elstein, A. S., et al. (2001). *Decision making in health and medicine: Integrating evidence and values*. Cambridge, UK: Cambridge University Press.

Confidence Intervals

Any decision in medicine is arrived at through a careful process of examining the evidence and deciding what would be the best course of action. There are many parts to this process, including the gathering of evidence that is of as high a quality as possible, the critical examination of this evidence, and a consideration of the interests of all those likely to be affected by the decision.

This entry concentrates on the process of critically examining the evidence and in particular the importance of confidence intervals to this process. Some key concepts will be defined, before discussing what is meant by statistical and clinical significance and then demonstrating the relevance and importance of confidence limits to medical decision making through examples from the literature.

Key Concepts

In classical statistical inference, the null hypothesis is the hypothesis that is tested. It is assumed to be true and is only rejected if there is a weight of evidence against it. The p value provides evidence in support of the null hypothesis. Technically speaking the p value is the probability of obtaining the study results (or results more extreme) if the null hypothesis is true. Thus a "small" p value indicates that the results obtained are unlikely when the null hypothesis is true and the null hypothesis is rejected in favor of the alternative hypothesis. Alternatively, if the p value is "large," then the results obtained are likely when the null hypothesis is true and the null hypothesis is not rejected. However, a large p value does not mean that the null hypothesis is correct: Absence of evidence does not equate to evidence of absence. The power of a study refers to the probability that a study will reject the null hypothesis if it is not true. While a nonsignificant p value may be indicative of the null hypothesis being correct, it may also be the result of the study lacking the power to reject the null hypothesis even though it is incorrect.

A result is said to be statistically significant if the p value is below the level set for defining statistical significance. This level is set before a study is undertaken. Conventionally, the cutoff value or two-sided significance level for declaring that a particular result is statistically significant is .05 (or 5%). Thus if the p value is less than this value, the null hypothesis is rejected and the result is said to be statistically significant at the 5% or .05 level.

For example, researchers in Australia (J. B. Dixon and colleagues) recently investigated whether adjustable gastric banding resulted in better glycemic control for type 2 diabetes compared with standard approaches to weight loss. At the end of the study, the 30 patients randomized to gastric band surgery (surgical) weighed, on average, 19.6 kg less than the 30 patients randomized to conventional weight loss therapy (standard), and the p value associated with this difference was less than .001. As this is less than .05, the authors were able to conclude that there was a statistically significant difference in the amount of weight lost between the two therapy groups.

However, a p value is not everything as it gives no information about the likely size of the result or the range of plausible values for it. This additional information is given by calculating a confidence interval for the result. Strictly speaking, a confidence interval represents the limits within which the true population value will lie for a given percentage of possible samples, and it can be calculated for any estimated quantity from the sample, including a mean or mean difference, proportion, or difference between two proportions. In practice, while not strictly speaking correct, it is not unreasonable to interpret; for example, the 95% confidence interval for the

mean as being the interval within which the true population mean is likely to lie with 95% certainty, or probability .95. For large samples (say greater than 60), the 95% confidence interval is calculated as

$$\bar{x} - 1.96 \times s/\sqrt{n} \text{ to } \bar{x} + 1.96 \times s/\sqrt{n},$$

where

$\bar{x}$ is the sample mean,

s is the sample standard deviation,

n is the number of observations in the sample, and

1.96 is the two-sided 5% point of the standard normal distribution.

The reason why we can use this simple formula is that, according to the Central Limit Theorem, the mean follows a Normal distribution. The Normal distribution is one of the fundamental distributions of statistics, and it is characterized such that the middle 95% of the data lie within +/– 1.96 standard deviations of its mean value. Conversely, only 5% of the data lie outside of these limits. The sample mean is an unbiased estimator of the true population mean, and while s is the sample standard deviation (for the data collected), the standard deviation of the mean is given by $s/\sqrt{n}$ and is often referred to as the standard error of the mean. Thus 95% of possible values for the true population mean will lie within $1.96 \times s/\sqrt{n}$ of the sample mean.

While the 95% confidence interval is the standard, it is possible to calculate a confidence interval to have greater or lesser coverage, that is, a 90% confidence interval or a 99% confidence interval, and this is done by changing the value of the cutoff point of the standard normal distribution in the expression above. For 90% limits, this changes to 1.64, and for 99% limits, this changes to 2.58.

For the above example, the 95% confidence interval for the mean difference in weight lost was 15.2 to 23.8 kg. Thus, the true mean difference in amount of weight lost between those with surgical intervention and standard therapy lies between 15.2 and 23.8 kg with 95% certainty.

Statistical Versus Clinical Significance

For medical decision making, in addition to statistical significance, it is essential to consider clinical significance, and it is in contributing to this that confidence intervals demonstrate their importance. A clinically significant difference is defined as a difference that is sufficiently large as to make a difference to patients or cause a change in clinical practice. Clinical significance is not a statistical concept, and its level cannot be set by a statistician. It must be arrived at through debate with knowledgeable subject experts, and the value set will depend on context. What is important to patients might be very different from what is considered important by policy makers or clinicians.

Even if a result is statistically significant, it may not be clinically significant, and conversely an estimated difference that is clinically important may not be statistically significant. For example, consider a large study comparing two treatments for high blood pressure; the results suggest that there is a statistically significant difference ($p < .001$) in the amount by which blood pressure is lowered. This p value relates to a difference of 2 mmHg between the two treatments, with a 95% confidence interval of 1.3 to 2.7 mmHg. Although this difference is statistically significant at the .1% level, it is not clinically significant as it represents a very small change in blood pressure and it is unlikely that clinicians and indeed their patients would change to a new treatment for such a marginal effect.

This is not simply a trivial point. Often in research presentations or papers, p values alone are quoted, and inferences about differences between groups are made based on this one statistic. Statistically significant p values may be masking differences that have little clinical importance. Conversely, it may be possible to have a p value greater than the magic 5% but with a genuine difference between groups, which the study did not have enough power to detect. This will be shown by the confidence interval being so large that it not only includes the null difference but also includes a clinically important difference.

There are two sides to clinical significance, depending on whether it is important to demonstrate that two treatments are different from one another (superiority) or whether it is of interest to demonstrate that their effect is the same (equivalence), and

confidence intervals have a part to play in both, as outlined below.

Importance in Medical Decision Making

Treatment Superiority

The importance of confidence intervals in studies to demonstrate superiority is best explained by reference to Table 1 and Figure 1. These display the results of seven (theoretical) studies comparing the same two treatments for superiority; that is, the object is to demonstrate that the two treatments are different. The table shows some possible point estimates of the effect size, together with the associated *p* values.

It is clear from this table that three of the studies are not statistically significant at the 5% level and four are. However, even assuming that a clinically important difference is two units on the measurement scale, it is impossible to tell which of these results are definitely of clinical importance, based on the *p* values and effect sizes alone. This information can only be obtained by reference to the confidence intervals as these will show the range of plausible values for the effect size, as shown in Figure 1. The left-hand vertical line in the figure represents the value that indicates the two treatments are equivalent and the right-hand line represents the value of a clinically important difference between the two treatments. Looking at Figure 1, it is clear that while four studies are statistically significant (C, D, F, and G), only one, Study G, is definitely clinically significant, as not only is the point estimate of the effect greater than the clinically significant difference of 2, but also the lower limit for the confidence interval is beyond this value. Of the other six studies, four, including two nonsignificant studies, may possibly be clinically significant as the upper limit of the confidence intervals includes the value set as being clinically important; however, given that the lower limit of the confidence interval is below the limit of clinical significance, clinical significance cannot definitely be inferred.

Table 1 Results of six studies examining the difference between two treatments

Study	*Size of Difference*	*p value*
A	0.8	>.05
B	0.8	>.05
C	0.8	<.05[a]
D	1.8	<.05[a]
E	2.5	>.05
F	2.5	<.05[a]
G	2.5	<.05[a]

a. These values are statistically significant.

This examination of the confidence intervals of the effect size is particularly important in the case of studies that do not reach statistical significance, as mentioned above. Even if the *p* value is greater than .05, it may be that the null hypothesis is genuinely true, or it may be that the study lacked the power to reject the null hypothesis. Looking at Figure 1, Study B represents a case of the former—this result is neither statistically significant nor clinically significant—while Study E is an example of the latter. The point estimate for E is larger than a clinically important difference, but the confidence interval is so large that it includes the null difference.

Treatment Equivalence

Confidence intervals are equally important in studies that examine whether two treatments are equivalent in their effect. For equivalence studies, conclusions will always be based on an examination of the confidence intervals. Before an equivalence trial is carried out the limits of equivalence are agreed on, so that after the trial a decision can be made as to whether the treatments are, to all intents and purposes, the same in their effect. These prespecified limits should be narrow enough to exclude any difference of clinical importance. After the trial, equivalence is usually accepted if the confidence interval for any observed treatment difference falls entirely within the limits of equivalence and includes a value of zero difference. If one of the limits falls outside the limits of equivalence, it would imply that one of the plausible values for the treatment effect was at least as large as a clinically important difference.

A study by I. F. Burgess and colleagues published in 2005 examined whether 4% dimeticone lotion was equivalent to phenothrin, the most commonly used pediculicide, for the treatment of head louse

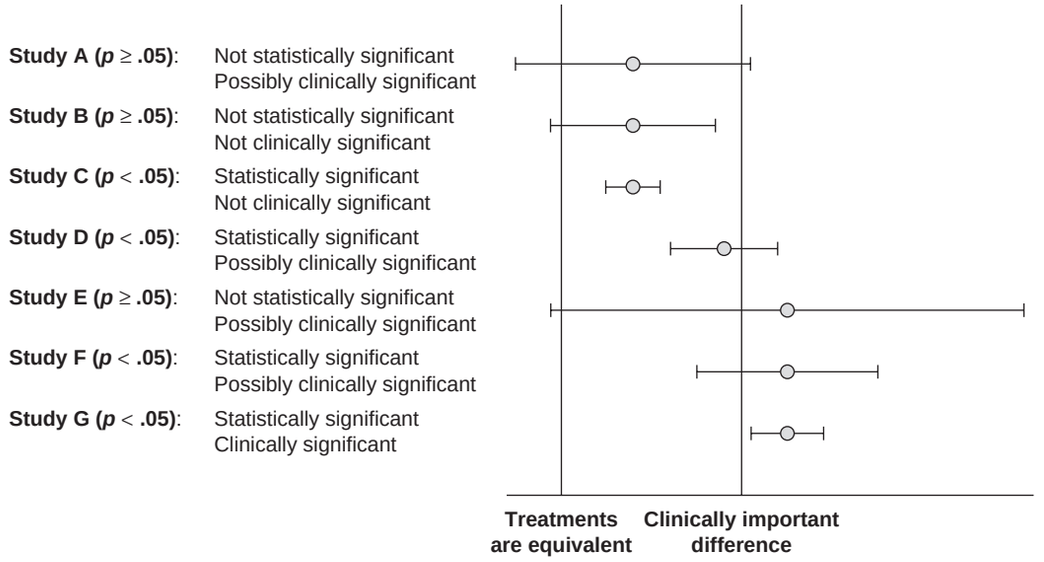


Figure 1 Statistical and clinical significance: Results of seven studies (point estimates together with confidence intervals)

infestation. The main outcome was cure of infestation or reinfestation after cure. Before the study began, it was decided that the two treatments would be declared equivalent if the results were within 20% between treatment groups, based on the 95% confidence intervals, that is, if the upper and lower limits for the 95% confidence interval for the difference between groups were both less than 20% either side of no difference. Of the 127 individuals randomized to receive dimeticone, 89 were either cured or reinfested after cure at follow-up (70%), while 94 of the 125 followed up in the phenothrin group were cured or reinfested after cure (75%). Thus, 5% fewer individuals in the dimeticone group were cured or reinfested after cure, and the 95% confidence interval for this difference was –16% to 6%. As these 95% limits were within the 20% limits of equivalence set before the study was undertaken, as illustrated by Figure 2, the researchers were able to conclude that the two treatments were equivalent to within 20%.

While a *p* value is a useful starting point, it would be ill advised to make a decision based on this single piece of information, and it is vital to examine the estimate of any effect and its associated confidence interval before making a decision. This will give a range of plausible values for the effect size and will assist one in deciding whether

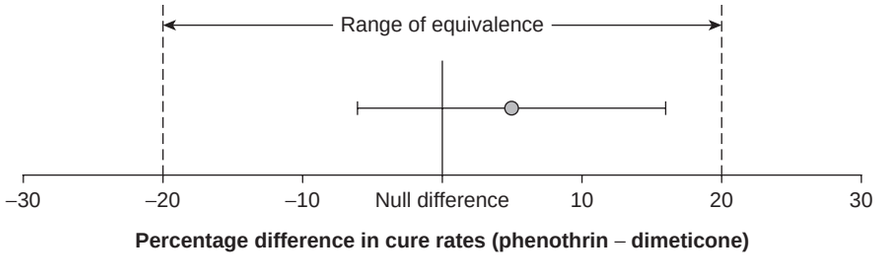


Figure 2 Estimated difference in cure and reinfestation after cure rates between dimeticone and phenothrin, together with the 95% confidence interval for the difference.

any difference found is of clinical importance or whether the study had sufficient power to reject the null hypothesis.

Jenny V. Freeman

See also Effect Size; Hypothesis Testing; Managing Variability and Uncertainty; Sample Size and Power

Further Readings

Altman, D., Machin, D., Bryant, T., & Gardner, S. (2000). *Statistics with confidence.* Oxford, UK: Wiley Blackwell.

Armitage, P., Berry, G., & Matthews, J. N. S. (2001). *Statistical methods in medical research* (4th ed.). New York: Blackwell Science.

Burgess, I. F., Brown, C. M., & Lee, P. N. (2005). Treatment of head louse infestation with 4% dimeticone lotion: Randomised controlled equivalence trial. *British Medical Journal, 330,* 1423–1426.

Campbell, M. J., Machin, D., & Walters, S. J. (2007). *Medical statistics: A textbook for the health sciences.* Chichester, UK: Wiley.

Dixon, J. B., O'Brien, P. E., Playfair, J., Chapman, L., Schachter, L. M., Skinner, S., et al. (2008). Adjustable gastric banding and conventional therapy for type 2 diabetes. *Journal of the American Medical Association, 299*(3), 316–323.

Kirkwood, B. R., & Sterne, J. A. C. (2003). *Essential medical statistics* (2nd ed.). Oxford, UK: Blackwell Science.

Petrie, A., & Sabin, C. (2005). *Medical statistics at a glance* (2nd ed.). Oxford, UK: Blackwell Science.

Confirmation Bias

Confirmation bias is the tendency for people to search for or interpret information in a manner that favors their current beliefs. This entry communicates psychological research on confirmation bias as it relates to medical decision making. This will help medical professionals, patients, and policy makers consider when it might pose a concern and how to avoid it. The focus is on choosing a test for a simple case of medical diagnosis. The first section discusses how inference and information search ought to take place; the second section discusses confirmation bias and other possible errors; the final section discusses how to improve inference and information search.

How Should Inference and Information Acquisition Proceed?

No choice of diagnostic tests can cause confirmation bias if the test results are assimilated in a statistically optimal manner. Therefore, this section first discusses how to incorporate test results in a statistically optimal (Bayesian) way. It then discusses various strategies to select informative tests.

Suppose that the base rate of a disease (d) in males is 10% and that a test for this disease is given to males in routine exams. The test has 90% sensitivity (true positive rate): 90% of males who have the disease test positive. Expressed in probabilistic notation, $P(\text{pos}|d) = 90\%$. The test has 80% specificity: $P(\text{neg}|\sim d) = 80\%$ (20% false-positive rate), meaning that 80% of males who do not have the disease correctly test negative. Suppose a male has a positive test in routine screening. What is the probability that he has the disease? By Bayes's theorem (see Figure 1, Panel A),

$$P(d|\text{pos}) = P(\text{pos}|d)P(d)/P(\text{pos}),$$

where

$$P(\text{pos}) = P(\text{pos}|d)P(d) + P(\text{pos}|\sim d)P(\sim d).$$

Therefore,

$$P(d|\text{pos}) = (.90 \times .10)/(.90 \times .10 + .20 \times .90) = .09/.27 = 1/3.$$

Alternately (see Figure 1, Panels B and C), it is possible to count the number of men with the disease and a positive test, and who test positive without having the disease:

$$P(d|\text{pos}) = \text{num}(d \ \& \ \text{pos})/\text{num}(\text{pos}) = 9/(9 + 18) = 1/3,$$

where

$$\text{num}(\text{pos}) = \text{num}(\text{pos} \ \& \ d) + \text{num}(\text{pos} \ \& \ \sim d).$$

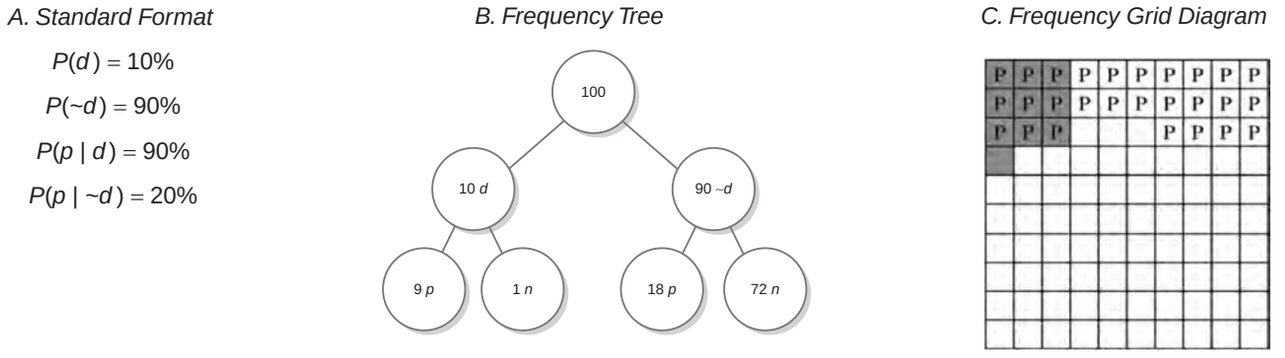


Figure 1 Different formats for presenting probabilistic information

Notes: The means by which probabilistic information is presented have a large impact on how meaningful the information is to people. The standard probability format (Panel A) is complicated for people to work with, although they can be trained to do so. Both the frequency tree (Panel B) and frequency grid diagram (Panel C) provide more meaningful representations of the information. The term "*d*" denotes the disease; "*~d*" absence of the disease; "*p*" denotes a positive test. In Panel C, shaded cells denote presence of the disease.

But how should a diagnostic test be chosen in the first place? The fundamental difficulty is that which test is most useful depends on the particular outcome obtained, and the outcome cannot be known in advance. For instance, the presence of a particular gene might definitively predict a disease, but that gene might occur with only one in a million probability. Another test might never definitively predict the disease but might always offer a high degree of certainty about whether the disease is present or not.

Optimal experimental design ideas provide a reasonable framework for calculating which test, on balance, will be most useful. All these ideas are within the realm of Savage's Bayesian decision theory, which defines the subjective expected usefulness (utility) of a test, before that test is conducted, as the average usefulness of all possible test results, weighting each result according to its probability.

In the case of a test T that can either be positive (pos) or negative (neg), the test's expected utility (eu) would be calculated as follows:

$$\mathrm{eu}(T) = P(\mathrm{pos}) \times u(\mathrm{pos}) + P(\mathrm{neg}) \times u(\mathrm{neg}),$$

where u corresponds to utility. Various optimal experimental design ideas quantify, in different ways, the usefulness of particular test outcomes. Suppose one wishes to use improvement in probability of correct diagnosis to quantify the usefulness of possible diagnostic tests. (This equates to minimizing error.) The probability gain (pg) of a test, with respect to determining whether or not a patient has disease d, is calculated as follows:

$$\mathrm{eu}_{\mathrm{pg}}(T) = P(\mathrm{pos}) \times [\max(P(d|\mathrm{pos}), P(\sim d|\mathrm{pos})) - \max(P(d), P(\sim d))] + P(\mathrm{neg}) \times [\max(P(d|\mathrm{neg}), P(\sim d|\mathrm{neg})) - \max(P(d), P(\sim d))].$$

Suppose the goal is to learn whether or not a patient has a disease that occurs in 10% of patients. Test 1 has 95% sensitivity and 85% specificity. Test 2 has 85% sensitivity and 95% specificity. Which test maximizes probability gain? Test 1 has probability gain 0, though Test 2 has probability gain .04. Although Test 1 has high sensitivity, its low specificity is problematic as the base rate of the disease is only 10%. Test 1 does not change the diagnosis of any patient, because, irrespective of whether it is positive or negative, the patient most likely does not have the disease. Test 2's much higher specificity, however, reduces false positives enough so that a majority of people who test positive actually have the disease.

It can be helpful, as an exercise, to consider possible tests' probability gain before ordering a test, in situations where the relevant environmental probabilities are known. In real medical diagnosis, additional factors, such as a test's cost and its potential to harm the patient, should also be taken into account.

Confirmation Bias and Other Errors

Do people typically reason following Bayes's theorem? Do physicians intuitively select useful tests for medical diagnosis? If human cognition and behavior are suboptimal, do they reflect confirmation bias?

From early research on Bayesian reasoning through the present, there has been evidence that people are either too conservative or too aggressive in updating their beliefs. Some research suggests that people make too much use of base rates (the proportion of people with a disease), as opposed to likelihood information (a test result). Other research suggests that people make too little use of base rates, relying on likelihood information too much.

Do these errors lead to systematically overweighting one's working hypothesis (e.g., the most probable disease)? Note that test results can either increase or decrease the probability of a particular disease. Because of this, neither being too conservative nor being too aggressive in updating beliefs in response to test results would consistently give a bias to confirm one's working hypothesis. Thus, while there is plenty of evidence that people (including physicians) sometimes update too much and sometimes too little, that does not necessarily imply confirmation bias.

If people have personal experience with environmental probabilities, their inferences are often quite accurate. In routine diagnostic and treatment scenarios, in which individual practitioners have previously experienced dozens, hundreds, or even thousands of similar cases and have obtained feedback on the patients' outcomes, physicians' intuitions may be well-calibrated to underlying probabilities. Little if any confirmation bias would be expected in these situations. In situations in which relevant data are available but practitioners do not have much personal experience, for instance because rare diseases are involved, intuitions may not as closely approximate Bayes's theorem.

Confirmation Bias in Inference

Apart from the general difficulty in probabilistic reasoning, how might people fall victim to confirmation bias per se? Below, several situations are described that might lead to confirmation bias.

1. If people obtain useless information but think it supports their working hypothesis, that could lead to confirmation bias. Suppose a physician asks a patient about the presence of a symptom that, if present, would support a particular disease diagnosis. Suppose the patient tends to answer "yes" in cases where the question is unclear, so as to cooperate. If the physician does not take the patient's bias to answer "yes" into account when interpreting the answer to the question, the physician could be led, on average, to be excessively confident in his or her diagnosis.

2. Sometimes a test's sensitivity (its true positive rate) is conflated with its positive predictive value (the probability of the disease given a positive result). In situations where the sensitivity is high, but specificity is low or the base rate of the disease is very low, this error can cause confirmation bias. For instance, among people from low-risk populations, a substantial proportion of people with positive HIV test results do not have HIV. Some counselors, however, have wrongly assumed that a positive test means a person has HIV.

3. There are many situations in which people want to reach certain conclusions or maintain certain beliefs, and they are quite good at doing so. Imagine that a physician has diagnosed a patient with a serious illness and started the patient on a series of treatments with serious side effects. The physician might be more likely than, say, an impartial second physician, to discount new evidence indicating that the original diagnosis was wrong and that the patient had needlessly been subjected to harmful treatments.

4. Finally, people sometimes interpret ambiguous evidence in ways that give the benefit of the doubt to their favored hypothesis. This is not necessarily a flaw in inference. If one's current beliefs are based on a great deal of information, then a bit of new information (especially if from an unreliable source) should not change beliefs drastically. Whether a physician interprets a patient's failure to return a smile from across the room as indicating the patient didn't see him or her or as a snub will likely be influenced by whether the patient has previously been friendly or socially distant. Similarly, suppose an unknown researcher e-mails his or her discovery that AIDS is caused by nefarious extraterrestrials. Given the outlandish nature of the claim, and the unknown status of the "researcher," it would be wise to demand a lot of

corroborating evidence before updating beliefs about causation of AIDS at all, given this report. The overriding issue is that one's degree of belief, and amount of change of belief, should correspond to the objective value of the evidence.

Information Acquisition and Confirmation Bias

Do people use statistically justifiable strategies for evidence acquisition, for instance when requesting a test or asking a patient a question? Are people prone to confirmation bias or other errors?

Psychological experiments suggest that people are very sensitive to tests' usefulness when deciding which test to order. Any testing strategy not solely concerned with usefulness will be inefficient. However, if test results are evaluated in a Bayesian way, then although some information acquisition strategies are more efficient than others, none will lead to confirmation bias. Thus, improving probabilistic inference is a first step toward guarding against confirmation bias.

Positivity and extremity are additional factors that may contribute to people's choices of tests. Positivity is the tendency to request tests that are expected to result in a positive result, or a "yes" answer to a question, given that the working hypothesis is true. Extremity is a preference for tests whose outcomes are very likely or very unlikely under the working hypothesis relative to the alternate hypothesis. The evidence substantiating people's use of these particular strategies is somewhat murky. However, use of these testing strategies, together with particular biased inference strategies, could lead to confirmation bias.

Improving Inference and Information Acquisition

Improving Inference

The means by which probabilistic information is presented are important, and evidence suggests that either personal experience or appropriate training can help people meaningfully learn particular probabilities. The literature suggests several strategies to improve Bayesian inference:

1. Present information in a meaningful way. Figure 1, Panels B and C, illustrates two means of presenting equivalent information, in which the information is presented in terms of the *natural frequencies* of people with (and without) the disease who have a positive or negative test. These formats better facilitate Bayesian reasoning than does the standard probability format (Figure 1, Panel A). Simulating personal experience and providing feedback may be even more effective.
2. Teach Bayesian inference. Although people do not intuitively do very well with standard probability format problems, people can be trained to do better, especially when the training helps people use natural frequency formats for representing the probabilistic information.
3. Obtain feedback. Feedback is critical for learning environmental probabilities, such as base rates of diseases, and distribution of test outcomes for people with and without various diseases. Feedback is also critical for learning when those probabilities change, for instance because of an outbreak of a rare disease. Both individual practitioners and policy makers could think about how to ensure that feedback can be obtained, and patients and citizens should demand that they do so.

Improving Information Acquisition

People are not adept at maximizing either probability gain or individually specified utilities when information is presented in the standard probability format. Taking care to ensure that known statistical information is meaningful may be the single most important way to improve practitioners' capacity for good inference and information acquisition in medical decision making. Use of personal experience and feedback to convey probabilistic information in simulated environments can also facilitate Bayesian performance.

Beyond Confirmation Bias

While confirmation bias in inference and information acquisition may exist, it should be seen in the broader context of statistical illiteracy and misaligned incentives. Those problems may be the root of what can appear to be confirmation bias, rather than any inherent cognitive limitations that people have. For instance, the desire to make a patient feel that he or she is being

treated well, and to guard against the possibility of litigation, might lead to ordering a medically unnecessary (and potentially harmful) CT scan following mild head trauma. At the level of basic research, the source of funding can influence the conclusions that are reached. From a policy standpoint, the goal should be to make individual and institutional incentives match public health objectives as closely as possible.

Jonathan D. Nelson and Craig R. M. McKenzie

See also Bayes's Theorem; Biases in Human Prediction; Cognitive Psychology and Processes; Conditional Probability; Deliberation and Choice Processes; Errors in Clinical Reasoning; Evidence Synthesis; Expected Utility Theory; Expected Value of Sample Information, Net Benefit of Sampling; Hypothesis Testing; Probability; Probability Errors; Subjective Expected Utility Theory

Further Readings

Baron, J. (1985). *Rationality and intelligence*. Cambridge, UK: Cambridge University Press.

Baron, J., Beattie, J., & Hershey, J. C. (1988). Heuristics and biases in diagnostic reasoning: II. Congruence, information, and certainty. *Organizational Behavior and Human Decision Processes, 42,* 88–110.

Baron, J., & Hershey, J. C. (1988). Heuristics and biases in diagnostic reasoning: I. Priors, error costs, and test accuracy. *Organizational Behavior and Human Decision Processes, 41,* 259–279.

Gigerenzer, G., Gaissmaier, W., Kurz-Milcke, E., Schwartz, L. M., & Woloshin, S. (2008). Helping doctors and patients make sense of health statistics. *Psychological Science in the Public Interest*, *8*(2), 53–96.

Gigerenzer, G., & Hoffrage, U. (1995). How to improve Bayesian reasoning without instruction: Frequency formats. *Psychological Review, 102*(4), 684–704.

Kahneman, D., & Tversky, A. (1972). Subjective probability: A judgment of representativeness. *Cognitive Psychology, 3*(3), 430–454.

Klayman, J. (1995). Varieties of confirmation bias. *Psychology of Learning and Motivation*, *32,* 385–418.

Klayman, J., & Ha, Y.-W. (1987). Confirmation, disconfirmation, and information. *Psychological Review, 94,* 211–228.

McKenzie, C. R. M. (2004). Hypothesis testing and evaluation. In D. J. Koehler & N. Harvey (Eds.), *Blackwell handbook of judgment and decision making* (pp. 200–219). Oxford, UK: Blackwell.

McKenzie, C. R. M. (2006). Increased sensitivity to differentially diagnostic answers using familiar materials: Implications for confirmation bias. *Memory and Cognition, 34,* 577–588.

Nelson, J. D. (2005). Finding useful questions: On Bayesian diagnosticity, probability, impact and information gain. *Psychological Review, 112*(4), 979–999.

Nelson, J. D. (2008). Towards a rational theory of human information acquisition. In M. Oaksford & N. Chater (Eds.), *The probabilistic mind: Prospects for rational models of cognition* (pp. 143–163). Oxford, UK: Oxford University Press.

Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology, 2*(2), 175–220.

Oaksford, M., & Chater, N. (2003). Optimal data selection: Revision, review, and reevaluation. *Psychonomic Bulletin & Review, 10,* 289–318.

Savage, L. J. (1954). *The foundations of statistics*. New York: Wiley.

Sedlmeier, P., & Gigerenzer, G. (2001). Teaching Bayesian reasoning in less than two hours. *Journal of Experimental Psychology: General, 130,* 380–400.

Conflicts of Interest and Evidence-Based Clinical Medicine

Physicians' financial interests may have an unconscious influence on their interpretation of the scientific evidence relevant to the treatments they choose for their patients. This indirect determinant of physician behavior has not been extensively studied, but the causal process can be sketched using general principles of cognitive and social psychology and of marketing. The influence of financial interests on clinician knowledge is a distinct topic from their influence on clinician action. The latter topic has been the main concern in discussion of conflicts of interest induced by gifts from pharmaceutical or medical device manufacturers, by managed care or insurance rules, or by payments from interested industries to the directors of nonprofit hospitals or accompanying the intrusion of the commercial management of hospitals and clinics into the doctor-patient relationship.

Ideally, the physician applies medicine's best treatments appropriately for each patient, rationally considering the scientific evidence concerning

the treatment's efficacy as well as the patient's unique circumstances. Evidence that would support a treatment generally includes scientific studies proving it works as well as or better than the alternatives, and evidence that the burden of side effects or costs, per unit of health improvement produced, is not excessive. Relevant considerations for the particular patient may include individual characteristics that change the probable success of a treatment, such as being more robust or more fragile than the typical patient. It is also rationally and ethically appropriate to consider the patient's financial or social resources, such as ability to pay for a treatment without bankrupting the family or the capability of adhering to required behavioral changes or medical care demands over the long term. Potential biases in physician judgments of patient resources are not considered further here.

Financial interests may exert an unconscious influence on physicians' interpretation of evidence relevant to the treatment of their patients. An individual clinician's reading of the literature regarding the benefits and costs of patients' treatments can be distorted by the fact that he or she is able to provide some treatments but not others. The same sort of process can sway the production and interpretation of professional association guidelines in a way that promotes the profession, as against the interests of the competing specialties, the patient, or society in general.

Sometimes, financial interests may induce in the physician a kind of psychological blindness that impedes the ability of otherwise ethical people to recognize that the scientific evidence does not support their way of practice. Physicians may not be aware that their judgment is distorted this way even though an objective observer might call their decisions "irrational" with respect to the evidence. In contrast, a physician who consciously chose to use treatments that she or he knew to be ineffective or harmful for the patient because she or he could collect higher fees would be described as "unethical."

Example of Irrational Treatment Decisions Due to Financial Motivation

An orthopedic surgeon opted to simplify his work life by concentrating on only a few types of back surgery, to be able to spend more time with his family. He put out the word seeking referrals and organized the clinic to allow himself to spend as little time in the office and as much in the operating room as possible. During 10 years of this arrangement, evidence accumulated in the literature that back surgery is indicated for a smaller proportion of those who complain of low back pain than had previously been thought and that an elaborate sequence of diagnostic measures and trial treatments can identify patients who likely won't be cured by surgery but may be helped by alternative measures. The physician nonetheless has continued doing the same familiar operations on most patients who come through his door, after only a brief discussion of the surgical options in an initial consultation in the office. There is a compelling financial motivation for the physician to maintain his surgical volume. His family depends on the current level of income, as do the clinic employees, his partners who co-own the clinic with him, and the bank. At this point in his career, doing these procedures is the only skill the physician has that can bring in this much income. However, he has never suggested, even in jest, that he is "just in it for the money," and his friends and coworkers know him to be honestly concerned about his patients. When asked about the studies suggesting more discerning assessment of the patient is required, the physician says they are not applicable to his practice because he is following professional guidelines.

Unconscious Irrationality Due to Conflict of Interest

The physician in the vignette, whose practice is concentrated on a few lucrative surgical procedures, ignored or dismissed the evidence suggesting the procedures are not indicated for many patients. Physicians often experience conflict between the demand for billable activity and their commitment to do what is best for the patient. When they weigh the evidence in resolving this conflict between the competing values, they may give unconscious priority to their own financial interests. It is better for both patient and physician if physicians can be conscious of motivations that may blind them to the scientific evidence.

To promote accurate physician self-awareness, it has been recommended that physicians should disclose their financial interests to patients, as well as the constraints imposed by their employers or the patients' insurance. Thus, the back surgeon could acknowledge that his fee covers his expenses and

supports him comfortably, and in exchange he does his best for his patients and makes ongoing efforts to keep up with the state of the art. Talking about what the fees buy—the physician's expert interpretation of the emerging evidence—is as pertinent as discussing the expected efficacy of the treatment and its alternatives, the probabilities of various morbidities, and how other people have adjusted to the outcomes, especially if there is the possibility that patient distrust may undermine adherence. The physician's frank discussion of his interests and his efforts helps the patient engage in informed decision making. Such conversations can also help the physician maintain rationality and integrity: when the motives are acknowledged, the physician is less likely to be unconsciously influenced to recommend a treatment of inferior efficacy just because it is convenient or profitable, or is what others in his subspecialty do. Anticipation of such conversations can motivate the physician to keep up with the evidence so he or she can say in good faith that he or she is offering the patient the best treatments known.

The Role of Professional Organizations in the Interpretation of Scientific Evidence

Physicians associate with similar physicians in professional organizations that provide mutual support, including information on the newest treatment modalities and guidelines on the manner of practice judged to help the members thrive through the appropriate use of their special knowledge and skills to care for patients. Delegating the burden of evaluating treatments in this way can muffle the physician's awareness of the balance of evidence regarding the recommended treatment modalities. As a result of such informational filters, a patient with localized prostate cancer, for example, might be given radiation if he visits almost any radiation oncologist, surgery if he visits any urologist, or expectant management if he visits any general internist. It is not simply that these options are in equipoise for all such patients; these contradictions highlight the impact of professional organizations' shaping of members' views.

Unconscious psychological processes may contribute to this influence at two stages, in the production and the utilization of the guidelines, as illustrated in Figure 1. First, members of professional organizations who have been honored with appointment to a task force that will produce a guideline statement may experience expectations related to solidarity with the group. The committee may manifest polarization, in which the group's decision may express a more extreme position on a shared value than most of the individuals would hold on their own. Figure 1 shows the relation between the guidelines (B) and the possible treatments consistent with the scientific evidence (A). The effect of polarization is that the guideline highlights only a subset of the supportable interpretations of the scientific literature, and it may extend a little beyond what the general field may find supported. Committees are less likely to produce self-serving guidelines (though it is still possible) when they adopt the discipline of formally meta-analyzing only randomized controlled trials. The competing guidelines authored by Gharib and Surks regarding screening for subclinical hypothyroidism illustrate this point. The second process has to do with the reader's comprehension and recall of the published guidelines. Most guidelines consist of a general recommendation and a list of exceptions or qualifications. When individual practitioners read the guidelines (C in the figure), and again when they recall them at the point of use (D), often the gist is recalled while the detailed exceptions are forgotten. Lines carefully drawn when a guideline statement was composed may be missed or forgotten by the reader.

Beneficial Effects of Financial Motivation

Financial motives do not always interfere with rational patient treatment. Administrative power, including monetary bonuses or penalties contingent on individual or clinicwide performance, is one of the most powerful tools available for changing physicians' behavior. Aligning the rewards with the evidence-based practices can be an effective component of a program to improve medical care. But financial reward schemes can have unintended consequences, distorting physician behavior without benefit to the patient. For example, in the United States, to sustain themselves, the private (e.g., insurance) and public (e.g., Medicaid) systems that pay for medical care impose limits on allowable charges per visit. This has the unintended side effect that low-priority concerns such as prevention may be neglected. To compensate for

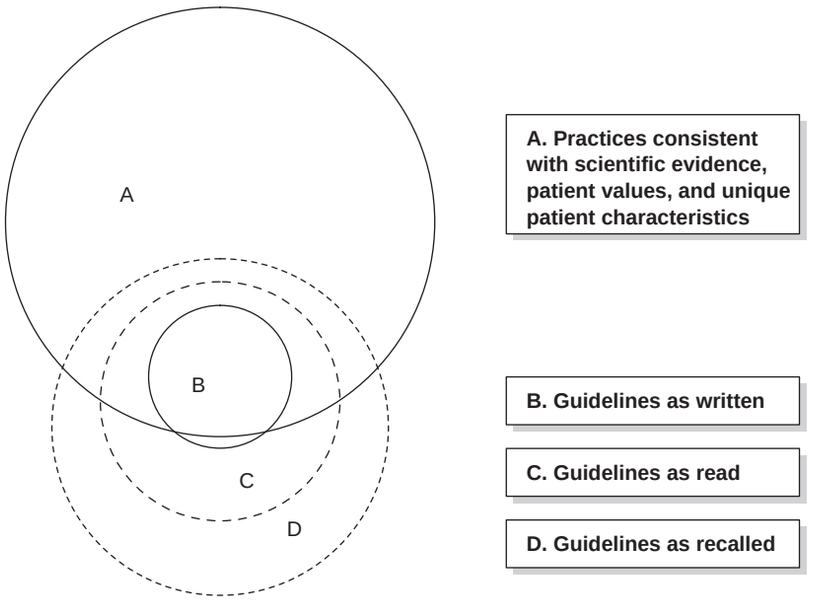


Figure 1 Psychological influences on guideline production and utilization

Note: Illustration of psychological influences on the guideline production and utilization process that may bias physician practices. A: the set of possible treatments supported by evidence. B: impact of group polarization on the guideline writers. C and D: impact of gist highlighting processes in comprehension and recall of guidelines.

this, local administrators may advise physicians to deal with just one patient concern during each visit, and schedule additional visits to address other issues. While this stratagem may help a clinic be financially viable, it imposes an additional burden on those patients responsible for co-payment or lacking transportation.

Contributing Factors

Individual clinicians' interpretation of the literature's evidence regarding the benefits and costs of their patients' potential treatments can be unconsciously distorted by how much they can be paid for providing the treatments. The rewards physicians receive for some practices may make it difficult for them to see that they need to give up those practices when a different method of treatment is proven better. Group polarization effects in the production of guidelines, and simplification processes in the comprehension and recall of recommendations, also contribute to the persistence of nonoptimal treatment practices that are financially rewarded.

Robert M. Hamm

See also Bias in Scientific Studies; Clinical Algorithms and Practice Guidelines; Evidence-Based Medicine; Irrational Persistence in Belief; Motivation

Further Readings

Bigos, S., Bowyer, O., Braen, G., Brown, K., Deyo, R., Haldemann, S., et al. (1994). *Acute low back problems in adults: Clinical practice guideline No. 14* (AHCPR Publication No. 95-0642). Rockville, MD: Agency for Health Care Policy and Research, Public Health Service, U.S. Department of Health and Human Services.

Brody, H. (2005). The company we keep: Why physicians should refuse to see pharmaceutical representatives. *Annals of Family Medicine, 3*(1), 82–85.

Bursztajn, H. J., & Brodsky, A. (1999). Captive patients, captive doctors: Clinical dilemmas and interventions in caring for patients in managed health care. *General Hospital Psychiatry, 21*(4), 239–248.

Cain, D. M., & Detsky, A. S. (2008). Everyone's a little bit biased (even physicians). *Journal of the American Medical Association, 299*(24), 2893–2895.

Gharib, H., Tuttle, R. M., Baskin, H. J., Fish, L. H., Singer, P. A., & McDermott, M. T. (2005). Subclinical

thyroid dysfunction: A joint statement on management from the American Association of Clinical Endocrinologists, the American Thyroid Association, and the Endocrine Society. *Journal of Clinical Endocrinology & Metabolism, 90*(1), 581–585.

Isenberg, D. J. (1986). Group polarization: A critical review and meta-analysis. *Journal of Personality and Social Psychology, 50,* 1141–1151.

Jones, J. W., & McCullough, L. B. (2007). Are ethics practical when externals impact your clinical judgment? *Journal of Vascular Surgery, 45*(6), 1282–1284.

Rivero-Arias, O., Campbell, H., Gray, A., Fairbank, J., Frost, H., & Wilson-MacDonald, J. (2005). Surgical stabilisation of the spine compared with a programme of intensive rehabilitation for the management of patients with chronic low back pain: Cost utility analysis based on a randomised controlled trial. *British Medical Journal, 330*(7502), 1239.

Smith, W. R. (2000). Evidence for the effectiveness of techniques to change physician behavior. *Chest, 118,* 8S–17S.

Sulmasy, D. P., Bloche, M. G., Mitchell, J. M., & Hadley, J. (2000). Physicians' ethical beliefs about cost-control arrangements. *Archives of Internal Medicine, 160*(5), 649–657.

Surks, M. I., Ortiz, E., Daniels, G. H., Sawin, C. T., Col, N. F., Cobin, R. H., et al. (2004). Subclinical thyroid disease: Scientific review and guidelines for diagnosis and management. *Journal of the American Medical Association, 291*(2), 228–238.

Confounding and Effect Modulation

The relationship between a predictor or study variable and an outcome variable may vary according to the value of a third variable, often called a confounding variable or an effect modulator. This entry clarifies the distinction between confounding and effect modulation (also called moderation or mediation) through the use of path diagrams. The statistical tests for establishing these three relationships are somewhat different (main effects model only for establishing confounding variables; main effects model with interaction term for establishing moderating variables and Sobel-like tests based on a series of regression for establishing mediating variables), so they are discussed separately and their interpretation clarified by example.

Overview

An important feature of a regression model is its ability to include multiple covariates and thereby statistically adjust for possible imbalances in the observed data before making statistical inferences. This process of adjustment has been given various names in different fields of study. In traditional statistical publications, it is sometimes called the *analysis of covariance,* while in clinical and epidemiologic studies it may be called *control for confounding*. Interactions between covariates may also be included in the model and regarded as *effect modifiers* in the sense that the effect on the outcome differs according to the level of the moderator variable. When an outcome is correlated with a study variable but the relationship disappears when adjusted by a third variable, the third variable is often called a mediating variable or mediator. In an epidemiological study of the strength of the association between smoking status and lung cancer, the relationship may be affected by other variables such as the drinking habits, extent of exposure to tobacco smoke, or age of the subject, or other personal or environmental conditions. Variables other than smoking that affect the relationship of smoking and lung cancer are often described as *modulating variables*. Modulating variables are further classified as *confounding variables, effect moderators, effect modifiers,* or *mediating variables* according to their finer properties. Some terms and interpretations used to distinguish different types of modulation are based on statistical definitions and may thus be measured and tested objectively. In other instances, judgments regarding causality will be required, thus introducing concepts not readily amenable to statistical analysis.

Definitions of terms such as *mediation, moderation,* and *confounding* have been questioned because of their implied dependence on the unquantifiable concept of causality. This entry illustrates these through simple statistical modeling and figures, and provides examples of the roles of confounding, mediating, and moderating variables. The first example illustrates the role of "helplessness" as a moderating variable where patients with

a "low" helplessness index could have decreasing depression even with an increasing "swollen joint count," whereas patients with a "high" helplessness index have the opposite relationship (increasing depression with an increasing swollen joint count). The second example illustrates the "mediating" role of pain and comorbidities on the association observed between body mass index (BMI) and total unhealthy days (TUD). The association vanishes when the mediating variables are adjusted in the model. The analyses for each of the examples are adjusted for various confounding variables.

Terms and Definitions

The purpose of many epidemiological studies is to determine the effect of a *predictor variable* or *risk factor X* on an *outcome variable Y* while accounting for the effects of another influential variable, denoted by *Z*. To facilitate the discussion, a simple statistical model can be written as

$$Y = a + bX + cZ + dXZ + \varepsilon,$$

where the unknown coefficients *a*, *b*, *c*, and *d* are to be determined by statistical model fitting and ε denotes a random error. The variable *X*, whose effect on *Y* is to be studied, may be called the study, experimental, or condition variable in an experimental study or a risk factor in an observational study. Variable *Z* provides information in addition to the study variable, which may affect the relationship of outcome *Y* to predictor *X*. The variable *Z* is generally described as modulating or modifying the effect of *X* on *Y* but may further be classified as a confounding, moderating, or mediating variable. The relationship of *Y*, *X*, and *Z* is often depicted schematically as in Figures 1, 2, and 3, respectively, for illustrating the three types of relationship.

When either the estimated coefficient of a variable in the model or the relevant (Pearson's) correlation is statistically different from zero (usually at the 5% level), the role of the variable is described as "significant." The variable *Z* may have any of the several different roles that may affect the relationship of *Y* to *X*. These roles are often described in the epidemiological and psychological literature using the following terms and definitions.

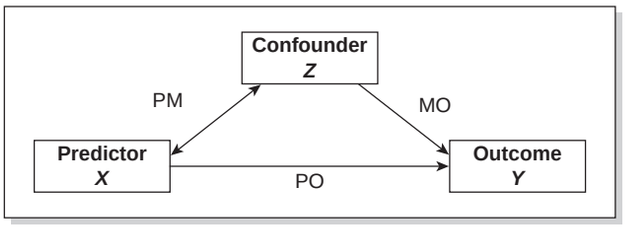


Figure 1 Conceptual model of confounding variable

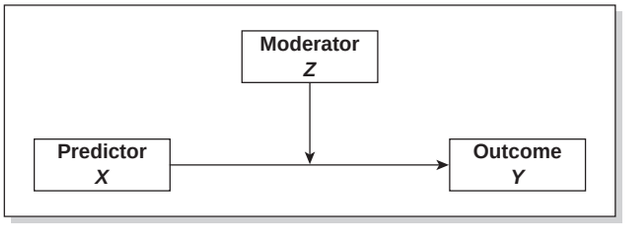


Figure 2 Conceptual model of a moderating variable

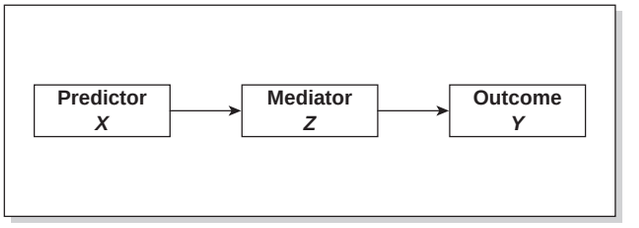


Figure 3 Conceptual model of a moderating variable

Confounding Variable

Variable *Z* is called a *confounding variable* or *confounder* of the effect of *X* on *Y* if *Z* is associated with *Y*, varies over the levels of *X* (with relationship both ways), and is not considered to be a cause of *Y*. This definition requires evaluation of the change in the relationship of *Y* and *X* due to *Z* and a subjective judgment that *Z* does not lie on the causal pathway between *X* and *Y*. A confounding variable may also be called a *lurking variable*. Since the relationship between *X* and *Y* changes with the value of *Z*, uncorrected confounding can result in the effect of *X* on *Y* being inappropriately increased or diminished or even reversed in direction. Confounding may be controlled by matching, stratifying on values of *Z*, or including *Z* in the statistical model.

Moderating Variable or Effect Modifier

Variable Z is called a *moderating variable* or *effect modifier* when its magnitude affects the magnitude or direction of the effect of X on Y and the interaction term XZ is statistically significant.

Mediating Variable

Variable Z is called a *mediating variable* or *mediator* of the effect of X on Y if X significantly affects Z, Z has a significant effect on Y, X affects Y in the absence of Z, and the effect of X on Y is diminished to nonsignificance when Z is added to the model. Some authors also require that Z be considered to be a cause of Y. A mediator is the same as a confounder except for the subjective judgment that the mediator is considered a "cause" of the outcome whereas a confounder is not. A mediator is a variable that is in a causal sequence between two variables, whereas a moderator is not part of a causal sequence between the two variables. The extent to which a variable may be considered a mediator may be assessed statistically by the Sobel-Goodman test. This test requires coefficients estimated from a separate regression fit of the path PM, MO, and PO.

Examples

Example of Mediating Effect

Obesity is an increasingly prevalent public health concern due to the increased risk of mortality associated with excess body fat and the increased risk of developing a variety of diseases such as type 2 diabetes, coronary heart disease, sleep apnea, knee osteoarthritis, and certain cancers. Obesity also has a substantial negative impact on a person's functional capacity and health-related quality of life (QoL). Heo and colleagues attempted to understand to what extent the association between obesity and QoL is mediated by those health problems that often arise in conjunction with obesity such as diabetes, hypertension, and (musculoskeletal) joint pain. In their article "Obesity and Quality of Life: Mediating Effects of Pain and Comorbidities," they hypothesized potential mediating effects of pain and comorbidities on the association between obesity and QoL and tested their hypotheses using data on 154,074 participants from the cross-sectional survey data from the 1999 Behavioral Risk Factor Surveillance Survey (BRFSS).

The predictor variable of obesity was measured by the BMI. This was calculated from the self-reported weight and height and was classified in six categories (< 18.5 kg/m^2, underweight; 18.5 to 24.9 kg/m^2, desirable weight; 25 to 29.9 kg/m^2, overweight; 30 to 34.9 kg/m^2, Obesity Class I; 35 to 39.9 kg/m^2, Obesity Class II; and ≥ 40 kg/m^2, Obesity Class III). Although they considered four outcome variables, for keeping the illustration simple here, we consider only one outcome variable of TUD dichotomized at 14 days. Potential mediator variables of joint pain (PAIN) were derived from the question "During the past 12 months, have you had pain, aching, stiffness, or swelling in or around a joint?" (0 = *No*, 1 = *Yes*) and obesity-related comorbidities (ORCs) were derived from the sum of responses to the nine dichotomous variables arising from questions such as "Have you ever been told by a doctor, nurse, or other health professional that you have high blood pressure?" (0 = *No*, 1 = *Yes*). Covariates consisted of the following characteristics: age, sex, marital status (married vs. other), educational attainment ($<$ high school vs. $\geq$ high school), annual income ($<$ \$25,000 vs. $\geq$ \$25,000), smoking status (current, former, never), and employment status (employed vs. other).

Figure 4 shows the conceptualization of the statistical analysis. To estimate and test the significance of the association between BMI and TUD, the authors ran multiple logistic regressions on the BMI-defined categories on TUD (Path A of Scheme a in Figure 4). To examine mediator effects of PAIN and ORCs on the BMI-TUD association if this association is significant, they followed the guidelines suggested by Baron and Kenny. Specifically, they assessed whether or not (1) BMI effects on PAIN and ORCs (Path B of Scheme b in Figure 4) are significant; (2) the effects of PAIN and ORCs on TUD (Path C of Scheme b in Figure 4) are significant; and (3) the effects of BMI classes on TUD are reduced when Paths B and C (Figure 4) are controlled for, that is, when PAIN and ORCs are added into the model of Path A (Figure 4). If all these conditions are met, the data are consistent with the hypothesis that PAIN and ORCs mediate the relation between BMI and TUD, supporting Scheme b in Figure 4.

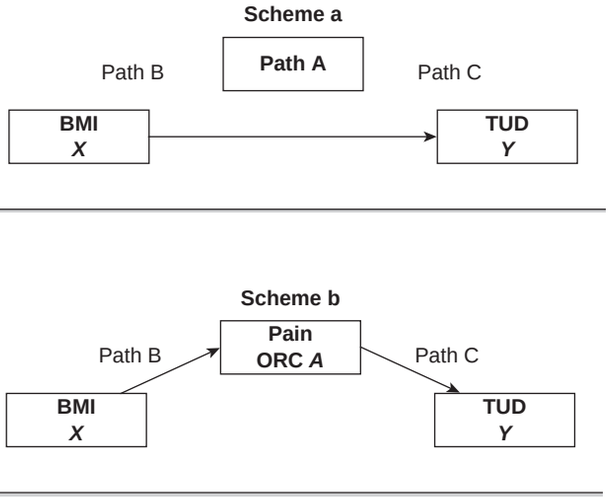


Figure 4 Schema of path diagrams of association between body mass index (BMI) and health-related quality-of-life (HRQOL) outcomes

Source: Heo, Allison, Faith, Zhu, and Fontaine (2003).

Collectively, from all the mediation analyses, the mediator effects of PAIN and ORCs on the relationship between high BMI and TUD are found significant. Moreover, controlling for the putative mediators resulted in nonsignificant effects of all BMI classes on TUD.

Example of Moderating Effect

Naidoo and Pretorius hypothesized that the stress-reducing function of helplessness (Z) has a moderating effect on the relationship between the rheumatoid arthritis (RA) health outcome of depression (Y) and the clinical measurement of the number of swollen joints (X) out of 28 joints in total. A cross-sectional study with 186 patients was undertaken for testing this moderating effect. The moderating variable of "helplessness" (Z) was measured by the Arthritis Helplessness Index (AHI). The AHI is a 15-item self-report inventory based on a 4-point Likert-type format that assesses the extent to which patients believe that they are able to control and cope with arthritis symptoms.

To test the hypotheses that Z moderates the relationship between X and Y, a regression model was fit with "depression" as outcome; swollen joint count (SJC) and helplessness as main effects; and an interaction term of SJC × Helplessness to test for moderation. Potentially confounding variables of age, sex, education, and income were also adjusted. Since the interaction term was found significant, the role of helplessness was established as a moderating variable. The patients with a "low" helplessness index were found to have decreasing depression even with an increasing number of swollen joint counts, whereas patients with a high helplessness index showed an opposite relationship (increasing depression with increasing number of swollen joint counts).

Causality

One of the fundamental goals of statistical design and analysis is to bring evidence based on data toward supporting causality. Understanding confounding and effect modulation are essential parts of getting as close as one can to causality, and separating the ideas of "confounding," "moderation," and "mediation" helps with use of the appropriate level of modeling.

Madhu Mazumdar and Ronald B. Harrist

Authors' note: Madhu Mazumdar was partially supported by the following grants: Center for Education and Research in Therapeutics (CERTs) (AHRQ RFA-HS-05-14), Clinical Translational Science Center (CTSC) (UL1-RR024996), and Collaborative Program in Nutrition and Cancer Prevention (NIGMS R25CA105012).

See also Causal Inference and Diagrams; Causal Inference in Medical Decision Making

Further Readings

Baron, R. M., & Kenny, D. A. (1986). The moderator-mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology, 51,* 1173–1182.

Bennett, J. A. (2000). Mediator and moderator variables in nursing research: Conceptual and statistical differences. *Research in Nursing and Health, 23*(5), 415–420.

Heo, M., Allison, D. B., Faith, M. S., Zhu, S., & Fontaine, K. R. (2003). Obesity and quality of life:

Mediating effects of pain and comorbidities. *Obesity Research, 11*(2), 209–216.
Hill, A. B. (1965). The environment and disease: Association or causation? *Proceedings of the Royal Society of Medicine, 58,* 295–300.
Judd, C. M., & Kenny, D. A. (1981). Process analysis: Estimating mediation in treatment evaluations. *Evaluation Review, 5,* 602–619.
Lipton, R., & Ødegaard, T. (2005). Causal thinking and causal language in epidemiology: It's in the details [Electronic version]. *Epidemiologic Perspectives & Innovations*, *2*(8).
MacKinnon, D. P., Fairchild, A. J., & Fritz, M. S. (2007). Mediation analysis. *Annual Review of Psychology, 58,* 593–614.
MacKinnon, D. P., Krull, J. L., & Lockwood, C. M. (2000). Equivalence of the mediation, confounding and suppression effect. *Prevention Science, 1*(4), 173–180.
Murray, D. M. (1998). *Design and analysis for group-randomized trials* (pp. 46–47). New York: Oxford University Press.
Naidoo, P., & Pretorius, T. B. (2006). The moderating role of helplessness in rheumatoid arthritis, a chronic disease. *Social Behavior and Personality, 34*(2), 103–112.
Pearl, J. (1998). *Why there is no statistical test for confounding, why many think there is, and why they are almost right*. UCLA Cognitive Systems Laboratory, Technical Report (R-256). Los Angeles: University of California.

Conjoint Analysis

Conjoint analysis (CA) is a quantitative technique used to elicit preferences. When faced with multiple alternatives, people often make decisions by making trade-offs between the specific features of competing products. CA derives preferences by examining these trade-offs through a series of rating, ranking, or choice tasks. Data generated from CA studies can then be used to determine which combination of features should be most preferred by each respondent.

CA was originally described by Luce and Tukey in 1964 and has since been widely used in market research, in economics, and most recently to examine preferences for competing programs, services, and treatment options in healthcare. This technique is based on three main assumptions. The first is that each product is a composite of different attributes and that each attribute is specified by a number of levels. For example, imagine that you are a researcher interested in eliciting patient preferences for competing pain medications. In this context, attributes might include specific medication characteristics such as route of administration, probability and magnitude of benefit, adverse effects, and out-of-pocket cost. The term *levels* refers to the range of estimates for each attribute. The levels for the attribute "out-of-pocket" costs for an insured population might range from $0 to $30.00 per month.

The second assumption underlying CA is that respondents have unique values, or utilities, for each attribute level. In this context *utility* is a number that represents the value a respondent associates with a particular characteristic, with higher utilities indicating increased value.

The final assumption underlying CA is that a subject's value for a specific product can be calculated by combining the discrete utilities associated with each attribute. Therefore, if the sum of a patient's utilities for the attributes of Medication A is greater than the sum of utilities for the attributes of Medication B, the patient should prefer Medication A to B.

Data generated from a CA study can answer important clinical questions, such as the following: Which attributes most strongly influence preferences? Which treatment is preferred and why? How much risk are patients willing to accept for a specified benefit? If cost is included as an attribute, CA can also estimate patients' willingness to pay.

Steps Involved in Performing a Conjoint Analysis Study

Step 1: Choose the Options, Attributes, and Levels

The investigator must first decide on the set of options to be evaluated. Is the objective to study preferences for all available treatment options for a particular condition or only those options appropriate for a particular subset of patients? Should hypothetical options representing potential future advances be included? If one is examining treatment preferences, are both pharmacologic and nonpharmacologic options to be included?

Once the set of options to be studied are identified, the investigator must choose which attributes and levels to include. Undoubtedly, this is the most difficult step in performing a CA study. Ideally, all the attributes required to choose between competing options should be included in the study. In some cases, the set of attributes is chosen based on data available from published studies. However, whenever possible, obtaining input from relevant stakeholders, via individual interviews or focus groups, is preferred. The estimates, or levels, for each attribute should be based on the best available evidence to date. With computerized programs it is possible to design separate versions of a survey to be able to present patients with individualized information.

Step 2: Choose a Conjoint Analysis Method

There are three main methods of conducting a CA study: full profile, choice-based, and adaptive (ACA, Sawtooth Software). These methods differ primarily in the way respondents are presented with information.

In full profile CA surveys, respondents are presented with complete profiles of hypothetical products that include a specified level for each attribute. Figure 1 describes two profiles from a hypothetical set of profiles examining preferences for pain medications.

Preferences are elicited by asking respondents to rate each profile or to rank a set of profiles. The main advantage of this technique is that it provides respondents with the most realistic descriptions of the products being evaluated. However, respondents tend to employ simplifying tactics to compensate for information overload when presented with full profiles using as few as four attributes, making this technique impractical for complex options.

	Drug 1	*Drug 6*
Route of administration	Cream	Pill
Probability of benefit	30%	80%
Risk of dyspepsia	10%	30%
Monthly cost ($)	$10	$20

Figure 1 Example of profiles used in full-profile conjoint analysis

Choice-based CA (CBC) is currently the most popular method of performing CA studies. As with the full profile approach, traditional CBC studies present respondents with profiles that include all attributes. Respondents are shown a choice set, usually composed of three or four profiles, and asked to indicate which they prefer. An example of a choice task evaluating treatment options for pain using the same attributes as those described above is provide in Figure 2.

CBC is preferred among many researchers because asking patients to perform a choice, rather than a rating or ranking task, is felt to be an easier task and more representative of how people make choices in the real world. In addition, CBC allows the investigator to include a "None" option—which enables respondents to refuse or defer.

ACA (Sawtooth Software, Inc., Sequim, WA) collects and analyzes preference data using an interactive computer program. This method is unique in that it uses individual respondents' answers to update and refine the questionnaire through a series of graded paired comparisons. Because it is interactive, ACA is more efficient than other techniques and allows a large number of attributes to be evaluated without resulting in information overload or respondent fatigue. This is an important advantage, since complex treatment decisions often require multiple trade-offs between competing risks and benefits. ACA surveys begin with a self-explicated set of questions that are followed by a set of paired comparison tasks. Figure 3 provides an example of the latter.

Step 3: Formulate an Experimental Design

The next step in developing a CA survey is to formulate an experimental design to decrease the number of scenarios each respondent evaluates. Imagine a very simple survey evaluating three attributes each having two levels. This small set of attributes and levels yields $2 \times 2 \times 2 = 8$ possible combinations. Increasing the number of levels by only one would yield $3 \times 3 \times 3 = 27$ possible combinations. Since most surveys include more than three attributes, experimental designs are required to identify an efficient subset of the total possible combinations of profiles to enable respondents to evaluate a practical number of scenarios. Fractional-factorial designs

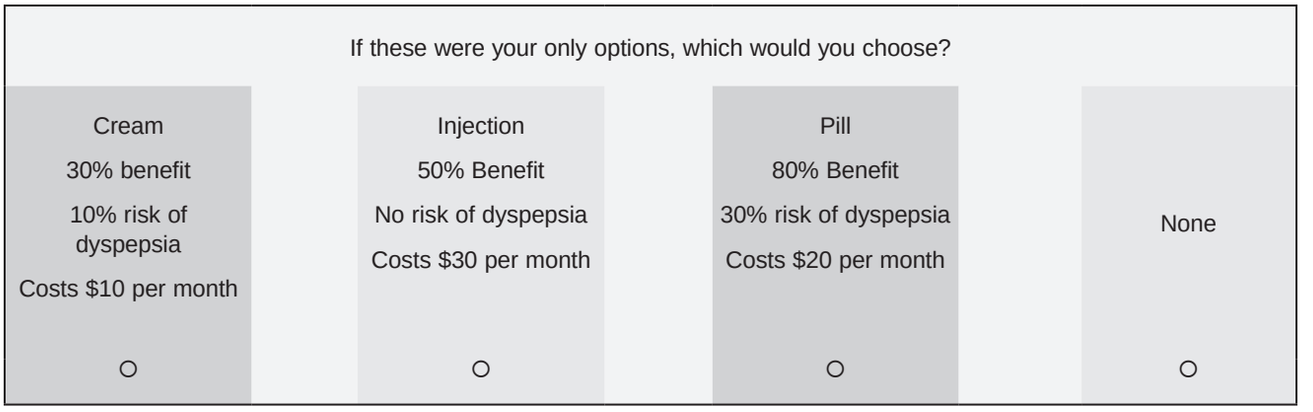


Figure 2 Example of a choice-based conjoint analysis choice task

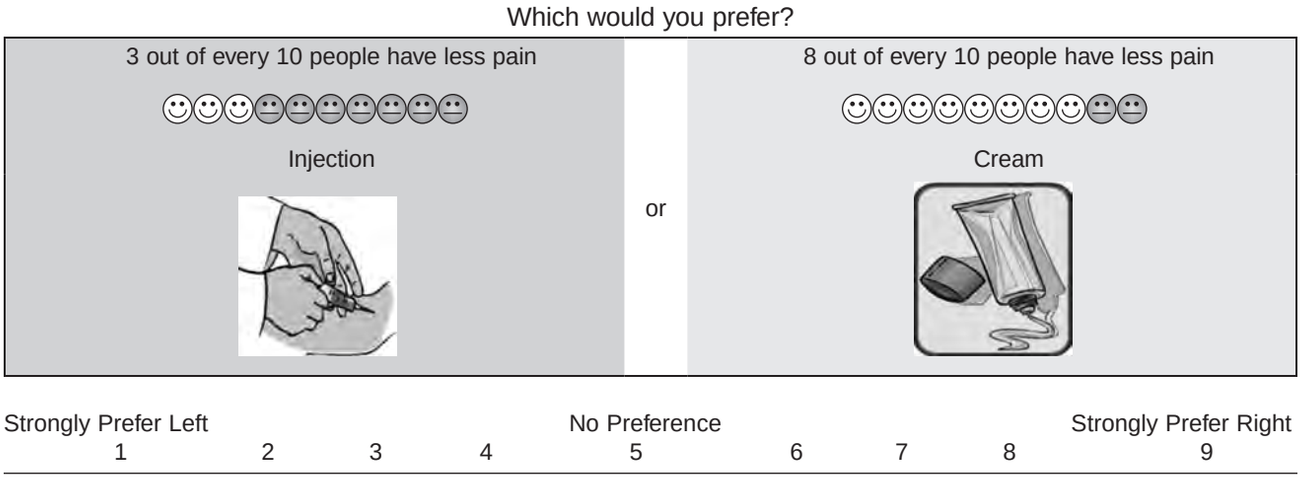


Figure 3 Example of an ACA paired-comparison task

can be generated using software programs, including SAS and Excel. There are also specialized CA software packages such as Sawtooth Software, Inc.

Step 4: Interpreting Conjoint Analysis Data

CA studies generate a utility or part-worth value for each level of each attribute. Part-worths can be calculated using several approaches. One of the most commonly used models is ordinary least squares regression. Recent advances include calculation of part-worths using Hierarchical Bayes Estimation. A full discussion of the models underlying CA is beyond the scope of this entry, however.

CA part-worths are scaled to an arbitrary constant within each attribute and are interval data. A set of hypothetical data are provided in Figure 4. In this example, the constant is the least preferred level of each attribute and is assigned a value of 0.

The significance of the part-worths is found within relative differences between the levels. Because the zero point is arbitrarily set, the absolute value of any specific level has no meaning. Therefore one cannot state that the utility or part-worth assigned to a 1% risk of dyspepsia is the same as that assigned to a $10 monthly cost. Nor can one state that an 80% chance of benefit is three times better than a 50% benefit. However, one can conclude that this respondent prefers pills over creams and that injections are least preferred.

Attribute	*Level*	*Part-Worth*
1. Route of administration	Cream	30
	Pill	40
	Injection	0
2. Probability of benefit	30%	0
	50%	30
	80%	90
3. Risk of dyspepsia	1%	20
	10%	5
	25%	0
4. Monthly cost	$5	30
	$10	20
	$30	0

Figure 4 Hypothetical part-worths generated by a conjoint analysis study

One can also conclude that for this respondent the value gained from changing a medication from a cream to a pill (10 additional utility units) is the same as that obtained by decreasing the monthly cost from $10 to $5.

CA surveys also allow the investigator to calculate the relative importance of each attribute. In this context, *relative importance* refers to the amount of importance respondents place on each treatment characteristic and is calculated by dividing the range of each characteristic (difference between levels) by the sum of ranges of all characteristics and multiplying by 100. These values sum to 100 and reflect the extent to which the difference between the levels of each characteristic affects each respondent's preferences. Relative importances are ratio measures and therefore support multiplicative functions. For example, based on the relative importances displayed in Figure 5, the respondent was influenced most by the probability of benefit and felt that route of administration was twice as important as the risk of dyspepsia. Of note, the relative importances are strongly influenced by the range of the levels chosen. For instance, in this example, one would expect cost to have a greater influence on preference if the maximum cost was $100 per month as opposed to $30.

In CBC studies, it is also possible to gain insight into respondents' preferences by counting the number of times each level was chosen. These data can be presented as proportions (with the denominator being the total number of times the level was presented in the survey). These proportions are ratio data and, unlike part-worths, can be compared within an attribute.

CA studies are most frequently used to predict preferences for available or hypothetical options defined by the researcher. For example, imagine that a researcher is interested in describing preferences for four treatment options for knee pain: capsaicin, acetaminophen, anti-inflammatory drugs, and cortisone injections. Using the attributes defined in Figure 4, the researcher defines each option by assigning an appropriate level to each attribute (see Figure 6).

Respondents' utilities are subsequently entered into a simulation model that yields a preference measure for each product. Sensitivity analyses to estimate the impact of changing specific characteristics on preference can also be performed. For example, using the example above, the investigator could examine how preferences for each of the four options are affected by changing cost, probability of benefit, or risk of toxicity.

Several simulator models are available, such as the first-choice and share-of-preference models. Each model uses different "rules" to estimate preferences. For example, in the first-choice model, the part-worths are summed and the respondent is assumed to choose the product with highest utility. In the share-of-preference model, preferences are calculated by first summing the utilities of the levels corresponding to each option. The utilities are then exponentiated and rescaled so that they sum to 100.

Previous studies of patient treatment preferences have (1) documented significant variability in treatment preferences, (2) found that patient preferences are frequently not aligned with treatment guidelines, and (3) shown that patient preferences may not be concordant with common medical practices. These findings each emphasize the importance of

Attribute	*Level*	*Part-Worth*	*Range*	*Relative Importance*
Route of administration	Cream	30	40	40/180 × 100 = 22
	Pill	40		
	Injection	0		
Probability of benefit	30%	0	90	90/180 × 100 = 50
	50%	30		
	80%	90		
Risk of dyspepsia	0%	20	20	20/180 × 100 = 11
	10%	5		
	25%	0		
Monthly cost	$5	30	30	30/180 × 100 = 17
	$10	20		
	$30	0		

Figure 5 Example of relative importances generated by a conjoint analysis study

Option	*Attribute 1 (Route)*	*Attribute 2 (Benefit)*	*Attribute 3 (Dyspepsia)*	*Attribute 4 (Cost)*
Option 1 (Capsaicin)	Level 1	Level 1	Level 1	Level 2
Option 2 (Acetaminophen)	Level 2	Level 1	Level 2	Level 1
Option 3 (Anti-inflammatory)	Level 2	Level 2	Level 3	Level 2
Option 4 (Cortisone injection)	Level 3	Level 2	Level 1	Level 3

Figure 6 Modeling preferences for knee pain

incorporating individual patient preferences into the medical decision-making process.

Important Features

CA has been used increasingly frequently to describe patient preferences for health-related services and treatment options. CA is also a means by which patients' views can be included in setting research priorities, designing trials, and developing policy.

CA has many properties that make it a valuable tool to elicit patient preferences and facilitate medical decision making:

- It can be designed to ensure that patients are made aware of all essential information related to appropriate treatment options and therefore should improve patient knowledge and informed consent.
- It improves the quality of decisions by making the trade-offs between competing options explicit. This is of direct clinical relevance since choices based on explicit trade-offs are less likely to be influenced by heuristics (errors in reasoning), which can lead to poor decisions.
- CA can be used to examine the amount of importance respondents place on specific treatment characteristics. This feature should enable physicians to gain insight into the reasons underlying their patients' preferences, tailor discussions to address individual patients' concerns, and ensure that decisions are made based on accurate expectations.
- It provides simulation capability. This feature allows the investigator to assess the impact of varying specific treatment characteristics on choice. For example, researchers can determine how much benefit patients require before accepting the risk of drug toxicity, whether decreasing the burden or inconveniences of therapy might increase patient acceptance of treatment, or which treatment option fits best with an individual patient's values.

Future Research

A reasonable body of evidence has now shown that CA is a feasible and valuable method of eliciting preferences in healthcare. Future research is now needed to determine if CA can be implemented as a decision support tool to improve informed decision making in medicine at the population as well as the individual patient level.

Liana Fraenkel

See also Utility Assessment Techniques

Further Readings

Kuhfeld, W. H. (2005). *Marketing research methods in SAS: Experimental design, choice, conjoint, and graphical techniques.* Cary, NC: SAS Institute. Retrieved May 15, 2008, from http://support.sas.com/techsup/technote/ts722title.pdf

Luce, D., & Tukey, J. (1964). Simultaneous conjoint measurement: A new type of fundamental measurement. *Journal of Mathematical Psychology, 1,* 1–27.

Orme, B. (n.d.). Hierarchical Bayes regression analysis: Technical paper. *Technical Paper Series*. Retrieved February 13, 2009, from http://www.sawtoothsoftware.com

Sawtooth Software Technical Papers Library: http://www.sawtoothsoftware.com/education/techpap.shtml

Srinivasan, V., & Park, C. S. (1997). Surprising robustness of the self-explicated approach to customer preference structure measurement. *Journal of Marketing Research, 34,* 286–291.

Wright, P. (1975). Consumer choice strategies: Simplifying vs. optimizing. *Journal of Marketing Research, 12,* 60–67.

Conjunction Probability Error

The conjunction rule applies to predictive judgment or forward conditional reasoning. It is a normative rule that states that the probability of any combination of events cannot exceed the probability of constituent events. For example, the probability of picking the queen of spades from a card deck cannot exceed the probability of picking a spade and a queen from the deck. Typically, people can successfully apply the conjunction rule to transparent problems such as the card selection problem. However, there is overwhelming evidence that when problems are less transparent, people often ignore

the rule and judge the conjunction of events as more probable than a constituent event, thereby committing the conjunction probability error. Because of the pervasiveness of the conjunction error and its clear violation of normative probability theory, it is important to understand conditions that tend to produce the error, procedures that may reduce its occurrence, and instances where it does not apply.

Conditions That Produce the Conjunction Error

The initial investigation of the conjunction error was conducted within the framework of understanding how heuristic thought processes may produce systematic biases in judgment and choice. In their seminal investigation, Amos Tversky and Daniel Kahneman first explored the conjunction error as resulting from the use of the representativeness heuristic for judging probabilities. According to this heuristic, people judge probabilities for specific outcomes by making a similarity comparison with a model of the population from which the outcomes were sampled. For example, knowing that a person is a member of a particular group, one may use a stereotype of that group as a model to predict behaviors or attributes of the person.

The Linda Problem

An often used example that has been shown to produce robust conjunction errors is the *Linda problem*. As described by Tversky and Kahneman, Linda is 31 years old, single, outspoken, and intelligent. Participants are told that when she was a philosophy major at school, she was concerned with social justice and participated in protests and demonstrations. This background establishes a model of Linda as a sophisticated individual concerned with social issues. After reading the description, participants typically rank the relative likelihoods of predicted occupations and activities that apply to Linda. Three key statements that may be evaluated include the following:

(U) Linda is a bank teller.

(L) Linda is active in the feminist movement.

(U & L) Linda is a bank teller and is active in the feminist movement.

The first statement is unlikely (*U*) based on the model of Linda and is given a relatively low probability ranking. The second statement is likely (*L*) based on the model of Linda and is given a relatively high probability ranking. The third statement is the key statement as it conjoins the unlikely and likely events (*U & L*). As such, it represents a subset of both these events and cannot have a higher probability than either of these. Yet nearly all participants indicate that the conjunction is more probable than the unlikely event. These results are obtained with both statistically naive and statistically sophisticated participants and in situations in which participants are directly assessing the relative likelihoods of the events. Furthermore, a majority of participants still commit the error even when they are asked to bet on these outcomes, implying the effect does not disappear with monetary incentives for correct application of the conjunction rule.

The conjunction error in the Linda problem is constructed by pairing an unlikely outcome from the model with a likely outcome from the model. In the probability calculus, the probability of the combined events can be expressed as follows:

$$\Pr(U \,\&\, L) = \Pr(U)\,\Pr(L|U).$$

This formula makes it explicit that the probability of Linda being a bank teller and active in the feminist movement, $\Pr(U \,\&\, L)$, must be less than or equal to the probability of her being a bank teller, $\Pr(U)$, as the probability of being active in the feminist movement given she is a bank teller, $\Pr(L|U)$, must be less than or equal to 1.0. But according to similarity-based heuristic thinking, combining an outcome that is dissimilar to the model with one that is similar to the model results in an evaluation of moderate similarity for the combined events. If probabilities are then based on such similarity evaluations, the mixed outcome case is judged as more probable than the unlikely constituent outcome, resulting in the conjunction error.

A Second Recipe for Conjunction Errors

The Linda problem used the recipe of combining an unlikely outcome with a likely outcome to produce the conjunction error. A second recipe for creating conjunction errors is to add an outcome that makes the other outcome more likely

or plausible. Tversky and Kahneman illustrated this recipe in a health-related example in which participants were told that a health survey had been administered to a large sample of adult males of all ages and occupations. They were then asked to indicate which statement was more likely of a randomly selected person from the survey:

1. This person has had one or more heart attacks.
2. This person has had one or more heart attacks and is over 55 years of age.

The majority of respondents chose the conjunction to be more probable in this instance. The specified age makes it easier for people to imagine this person having had one or more heart attacks. More generally, this type of conjunction error may be attributed to scenario thinking. In the first case, there is no reason to think that the selected individual might have had a heart attack. In the second case, the age-related information fills in some of the causal linkages that make the scenario more plausible and hence seem more probable. This type of scenario-based conjunction error can occur whenever a conjoined outcome provides a causal mechanism for the occurrence of the other outcome.

Application to Medical Decision Making

Tversky and Kahneman also demonstrated the applicability of the conjunction error directly to medical decision making. One of the problems they administered to two different groups of internists indicated that "A 55-year-old woman had pulmonary embolism documented angiographically 10 days after a cholecystectomy." The doctors were asked to rank order the probability that the patient would be experiencing each of a set of conditions. These included "dyspnea and hemiparesis" and "hemiparesis." Across the two samples, 91% indicated that the conjunction of conditions was more likely than the constituent condition. When physicians in an additional sample were confronted with their conjunction errors, they did not try to defend their decisions but simply indicated their surprise and dismay at having made such elementary errors. This last result suggests that the conjunction error is not simply due to misunderstanding how the alternatives are presented in the problem but instead represents a serious threat to risk assessment that can take place with experts within their own domain of expertise.

Procedures That May Reduce the Conjunction Error

Several criticisms of the work on the conjunction effect have been leveled over the 25 years since it was first reported. These criticisms focus on various features of how the problems are presented. One class of criticisms suggests that the problems may be ambiguously stated so that errors are due to participants misunderstanding what the experimenter is trying to communicate. For example, in several versions of the Linda problem, one simply chooses which is more probable, that "Linda is a bank teller" or that "Linda is a bank teller and is active in the feminist movement." One might argue that the pragmatics of conversation norms lead individuals to interpret the first statement as meaning "Linda is a bank teller and is not active in the feminist movement." Through the years, numerous ways of clarifying the options have been explored. The bottom line, however, is that although some versions may lead to fewer conjunction errors, they generally do not eliminate conjunction errors (i.e., the majority of participants still commit the error even with the reworded statements).

Another criticism has been directed against the normative force of the conjunction error. This argument is based on a strict frequentistic interpretation of probability, which states that it is reasonable to judge probabilities for samples from a population but it is not reasonable to judge probbilities for propensities of unique events. In the Linda problem, either she is or she is not a bank teller, and hence probability is not applicable. Rather than resolve this interpretation at the normative level, researchers have probed whether the conjunction error occurs when people are evaluating probabilities of samples from a population. For example, we can conceive of 100 women fitting Linda's description and estimate the probability that a random sample from this population would have these characteristics. Although some studies have shown a marked reduction of conjunction errors in this case, most have demonstrated very strong conjunction errors still occur. The health survey example discussed above is one case in point.

Related to the issue of interpreting probabilities is the assertion that probabilities are not a natural way of processing frequency information and so people will make errors when forced to consider probabilities rather than frequencies. Several researchers have tested this idea by comparing performance on problems requiring probability assessments versus frequency assessments. Note that the frequency assessment requires that one talk about sampling from a population rather than talk about propensities of individuals. The health survey problem described above has been formulated in frequency terms by asking participants to estimate how many of a sample of 100 individuals from the survey would fit each description. In general, the response format of estimating frequencies sampled from a large population has led to a significant reduction of conjunction errors, with the majority of participants not committing the error. This method would then appear to be a good way to reduce reasoning errors and de-bias judges.

However, a closer look at the pattern of results across numerous studies indicates that it is not the frequency format itself that is strongly reducing conjunction errors; rather, it is the requirement of making estimates that is critical. Numerous studies have shown that choosing which alternative would result in the highest sampled frequency does little to reduce conjunction errors. It is only when estimates must be generated for each option that conjunction errors are dramatically reduced. This occurs even when the estimates are of probabilities rather than of frequencies. This result supports the idea that people have at least two distinct ways to process probability information. One may be more qualitative and heuristic-based and the other more numerical and algorithmic. When the response mode is qualitative in nature, as in ranking and choice, people tend to apply the qualitative heuristic mode of thought and commit conjunction errors. When the response mode requires numerical assessments, people are more inclined to apply the quantitative algorithmic approaches and hence reduce conjunction errors.

Applicability of the Conjunction Rule

It is important to note when the conjunction rule does and does not apply when considering the various tasks associated with assessing probabilities. The conjunction rule applies to predictive judgment or forward conditional reasoning. In this type of reasoning, events are conditioned on a premise represented as a hypothesized model or hypothesized sampling procedure. In the medical decision-making context, it applies to predicting symptoms given a disease or outcomes given a procedure. In these cases, one must be careful to consider whether probability assessments are being inappropriately increased by the consideration of a conjunct that makes a particular outcome easier to envision. It is important to avoid scenario thinking or similarity-based thinking in making these assessments.

The conjunction rule does not apply to diagnostic judgment or backward conditional reasoning. In this kind of reasoning, one is inferring the probability of a hypothesis based on an outcome or a conjunction of outcomes. In medical decision making, this is by far the more common type of assessment. Given a particular set of symptoms one must estimate the likelihood of a given disease as the cause. Here, Bayesian updating applies so that conjoining a diagnostic symptom with a nondiagnostic symptom should lead to an increase in the overall probability of the disease. One possibility is that people commit the conjunction error because they do not correctly differentiate between these two tasks and hence incorrectly apply diagnostic reasoning to a prediction task.

Douglas H. Wedell

See also Bayesian Evidence Synthesis; Biases in Human Prediction; Frequency Estimation; Heuristics; Probability Errors

Further Readings

Gigerenzer, G. (1996). On narrow norms and vague heuristics: A reply to Tversky and Kahneman. *Psychological Review, 103,* 592–596.

Hertwig, R., & Chase, V. M. (1998). Many reasons or just one: How response mode affects reasoning in the conjunction problem. *Thinking and Reasoning, 4,* 319–352.

Sloman, S. A., Over, D., Slovak, L., & Stibel, J. M. (2003). Frequency illusions and other fallacies. *Organizational Behavior and Human Decision Processes, 91,* 296–309.

Tversky, A., & Kahneman, D. (1983). Extensional versus intuitive reasoning: The conjunction fallacy in

probability judgment. *Psychological Review, 90,* 293–315.
Wedell, D. H., & Moro, R. (2008). Testing boundary conditions for the conjunction fallacy: Effects of response mode, conceptual focus and problem type. *Cognition, 107,* 105–136.

Constraint Theory

Diagnosis—the process by which examination and existing knowledge are used to establish the nature and circumstances of a particular condition—has always been the basis of the practice of medicine. It is the focal point of the doctor-patient relationship. But given rapid advancements in medical science and technology over the past 50 or so years, diagnostic methods have emerged elsewhere across the healthcare system: in all manner of relationships, decision-making processes, management structures, and work in general. Individuals and organizations have, as a consequence, sought new ways to manage these transformations. One such way has been through the application of constraint theory, through which one can organize existing thoughts about complex systems and communicate them through scientific algorithms.

Information Is the Transformation

The basic features, challenges, and opportunities of today's healthcare system are not all-too-different from those of the past. Since the early days of the American republic, there has been plenty of thought about how to integrate a host of general, though interconnected needs related to patients, physicians and other practitioners, government, taxes, insurance, access to and quality of care, and business interests, to name a few.

Yet, in the interim, the vast improvements in science and technology that have been applied to the practice of medicine have produced an increasing amount of data. This has both benefited and swamped the system.

The amount of data available to people throughout the healthcare system, and the capacity to process it, is ever increasing. But for the data to be useful, it must be converted into information; for this to happen, the data must be oriented toward a particular purpose or given some relevance. As medicine becomes more specialized and healthcare more complex, the importance of information and how it flows—whether information is good or not, how people choose to share it, and whether it passes easily between people—makes a difference in the performance of people and organizations throughout the system. In response, a number of models have been developed in an attempt to effectively process data, convert it into information, and capture the flow of that information so that it can be used to make the right decisions.

Quality Improvement Methods

Flawless performance and attention to detail are highly regarded qualities in the medical and healthcare professions. This reality, plus today's economic dimensions and the fact that organizations within healthcare have grown larger and more difficult to manage, has increased the demand for nonmedical professionals who have experience in management techniques that could be applied to improve organizational behavior and, thereby, patient care. High on the list of innovative concepts that have infiltrated the workings of the modern healthcare organization is the implementation of quality improvement programs. Adapted from the engineering and service-outcome approaches that are popular, especially in the business of manufacturing, these programs are highly disciplined, statistically driven methods by which to measure and eliminate any number of "defects" in a production process. In form and function, they require a high level of systematic predictability and a low tolerance for human fallibility. Such programs are intended to be a reliable means by which to use input and output data to achieve the goal of delivering to customers a product or service that satisfies their needs.

A good portion of these programs come out of the Total Quality Management (TQM) philosophy developed in the mid- to late 20th century by, among others, W. Edwards Deming, Joseph Juran, and Kaoru Ishikawa. The TQM movement generally considers that every person and all activities in an organization must be managed toward customer requirements for a product or service. To accomplish as much, especially over the long term, TQM programs begin with four basic assumptions

about quality, people, organizations, and management. These assumptions include the beliefs that (a) the production of quality products and services is preferred over compromising quality in an attempt to keep costs low; (b) employees care about the quality of their performance and will work to improve it so long as management pays attention to their ideas, provides them with the means necessary for improvement, and creates a positive work environment; (c) organizations are constituted of interdependent parts that must function as a system; and (d) senior management is responsible for the creation, organization, and direction of the overall system that leads to quality outcomes.

From there, the interventions intended to actually bring about change and improve quality must focus on work processes, analysis of variability and variation in those processes, systematic collection and analysis of data at precise points in the processes, and a commitment to learning and "continuous improvement." Regard for these factors permits the development of new and better methods for performing work, which in turn improves the quality of the product or service being worked on. In all, the outcome relies on giving meaning to an increasing number of variables that must be variously and appropriately integrated into decisions across and through the system.

One of today's acknowledged, though controversial, interventions for service outcome and quality improvement is the theory of constraints, developed by Eliyahu Goldratt. It is grounded in the notion of a "weakest link" in any complex system. That is, at any point in time, there is some phenomenon that limits the function of the system to move beyond its current capacity and closer to achieving its goal. In this cause-and-effect relationship, the phenomenon—the constraint—must be identified and the entire system managed accordingly if the system is to improve. Yet the theory of constraints should not be confused with constraint theory.

Theory of Complexity and Constraint Theory

In complex systems, while the prevailing conditions of a certain environment are stable and predictable, the actions and effects of the elements within it are not. Over the past four decades, with the rise of digital computation and data processing, mathematical proofs have been used to show that complex systems are determined by numberless internal and external factors. These factors are not necessarily statistically significant and, therefore, do not allow prediction in the classical sense. And it may be that one of the statistically insignificant factors turns out to be that which has the greatest impact on the entire system. There have lately emerged across the study of modern mathematics several theorems that clearly identify such factors, including the constraint theory cast by George J. Friedman.

The traditional way to achieve a complete, correct, and consistent method for managing a complex system has been to divide a specific model into submodels that could be refined by specialists and later connected into an aggregate model. But Friedman's contention is that there is no assurance of consistency in the aggregate model even if there is consistency in every submodel. Through the development of constraint theory, he has shown that model structure can be used by cross-functional teams, analysts, and managers to discern inconsistencies in an aggregate model.

Friedman's constraint theory uses the "Four-Fold Way," which is a progressive collection of "views": set theoretic; family of submodels; bipartite graph; and constraint matrix. It separates a given model from computations and chronicles the existence and flow of constraints throughout the model. Each constraint may be tagged and valued as an *overconstraint*—an instance in which more variables exist than the number required for solving a group of equations—or an *underconstraint,* in which fewer variables exist than the amount required for solving a group of equations. The entire operation is typically represented on a bipartite graph (see Figure 1), with a nodes vertex that denotes the relations within the model and a knots vertex that signifies its variables; nodes are represented by squares and knots by circles, and a knot will be connected to a node by an edge if and only if the corresponding variable is present in the corresponding functional relationship of a model. In effect, the nodes are central points, and the knots are points at which the values of the variables pass from one central point to another. The end result is that visualization of the system, before returning

to the original group of model equations, can benefit the development of a strategy that ideally would lead to some solution.

The essence of constraint theory is that it enhances the use of computer assistance to bring some level of control to numberless variables in a system. It intends to identify decisive factors, yet does not, as a rule, convey how to eliminate extraneous ones. This, in any case, helps one more accurately analyze the behaviors and performance of the people and forces within the system—and at its various stages, under its various criteria, and with respect to its various needs of integration and design. But for anyone to become proficient—or at least—in the technique of constraint theory requires that one first comprehend the basic concepts of set theory and graph theory, which is more often the domain of mathematicians and engineers than of physicians and healthcare professionals. That is, while they could well have the capacity and knowledge to grasp the particulars of constraint theory, it is more likely that physicians and other healthcare professionals' time, contributions, strengths, and priorities are better invested in the tasks and practices specific to their work.

There is no question that new realities during the mid- to late 20th century—primarily, advances in medical science and technology and the advent of managed care—have necessitated new applications of new knowledge. Nor is there doubt that with such transformations there is a need to incorporate every relative complexity—however overt or subtle, close or remote, old or new—into capable analysis. Being able to understand and act on constraints at a given point in a system is especially imperative today as the healthcare system increasingly relies on the use of knowledge and learning as a basis for skills that allow its highly specialized, productive work to be performed. It therefore needs programs that demand that decision makers be precise in every decision-making capacity, that they know precisely what to do with the available information, and that they are listening to and asking questions that encourage critical thinking and the careful development of ideas. Only then can medical and healthcare professionals tend to the care for and cure of the sick patient.

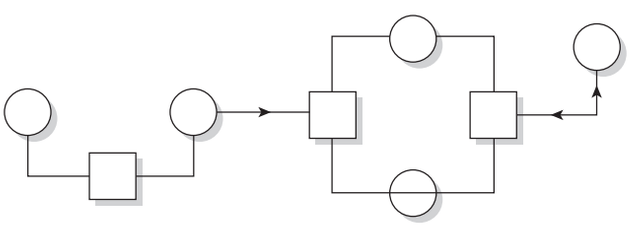


Figure 1 Bipartite graph: A nonspecific, simple example (see Friedman, 2005, for specific and complex examples)

Lee H. Igel

See also Chaos Theory; Complexity; Computer-Assisted Decision Making

Further Readings

Deming, W. E. (2000). *Out of the crisis*. Cambridge: MIT Press.

Drucker, P. F. (2003). *The new realities*. New Brunswick, NJ: Transaction.

Friedman, G. J. (1976). Constraint theory: An overview. *International Journal of Systems Science, 7*(10), 1113–1151.

Friedman, G. J. (2005). *Constraint theory: Multidimensional mathematical model management*. New York: Springer.

Goldratt, E. M. (1999). *Theory of constraints*. Great Barrington, MA: North River Press.

Goldratt, E. M., & Cox, J. (2004). *The goal: A process of ongoing improvement* (3rd ed.). Great Barrington, MA: North River Press.

Hackman, J. R., & Wageman, R. (1995). Total quality management: Empirical, conceptual, and practical issues. *Administrative Science Quarterly, 40*(2), 309–342.

Ishikawa, K. (1991). *What is total quality control? The Japanese way*. Englewood Cliffs, NJ: Prentice Hall.

Juran, J. M. (1995). *Managerial breakthrough: A new concept of the manager's job* (Rev. ed.). New York: McGraw-Hill.

Warfield, J. N. (2003). A proposal for systems science. *Systems Research and Behavioral Science, 20*(6), 507–520.

Construction of Values

Construction of values refers to the process whereby an individual's preference for a particular health state or more generally, a "good" of any

sort, is developed or built (constructed) at the time when that state or good is encountered, either actually or hypothetically. Preference in this context refers to the desirability or undesirability of something, from the subjective perspective of the person assessing or evaluating it. Preferences are the external manifestation of underlying values, and are commonly referred to interchangeably. Value construction occurs in clinical contexts when medical decisions are imminent or in forecasting decision making. In research settings, value construction occurs usually in the consideration of hypothetical choices, such as in preference elicitation surveys or choice experiments. Value construction can be contrasted with value retrieval, in which values already exist and are known to the individual, and are simply retrieved from memory.

Definition

Preferences about goods, health states, or even issues are thought to exist on a continuum, from those that are basic to those that are highly complex. Basic values are easily known and expressed by an individual; these values are quite possibly innate. Complex values require extensive cognitive effort to understand and express and may not be immediately available to an individual. For example, an infant's preference for its mother over another person could be considered a basic, innate preference or value. Similarly, the value one person places on Pepsi versus Coke is basic and known and easy to express. At the other extreme, the value placed on a painful and debilitating yet life-extending therapy might not be known to an individual without the benefit of extensive thought, consideration, and deliberation. The value for this therapy is based on more basic values but is some combination of many considerations and preferences, including trade-offs among conflicting values, resulting in a complex preference. The construction of values refers to this latter process in which an individual uses information and more basic values to construct, or build, the more complex value.

Construction Process

Values are constructed at the time when an individual is faced with a situation that demands knowledge or expression of his or her values. In general, values are called on every time an individual makes a choice or decision, from purchasing one brand versus another to casting a ballot. In the context of health and medicine, values are usually called on when an individual is faced with a decision about a medical intervention or treatment, from something as simple as receiving a flu shot to consenting to surgery. Values are also invoked during surveys and experiments asking about choices and decisions, wherein much of our knowledge about preference construction has been demonstrated.

The construction process generally begins when an individual is faced with a choice or decision that defies basic values. For example, if a person is asked which political party he or she supports, he or she may reply "Democrat" or "Republican." If a person is asked whether she supports Candidate A or Candidate B, she may ask about the candidates' positions on an issue important to her, such as environmental protection. On learning of the candidates' positions, she will choose A or B. If then she is told Candidate A is female and Candidate B is male, and this person prefers to support a female candidate, she will have to consider both the candidates' genders and their positions on environmental protection to make a choice. If the male candidate is a stronger proponent of environmental protection, the person has to weigh the importance of her gender preference against her environmental protection preference to arrive at a decision. This type of choice would be considered to invoke complex values because it is not readily apparent what choice would be dictated from the basic values regarding gender and environmental protection. Values for the candidates hence would be *constructed* from the information and basic values. Value construction occurs when basic values would suffice but information is unknown, when basic values do not exist for the options encountered, or when the complexity of the choice involves combinations of or trade-offs among basic values.

Value construction in medical decision making often involves multiple and conflicting trade-offs, and information is often lacking. Basic values regarding medical decisions can be well-known and accessible, such as the value placed on quality of life or longevity. Yet these values are often

encountered in contradiction, making complexity inherent in most medical choices.

Elicitation Process

While value construction occurs implicitly when choices are made, the process becomes more explicit when values or preferences are specifically elicited for decision making or in the context of surveys. Because complex values are based on basic values and potentially information, time and consideration are necessary components of the construction process, though it can be highly person and situation specific. Lacking any of these elements, complex values may be expressed inaccurately as their more basic components, or as entities entirely different from those that would be articulated given sufficient information and consideration. Such occurrences have been demonstrated as preference reversals, which are basically situations in which a person directly contradicts himself or herself in matched choices, or as framing effects, in which the context in which a choice is presented unduly influences the outcome. Such incidents are not adequately described by the theories that underlie decision making and in this context indicate the need for value construction to maximize expression of true preferences, unaffected by context and other external factors. Value elicitation processes should therefore take the necessary elements of value construction into account to produce valid and stable expressions of complex values.

Importance in Medical Decision Making

Acknowledging that values are constructed implies that a process should be followed when decisions are made. Since medical decisions commonly involve complex values and multiple trade-offs, the elements necessary for value construction should be provided to ensure fully informed and formed choices. Complete information and thorough consideration may enable a construction process that leads to decisions that accurately reflect underlying basic values. Understanding of the processes that motivate value formation can provide guidance in eliciting and articulating quality decision making in health and medicine.

Eve Wittenberg

See also Decision Quality; Preference Reversals; Utility Assessment Techniques

Further Readings

Gregory, R., Lichtenstein, S., & Slovic, P. (1993). Valuing environmental resources: A constructive approach. *Journal of Risk and Uncertainty, 7,* 177–197.

Lichtenstein, S., & Slovic, P. (Eds.). (2006). *The construction of preference.* New York: Cambridge University Press.

Payne, J. W., Bettman, J. R., & Schkade, D. A. (1999). Measuring constructed preferences: Towards a building code. *Journal of Risk and Uncertainty, 19,* 1–3, 243–270.

Consumer-Directed Health Plans

Consumer-directed healthcare is an approach to financing healthcare services wherein individuals are given fixed allowances with which they can purchase specified services. Most plans couple this with catastrophic coverage, usually with a high deductible. These plans often receive tax advantages. Variants may be termed *medical savings accounts, health savings accounts,* or *flexible spending accounts.* They have been employed in a number of countries, including the United States, Singapore, South Africa, and China. There are differences in terms of such details as who contributes (employer, employee, or both), the levels of deductibles and co-payments payable, which services can be purchased with these funds, and whether unused contributions can be carried over to subsequent years. Some models employ a "use it or lose it" approach, whereas others allow savings to be accumulated, often tax-free. Consumer-directed models are predicated on the assumption that potential users of care should be the ones making the decisions about what care to receive and from whom.

The Case For

Market Approaches to Allocation

Consumer-based models are based on the premise that, like other commodities, healthcare is a

market good and as such its utilization is subject to the predictions of economic theory. In economics, price is the signal that ensures a balance between supply and demand. Economic theorists would thus predict that reducing price would increase demand. In addition, they note that insurance may create what is termed *moral hazard*, a term referring to the prospect that insulating people from risk (a major purpose of insurance) may make them less concerned about the potential negative consequences of that risk than they otherwise might be. For example, those with flood insurance may be more willing to build in flood plains, in the confidence that insurance would cover their losses. Similarly, economic theory would predict that those with health insurance, because they do not have to pay the full cost of any care they receive, would have an incentive to over-use it. Advocates thus argue that consumer-driven models are the best way to achieve cost control because wise consumers will shop around for the best buy, measured in terms of both quality and price. They suggest that an additional benefit of high deductible plans is that insurers will save money by not having to process small claims.

In contrast, other theorists argue that utilization of health services differs from purchases of consumer goods in that it is (or at least should be) based on need rather than demand. Because need is defined by experts rather than by consumers, they further argue that those individuals receiving care are not always in the best position to make treatment decisions, for a number of reasons, including "asymmetric information."

Who Is the Decision Maker?

Consumer-directed models are often presented as an alternative to managed care, which is described as representing control by technocrats, who inhibit innovation, and instead attempt to control costs with "just say no" policies, to the detriment of both patients and providers. Others note that they also represent a rejection of agency models, whereby expert providers are expected to determine what care their patients need, in favor of models wherein the recipients of care act as the decision makers about both what care to purchase and from which providers. Consumer-directed plans are therefore justified as empowering users of services and being linked to informed decision making. Discussion of these models is thus often associated with language speaking of patient empowerment and of putting patients in control.

Who Pays for What?

In contrast to approaches that pool risks and guarantee coverage for "necessary" services, consumer-directed models try to minimize the extent of cross-subsidization. Consumer-directed care is accordingly associated with a major shift of costs from insurers to consumers, in the form of high deductibles and co-payments; this shift is justified as necessary to make individuals act as informed consumers. In this model, insurance is reserved for catastrophic costs, with the more predictable costs expected to be covered through personal savings. To encourage that transition, governments may define minimum or maximum levels for deductibles and give preferable taxation treatment to the savings account components. Plans may also extend the range of insured benefits if they allow savings to be used for services not traditionally covered by insurance.

The Case Against

Opponents contest most of the aforementioned assumptions.

Impact of Cost Sharing on Utilization

One data source, referred to by both sides of the debate, is the RAND Health Insurance Experiment (HIE), a randomized experiment of various cost-sharing arrangements conducted between 1971 and 1982. The researchers found that cost sharing reduced the use of nearly all health services among study participants (which excluded the elderly and many of those with preexisting serious health conditions). Extrapolating these findings, proponents argue that consumer-directed care will reduce costs and increase efficiency. However, as the RAND group has itself pointed out, this reduced use of services resulted primarily from decisions not to seek out care. Once in the healthcare system, there were only modest effects on the cost of an episode of care. Cost sharing was equally likely to deter appropriate (and effective) care as to deter more

marginal (ineffective) visits. In general, the reduction in services did not lead to adverse health outcomes, at least in the short run. However, there were exceptions, particularly for the poorest patients. This evoked concerns that cost sharing might deter preventive and follow-up care and ultimately lead to higher costs and worse outcomes. The experiment did not find any discernible differences in the quality of care, or in how well people took care of themselves. Patient satisfaction tended to be lower in the plans with higher cost sharing. Extrapolating these findings, opponents worry that needed care will not be received. Advocates suggest that certain services (including some preventive care) can be exempted from cost sharing requirements.

Availability of Information for Decision Making

A related set of arguments stresses agency relationships, and the difficulty of individuals attempting to be wise purchasers in areas requiring expertise. Some argue that, left to their own devices, individuals may delay receiving appropriate care. Others respond that this objection is paternalistic and can be overcome if good information is made available about costs and quality.

Adverse Selection

Another set of arguments relates to the highly skewed nature of health expenditures. As studies in the United States and Canada have confirmed, a very small proportion of individuals represent the bulk of health expenditures. The lowest spending 50% account for less than 5% of costs, and similar patterns apply within every age-sex category. Insurers have a strong incentive to avoid those individuals likely to generate high costs, a phenomenon referred to as adverse selection. Similarly, consumer-directed plans are likely to be most attractive to those with better health status. To the extent that risk pooling breaks down, these authors note that there is likely to be a negative impact on the sustainability of an insurance model, with the healthier benefiting from lower premiums, and the sick finding themselves uninsurable.

Choice

Another set of arguments relates to the meaning of patient choice. To the extent that market-based models assume enough excess capacity to react to increases in demand, choice may be illusory. This may apply where there are not multiple potential providers, including in rural/remote areas, and for certain highly specialized services. It may also apply when individuals do not have sufficient resources to purchase care.

Empirical Results

Consumer-directed plans are relatively recent, and evaluation is therefore limited. The international evidence is mixed, with growing suggestions that they create gaps in access. In the United States, they represent a small proportion of insured individuals (about 3%) but are growing rapidly. The literature suggests a mixed picture. The empirical evidence to date suggests that the bulk of the population—which tends to be healthier—may well reduce use without adverse health effects but that already vulnerable populations (by income, and by health status) may show worse results. Because costs are so highly skewed, the overall savings are likely to be minimal and potentially offset by higher costs among those not receiving necessary care. Premiums are lower, which is to the advantage of those paying for coverage (employers or potential consumers). However, coverage is less, and out-of-pocket costs can be considerable; Bloche estimates that they can exceed $10,000 per year for families. To date, analysts have not yet found impacts on quality of care and have found that few individuals feel confident with the information available to them to date.

The Government Accountability Office Report

A 2006 review by the U.S. Government Accountability Office (GAO) surveyed early experiences with one kind of plan—Health Savings Accounts (HSAs). They found the following:

- The sorts of services covered were similar.
- Those enrolled were much more likely to have higher incomes (51% vs. 18% of all tax filers younger than age 65).
- Costs for enrollees were higher than for those enrolled in traditional (PPO) plans when extensive care was used but lower when use was low to moderate.

- Few participants researched costs before obtaining services; if consumerism were to increase, it "will likely require time, education, and improved decision support tools that provide enrollees with more information about the cost and quality of health care providers and services" (p. 30).
- "Most participants were satisfied with their HSA-eligible plan and would recommend these plans to healthy consumers but not to those who use maintenance medication, have a chronic condition, have children, or may not have the funds to meet the high deductible."

Outlook

Given ongoing problems with both access and cost control in the United States, consumer-directed health plans are likely to play a role. The extent to which they can fulfill their stated goals, however, remains unclear. More evidence is clearly needed, but to date the claims of advocates appear problematic, both in their assumptions about the nature of decision making in healthcare and about the differences between medical care and other consumer goods.

Raisa Deber

See also Decisions Faced by Patients: Primary Care; Patient Rights

Further Readings

Berk, M. L., & Monheit, A. C. (2001). The concentration of health care expenditures, revisited. *Health Affairs, 20*(2), 9–18.

Bloche, M. G. (2006). Consumer-directed health care. *New England Journal of Medicine, 355*(17), 1756–1759.

Davis, K., Doty, M. M., & Ho, A. (2005). *How high is too high? Implications of high-deductible health plans*. New York: Commonwealth Fund.

Deber, R., Forget, E., & Roos, L. (2004, October). Medical savings accounts in a universal system: Wishful thinking meets evidence. *Health Policy, 70*(1), 49–66.

Forget, E. L., Deber, R., & Roos, L. L. (2002). Medical savings accounts: Will they reduce costs? *Canadian Medical Association Journal, 167*(2), 143–147.

Herzlinger, R. E. (Ed.). (2004). *Consumer-driven health care: Implications for providers, payers, and policymakers*. San Francisco: Jossey-Bass.

Jost, T. S. (2007). *Health care at risk: A critique of the consumer-driven movement*. Durham, NC: Duke University Press.

Newhouse, J. P., & The Insurance Experiment Group. (1993). *Free for all? Lessons from the RAND health insurance experiment*. Cambridge, MA: Harvard University Press.

Robinson, J. C. (2005). Managed consumerism in health care. *Health Affairs, 24*(6), 1478–1489.

U.S. Government Accountability Office. (2006). *Consumer-directed health plans: Early enrollee experiences with health savings accounts and eligible health plans*. Report to the Ranking Minority Member, Committee on Finance, U.S. Senate, August.

Context Effects

Normative decision theory is often formulated to assume that decision makers have perfect information, a perfect grasp of their objectives, and the perfect ability to use that information to make uncertain decisions and further their objectives. It is common for psychologists to criticize the use of such strong assumptions as indefensible because they ignore the effects of important situational and contextual factors. In this respect, the term *context* can be defined in two distinct but conceptually related ways: (1) context as the presentation (description), or *framing*, of the decision problem, which determines how the task is conceptualized by the individual, and (2) context as the set of available choice options (e.g., in decision making under risk). Both types of context affect how the decision problem is cognitively represented by the agent, which in turn affects the outcome of the decision making process. Here, these two types of context effects are discussed separately.

Context Effects Caused by Task Framing

In these accounts, the term *context* refers to a set of facts describing a particular situation from a specific point of view. There is evidence that minor changes in the presentation or framing of risky choice problems can have dramatic impacts on choices. Such effects are failures of description invariance because different answers are elicited if decision problems are presented in different but

logically equivalent forms, or contexts. A famous example of framing effects is a study by Tversky and Kahneman, in which two groups were presented with an Asian disease story and their choice was between two probabilistically equivalent medical policies—one with a certain outcome and one with a risky outcome having higher potential gain. However, the description for the first group presented the information in terms of lives saved while the information presented to the second group was in terms of lives lost. There was a striking difference in responses to these two presentations: 72% of participants preferred the first policy when it was described as lives saved, while only 22% of participants preferred this option when it was in terms of lives lost. Such failures of description invariance appear to challenge the very idea that choices can, in general, be represented by any single preference function.

Prospect theory was proposed as a psychological account of such framing effects on behavior toward risk. In this theory, choices among prospects are determined by a preference function, in which outcomes are interpreted as gains and losses relative to a reference point (e.g., status quo wealth). Empirical estimates find that losses are weighted about twice as strongly as gains, that is, the utility function is steeper for losses than for gains, which means that the disutility of losing $100 is twice the utility of gaining $100. In the Asian disease problem, when outcomes were framed as lives saved, the majority of choosers were attracted to a sure gain of lives; when framed as losses the majority rejected the sure loss of deaths, which according to the loss function hurts much more, preferring instead to take the more risky policy.

Consistent with prospect theory, the rating of different health states varying in severity is influenced by the perspective of the rater (i.e., his or her own current health relative to the rated health conditions). For example, a mild lung disease scenario and a severe one are rated differently by lung disease patients, whereas healthy nonpatients rate the two scenarios as much more similar. Because patients and nonpatients have a different status quo reference point, they have different perceptions of the same health condition. For a patient suffering from a moderately severe lung disease, a milder case of the same disease would represent a gain in health generating a steep improvement in life quality, whereas a severe case of lung disease would represent a loss in health with a steep cost in quality. In contrast, for a healthy person, both mild and severe cases of lung disease would represent a loss in health.

A similar test of the validity of prospect theory in medical context showed that hospitalization causes a decline in patients' desire for very unpleasant life-sustaining treatment (i.e., individuals express different treatment preferences when they are healthy compared with when they are ill). Thus, direct experience with the discomforts of hospitalization changed patients' attitudes about the value of extending life via aggressive medical treatment. Therefore, the task of divining a patient's "true" end-of-life wishes becomes difficult because decisions to receive life-sustaining treatment stated by healthy individuals may be particularly susceptible to contextual change.

In summary, these recent studies are examples of expanding research questioning the stability of treatment preferences over time and across changes in an individual's health condition, and the general ability of individuals to predict accurately their future feelings and behavioral choices.

Context Effects Caused by the Choice Set

A number of decision experiments have investigated the effect of the context defined in terms of the set of available options. This research draws attention to a general and pervasive feature of human cognition, which is related to how people judge the magnitudes of attributes of choice options such as utilities, payoffs, and probabilities, which are essential ingredients of every decision problem. The basic question is whether there is a cognitive ability to represent absolute cardinal scales on any magnitude, and judgments involving such magnitudes are determined solely by the context. The research is based on evidence from psychophysics and perceptual judgment, which shows that people are not able to represent the absolute magnitudes of the attributes of any stimuli, for example, light, brightness, weight, loudness, happiness, satisfaction, and so on, and instead, they represent such magnitudes on a ordinal scale purely in relation to other magnitudes. For example, people were asked to choose a tone half as loud as a comparison tone. Some people were given a set of candidate tones that included the half-as-loud tone but were mostly quiet. Another

group was given a set of tones that also included the half-as-loud tone but were mostly loud. In both groups, people just selected a tone in the middle of the range, so in the quiet group people's estimates of the half loudness were much lower than in the loud group. The conclusion is that people have no real grip on absolute loudness. Other similar findings are consistent with the idea that people are unable to make reliable decontextualized judgments of absolute magnitudes.

A closely related phenomenon indicates that such psychophysical principles carry over to choice. In one study, people choose to trade off risk and return by choosing a gamble (of the form "p chance of x") from a varying range of options that was found to almost completely determine the choice. That is, people chose based not on absolute risk-return level but on the risk-return level relative to the other gamble options available. Parallel work on game playing and financial decisions found similar effects of skew and range, in line with the range-frequency theory of magnitude judgment. This pattern of responses (causing preference reversals) cannot be explained (produced) by any absolute measure of utility or related concepts such as the value-function in prospect theory and rank-dependent utility models.

Similar effects are discovered in medical decision making, in which the context of the rating task was found to influence the way participants distinguish between mild and severe scenarios. In one such study, both patients and nonpatients gave less distinct ratings to the two scenarios when each was presented in isolation than when they were presented alongside other scenarios that provided contextual information about the possible range of severity for lung disease. These results raise continuing concerns about the reliability and validity of subjective quality-of-life ratings, which appear sensitive to the particulars of the rating task. These effects are all predicted by the relativistic (contextual) judgment effects in psychophysics and risky decision making.

An extensive review of the literature also shows that people's judgments about the effectiveness of treatments and the healthcare decisions they make seem to be influenced by the different ways in which evidence from clinical trials can be presented. In particular, three different formats of data presentation have been the focus of a number of research studies: relative risk reduction, absolute risk reduction, and number of people who need to be treated to prevent one adverse event. For example, people gave higher mean ratings of a medical intervention's effectiveness when the benefits were described in terms of a relative risk reduction (34% relative decrease in the incidence of fatal and nonfatal myocardial infarction) rather than as an absolute risk reduction (1.4% decrease in the incidence of fatal and nonfatal myocardial infarction—2.5% vs. 3.9%) or a number-needed-to-treat format (77 persons must be treated for an average of just over 5 years to prevent one fatal or nonfatal myocardial infarction). This tendency is a robust finding across respondents (physicians, health professionals, patients, and the general public) and medical domains. These results can be explained by the relativistic account presented above. Due to a lack of stable underlying scales, people use the lower and upper bounds of 0% (the worst treatment available) and 100% (the best treatment available) of the probability scale as some sort of natural reference scale to map onto when they evaluate the attractiveness of something. Thus, both the relative and absolute risk reductions are evaluated with reference to the same 0% to 100% scale (and usually the former is bigger than the latter as in the example above). The number-needed-to-treat format presents an unbound psychological scale without natural upper limit on that number, and hence 1 out of 77 does not sound as convincing as 34% out of 100%. Similar effects are found in the marketing literature on price perception, where price reductions presented as percentages (save 10%) have stronger effect than amounts (save $3).

Implications

The accumulated evidence suggests that decision making is fundamentally context-dependent and judgments of the value of choice options are context-specific. The implications of this cognitive limit in medicine and public policy are serious, because they strike at the central methodologies used to measure preferences. Popular methods such as functional measurement and conjoint analysis measure trade-offs by asking respondents for attractiveness ratings of stimuli (e.g., policies) consisting of pairs of attributes (e.g., a reduction of x% in the annual risk of death for $$y$). Ratings of this sort are useful if the trade-offs are independent of what

other options are available. Such rationally irrelevant contextual factors are, for example, the range of values on each attribute within the session. Thus, if policy makers judge that a decrease from 20% to 15% in the annual risk of death is worth an expenditure increase from 10% to 30% of the medical (healthcare) budget, then this should be true regardless of whether the range of available expenditure options is from $10 million to $30 million or from $1 million to $100 million. Utility should depend on what happens (i.e., the actual outcomes in terms of 5% risk reduction and $20 million expenditure increase), not what options were considered. However, such independence is often not found and depends on various contextual factors. Therefore, professionals practicing medical decision making should be aware of such context effects to minimize the detrimental impact on clinical outcomes.

Ivo Vlaev

See also Attraction Effect; Bias; Contextual Error; Decision Psychology; Hedonic Prediction and Relativism; Heuristics

Further Readings

Covey, J. (2007). A meta-analysis of the effects of presenting treatment benefits in different formats. *Medical Decision Making, 27,* 638–654.

Ditto, P. H., Jacobson, J. A., Smucker, W. D., Danks, J. H., & Fagerlin, A. (2006). Context changes choices: A prospective study of the effects of hospitalization on life-sustaining treatment preferences. *Medical Decision Making, 26,* 313–332.

Kahneman, D., & Tversky, A. (Eds.). (2000). *Choices, values and frames*. New York: Cambridge University Press.

Lacy, H. P., Fagerlin, A., Loewenstein, G., Smith, D. M., Riis, J., & Ubel, P. A. (2006). It must be awful for them: Perspective and task context affects ratings for health conditions. *Judgment and Decision Making, 1,* 146–152.

Stewart, N., Chater, N., & Brown, G. D. A. (2006). Decision by sampling. *Cognitive Psychology, 53,* 1–26.

Contextual Error

Overlooking contextual information in the process of medical decision making can have predictable and avoidable adverse effects as significant as those that result from overlooking biomedical signs of a pathophysiologic condition. The failure, for instance, to recognize that a patient is not able to take a medication correctly (e.g., because of cognitive disabilities or cost) may have the same consequences as the failure to prescribe the medication correctly. While the latter type of error has been termed a diagnostic or medication error, the former is designated a *contextual error*.

Contextual Error Versus Biomedical Error

According to the Institute of Medicine (IOM), misguided clinical decision making or care delivery rises to the level of medical error when it results in either a wrong plan to achieve an aim (i.e., error of planning) or the failure of a planned action to be completed as intended (i.e., error of execution). Errors may be due either to failures to elicit essential information during the clinical encounter or, if elicited, to recognize the significance of essential information when formulating or implementing a plan of care.

Medical errors may be classified as contextual when they occur because of inattention to processes expressed outside the boundaries of a patient's skin (i.e., to processes that are part of the context of a patient's illness). They are distinguishable from biomedical errors, which are due to inattention to biomedical processes (i.e., to processes that occur within the patient). For instance, treating poor glucose control in a diabetic with metformin is a biomedical error if the patient has concomitant severe diabetic kidney disease because metformin can cause lactic acidosis in patients with poor renal function. Insulin is an acceptable alternative. On the other hand, prescribing self-administered insulin is a contextual error if the patient's poor control is due to dementia because dementia renders this approach unreliable and unsafe. Note that although dementia has biomedical origins, it is its expression outside the skin in the actions (or inactions) of the patient that are relevant here. It is a part of the context of his or her diabetes management.

Figure 1 presents a framework for comparing contextual error with biomedical error, organized according to failures to elicit or incorporate clinically significant information and to the IOM's

classification of medical error. Consider, for instance, examples of mechanisms B1 and C1: Overlooking signs of congestive heart failure in an asthmatic patient who is short of breath, unaware that he or she also has heart disease, is due to "incorrect/incomplete biomedical information." Overlooking medication nonadherence in a patient who is failing to respond to a medical therapy, unaware he or she is uninsured, is due to "incorrect/incomplete contextual information." Although one is a biomedical oversight and the other a contextual one, both cognitive processes lead to errors of planning in the IOM framework.

A typology of error that includes contextual error also illustrates the interdependence of biomedical and contextual information: Note that in mechanisms B2 and C2, biomedical and contextual information each determine whether the other is correctly processed. For example, attributing weakness to a patient's congestive heart failure, unaware that he or she fears exercising after his or her heart attack, represents "overlooking contextual information because of an incorrect biomedical explanation" (C2). Conversely, disregarding signs of dementia in a patient who is not taking his or her medication correctly, assuming he or she is

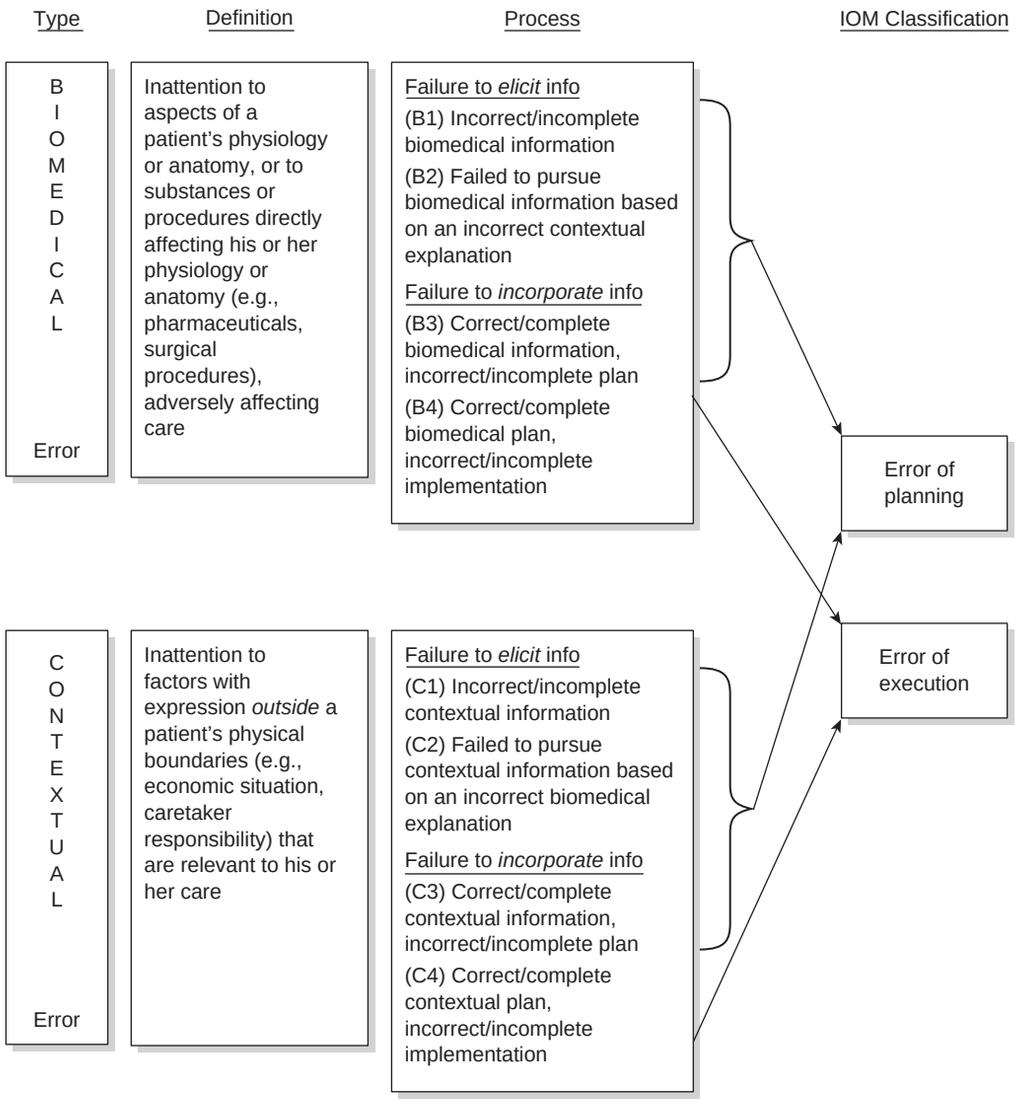


Figure 1 Biomedical versus contextual error

just not interested in complying with the recommended treatment plan, represents "overlooking biomedical information because of an incorrect contextual explanation" (B2).

B3 and C3 errors occur when correctly elicited information (biomedical or contextual) is not incorporated into the care plan. Finally, B4 and C4 pertain to errors in the implementation process. They represent errors of execution, per the IOM definition.

The problems caused by contextual and biomedical errors are remarkably similar. Problems caused by biomedical error have been classified as overuse, misuse, or underuse of medical services. Contextual errors may be classified similarly. For instance, overlooking poor medication adherence leads to overprescribing of additional medication. Sending a patient for elective surgery when he or she is unable to care for himself or herself postoperatively and lacks social support constitutes misuse. And not recognizing when an elder patient warrants evaluation for driving safety results in underuse of services.

Contextual Reasoning and Cognition

Contextualizing care requires cognitive skills distinct from those applied to biomedical decision making. Whereas biomedical reasoning classifies patients into known categories for which there are specific therapies, contextualized decision making explores how they differ from others with similar conditions in ways that require individualized care. Identifying when a diabetic patient with poor glucose control requires the addition of a second medication based on American Diabetes Association guidelines reflects the former; unmasking that the problem is, instead, poor medication adherence related to a diminishing capacity for self-care reflects the latter.

Categorization is generally arrived at through hypothesis testing: During an encounter, the clinician suspects that a patient has a particular condition that requires an accepted approach to care. The hypothesis is tested through clinical or laboratory examination. The process of reducing uncertainty continues until the patient falls into a sufficiently discrete category to prompt initiation of a specific therapy. Such an approach is essentially algorithmic.

The challenge of contextualization is in then discovering from the infinite complexity of the patient's life that which is unique to his or her life situation and relevant to the considered plan of care. As such, it requires moving from a deductive to a theory building approach to clinical reasoning in which unique elements of a patient's life are uncovered and assessed for clinical relevance. It involves a transition from asking "How is this patient similar to others?" to "How are they different?" Having asked and answered the question "Does this patient have diabetes?" one is now asking "Is there anything special about this individual's situation that is relevant to their diabetes management?"

Avoiding contextual errors requires considering contextual factors essential to planning patient care. Broadly these factors have been grouped into 10 categories to consider for each patient: cognitive abilities, emotional state, cultural beliefs, spiritual beliefs, access to care, social support, caretaker responsibilities, attitude toward illness, relationship with healthcare providers, and economic situation. Such factors may or may not have contextual relevance, depending on their relationship to the clinical problem. Simply getting to know a patient is not the objective here; rather, it is understanding how his or her life situation relates to his or her care.

When contextual factors are identified in the course of evaluating a clinical problem, they should prompt further inquiry. For instance, in the setting of deteriorating medication adherence, the clinician might ask of a patient with progressive dementia, "Is she still capable of taking these medications correctly?" If the context is economic, the question could be "Should I choose another medication because of the cost?" For social support, one might ask, "Now that he is weaker, will his wife still be able to care for him at home?" For spiritual beliefs, "Could her minister help her reach a decision?" The goal of these questions is not to place the patient into a predefined category for which there is a preconceived solution. Rather they are to unmask the particulars of a patient's life situation, pointing the way to an individualized plan of care.

Identifying Contextual Errors

The first challenge to identifying contextual errors is defining them. For many medical errors *res ipsa loquitur*, "the thing speaks for itself." If a surgeon

operates on the wrong limb or a pediatrician overlooks laboratory evidence of a serious infection in a newborn, there can be little disputing that an error occurred. It may be less clear when a physician's inattention to contextual factors also constitutes a medical error.

A second challenge is finding them. Many medical errors can be discerned from record reviews and incident reports. Contextual errors, however, rarely leave a footprint. The problem is that such errors are, by definition, errors only in a particular context. That context is the patient's life situation, and, if the error occurred, the relevant contextual factors were likely overlooked or their significance unrecognized and undocumented. Hence it is not feasible to identify the presence or absence of contextual errors by examining the medical record. For instance, two patients with a history of atrial fibrillation on warfarin may both meet evidence-based guidelines for anticoagulation; however, one of them may also have contextual contraindications such as transportation difficulties that compromise safe monitoring of the medication, a process that requires frequent blood draws. The clinician who did not attend to the transportation problems is also unlikely to have documented them.

It may not therefore be possible to define and identify contextual errors in clinical practice. An alternative, however, is an experimental rather than observational approach: Rather than looking for errors, one can create simulated situations where errors could occur and then see whether and how often they do. Current research employs incognito or unannounced, standardized patients (USPs) to present as if they are real patients in physician practices with scripted cases embedded with contextual information that is essential to care. If the provider fails to incorporate the contextually relevant factors into the plan of care, he or she will cause a medical error. Since the patient is only an actor, no real harm is done.

A critical component of the method is a protocol for validating each case as an instrument for assessing physician performance at contextualizing care: First a script is drafted based on a real scenario in which contextual factors seem essential to planning appropriate care. Then the narrative is presented to board certified clinicians with content expertise who are randomly assigned to review the text either with or without the critical contextual information. For instance, if the case involves unexplained weight loss in an impoverished homeless man, 10 reviewers are informed that the patient had inadequate access to food and the other 10 are not given this information. Both groups are told that all clinically relevant information has been provided and each clinician is instructed to propose appropriate care. The contextual information (i.e., inadequate access to food) is confirmed as clinically essential when all reviewers with the information propose an alternate plan from those without it. None of the reviewers may confer with one another about the cases. A case is considered validated when the two groups are internally consistent but 100% discordant in their recommended plans of care.

The use of standardized patients and validated cases addresses the challenges of defining and identifying contextual errors outlined above. Such an approach also enables comparison of physician performance across multiple providers in the same discipline. Standardized patients are intrinsically risk-adjusted in that every physician sees the same subject with the same narrative, providing an equivalent and objective standard for comparing practicing physicians.

Preventing Contextual Errors in Medical Decision Making

Considering psychosocial factors in the process of planning care is, of course, not new. In his seminal writing on the biopsychosocial model, George Engel introduced general systems theory as a framework for broadening the biomedical perspective to include social, psychological, and behavioral dimensions. In subsequent writing, he illustrated how perturbations in biomedical and psychosocial systems affect one another. Engel's model has stimulated many projects to define and describe the medical interview in a manner that incorporates psychosocial and biomedical elements into patient care. What has been missing, however, is a benchmark and metric for assessing how well clinicians perform at contextualizing or individualizing care. Contextual error is a discrete phenomenon that reflects the failure of the clinician to adequately integrate psychosocial with biomedical aspects of patient care.

With a metric it becomes possible to identify physician and practice characteristics that are

associated with contextual error making and to test interventions that may prevent it. Empirical research is limited but evolving. One recent pilot study of standardized patients and internal medicine residents demonstrated that about two thirds of clinicians in training made contextual errors involving cases with common ambulatory complaints when contextual information was essential to medical decision making. Remarkably, over half were due not to failures to elicit the information but to failures to incorporate it into the plan of care. Obtaining basic knowledge of how and why contextual errors occur should be invaluable to any subsequent effort to prevent their occurrence and ultimately improve patient outcomes.

Saul J. Weiner

See also Clinical Algorithms and Practice Guidelines; Cognitive Psychology and Processes; Medical Errors and Errors in Healthcare Delivery

Further Readings

Engel, G. L. (1977). The need for a new medical model: A challenge for biomedicine. *Science, 196*(4286), 1130.

Kohn, L. T., Corrigan, J. M., & Donaldson, M. S. (Eds.). (2000). *To err is human: Building a safer health system*. Washington, DC: Institute of Medicine, National Academy Press.

Weiner, S. J. (2004). Contextualizing medical decisions to individualize care: Lessons from the qualitative sciences. *Journal of General Internal Medicine, 19,* 281–285.

Weiner, S. J. (2004). From research evidence to context: The challenge of individualizing care. *ACP Journal Club, 141,* A11.

Weiner, S. J., Barnet, B., Cheng, T. L., & Daaleman, T. P. (2005). Processes for effective communication in primary care. *Annals of Internal Medicine, 142,* 709–714.

Weiner, S. J., Schwartz, A., Yudkowsky, R., Schiff, G. D., Weaver, F. M., Goldberg, J., et al. (2007). Evaluating physician performance at individualizing care: A pilot study tracking contextual errors in medical decision making. *Medical Decision Making, 27*(6), 726–734.

Contingent Valuation

Contingent valuation (CV) is a survey-based method to derive monetary values for the benefits of goods that are not available for purchase in the market. It specifies a hypothetical market whereupon the provision of the good is contingent on the respondent's maximum willingness to pay (WTP) for it (or, in a minority of cases, the minimum compensation they are willing to accept to be deprived of it). A hypothetical market is the construction, specification, and presentation of the imagined scenario on which respondents value the nonmarketed good. Individual values are aggregated to arrive at an overall societal value of the good. This value can then be compared with the societal cost of providing the good, in a cost-benefit analysis.

Why the Interest?

Interest in CV reflects dissatisfaction with other outcome measures, especially quality-adjusted life years (QALYs), in two principal respects. First, QALYs are based on preferences for *health outcomes* only, whereas CV imposes no restriction on which attributes of a program generate value, encompassing (a) health outcomes, including health state, duration, and probability; (b) other attributes, related to the process of care; (c) maintaining the good as an option for future consumption rather than for current consumption (option value); and (d) obtaining satisfaction from others, in addition to or rather than oneself consuming the good (externalities). Second, CV values benefits in the same unit as costs. This is required to assess whether the good represents an overall benefit in absolute terms (allocative efficiency), rather than a benefit relative to another option (technical efficiency). However, the reality is that few CV studies achieve these advantages in practice. Most studies use current patients, so they tend to capture only health outcomes, and few studies use their results to perform a cost-benefit analysis. The theoretical superiority of CV is thus seldom realized in practice.

How Has Contingent Valuation Developed?

CV has been used extensively in transport and environmental economics since the 1960s. It was first applied to healthcare in the mid-1970s, but only a handful of studies were completed before the late 1980s. The development of CV in health economics was led by researchers in the United

States, the United Kingdom, Canada, and Sweden, largely focused on cardiovascular disease. Since 2000, CV studies have been conducted in 35 countries, covering a vast range of diseases and interventions, although the single largest number of applications has been for pharmaceutical interventions (33%). However, CV studies remain rare, with only 265 studies published (as of December 31, 2005) compared with more than 35,000 other forms of economic evaluation on the OHE Health Economic Evaluation database.

Why So Few Studies?

Contingent valuation studies are incredibly complex, difficult, time-consuming, and costly to do well. This is because such studies face a number of methodological issues, for instance, framing effects (how the scenario is described), scale or scope biases (where WTP values are insensitive to the size or range of benefits described), payment vehicle and mode effects (where WTP values are affected by the payment method, e.g., taxation, out-of-pocket payment, or insurance) and payment frequency (e.g., weekly, monthly, or annually), and question order effects (where question order can affect results). These issues can be dealt with through adequate specification and administration of the market so that incentives to answer honestly are maximized. However, the issue of hypothetical bias, where respondents who do not actually have to part with money may state unrealistic valuations, may still be an issue even in a well-designed study since few opportunities exist to test this in practice in healthcare.

The most critical component is the specification and administration of the hypothetical market itself. *Specification* refers to, among other things, detailed information on the health problem, specifying the (attributes of the) good valued, determining the appropriate payment vehicle, how any element of uncertainty will be presented (as individuals are generally not risk-neutral), the relevant time period for valuation (which provides the foundation for the respondent's budget constraint), and the questionnaire format. This last aspect is especially controversial, and there remains considerable debate over the relative benefits of the five principal elicitation formats: (1) *open-ended,* where respondents are asked directly for their maximum WTP; (2) *bidding,* where respondents who accept or reject a given amount are bid up or down until maximum WTP is achieved; (3) *payment card* (or *categorical scales*), where a specified range of values is presented and respondents are asked to indicate which they would pay; (4) *dichotomous choice,* where respondents are presented with a single WTP value that they either accept or reject; and (5) *multibounded* dichotomous choice, where a single-bound dichotomous-choice question is followed with subsequent questions. The greatest difference is between the former three and latter two formats, where these surveys require different subsamples to be offered different values and logistic regression to be used to estimate the societal WTP.

Values drawn from a CV survey are determined by the characteristics of the hypothetical market specified, as above, as well as the characteristics of the respondent (preferences and income). The key to ensuring that only the latter varies is to undertake behavioral, rather than attitudinal, surveys. Behavioral surveys generate values that, although hypothetical, are substantive rather than formal and require a clearly defined market. This requires researchers to give detailed thought to what and how information is presented to respondents in the survey.

Administration of the hypothetical market refers to the use of face-to-face interview, remote interview (usually by telephone), and self-complete questionnaires (typically postal). In determining the mode of administration to be used, there is a balance to be struck between three factors: (1) the response rate (nonresponse is problematic if the sample not responding is likely to have a significantly different WTP compared with those who did respond); (2) the perceived validity of results (generally that a respondent's WTP will be more valid where respondents are encouraged to consider carefully the questions and their answers); and (3) the cost of the survey. Face-to-face interviews are overwhelmingly recommended to address points 1 and 2 but are very costly, and in health economics other methods are more typically used.

Analysis of results is also complex, particularly ensuring validity and reliability. *Validity* refers to the correspondence between what one wishes to measure and what is actually measured. Ideally validity is determined by comparing the measurement of interest to another measurement that is,

a priori, known to be correct (criterion validity)—in this case some form of market value is usually taken to be this external gold standard, reflecting the amount the individual would actually pay. Unfortunately, such a market value with which CV measurements can be compared rarely exists—which is the reason for conducting the CV survey, of course. Research has thus mostly looked to two different approaches to infer validity: construct validity (how well the measurement is predicted by factors that one would expect to be predictive a priori, e.g., that WTP is positively associated with income) and convergent validity (how comparable the values are from two different techniques for the measurement of a phenomenon, such as comparing the implied WTP ranking with ordinal ranking). *Reliability* refers to the reproducibility and stability of a measure. This may be cross-sectional (i.e., results are replicable when administered to independent samples) or temporal (i.e., results are stable when administered to the same sample at two different points in time). The first measure of reliability concerns the reliability of the measurement instrument itself—the instrument obtains the same information on repeated samples. The latter is a measure of the reliability of the WTP values themselves, commonly assessed using the test-retest method, where an initial sample of respondents is later reinterviewed using the same survey instrument. It is the latter that is important for policy purposes.

Although these issues should have been considered throughout the design and development of the study, surprisingly little work has been undertaken in these areas with respect to the use of CV in healthcare.

How Useful Is Contingent Valuation?

The ultimate purpose of conducting CV studies is to assist in medical decision making. However, there is a significant method-policy gap. While studies are increasingly being undertaken, most do not combine CV values with cost, so that a cost-benefit analysis cannot be undertaken. Furthermore, CV values themselves are not comparable due to considerable heterogeneity of methods.

In addition to incorporating cost information with CV studies, an obvious step in tackling the heterogeneity of methods is the development of guidelines. Such an agenda has already been applied in the cost-per-QALY arena, with conventions widely known and used. The closest steps made toward this in CV for health economics have been the five recommendations, made by Richard Smith, that need to be met by good-quality CV studies (response rate, association between WTP and socioeconomic status, sensitivity of WTP to scale and scope of the good, predictive validity, and reliability of elicitation methods), although even if these were met, studies could still fall short of providing the information needed to actually use the values elicited.

An objection to the development of guidelines could be the continued uncertainty around "best practice." However, guidelines for QALY studies were proposed despite methodological uncertainties. While it might be argued that uncertainties surrounding CV studies are larger, and that medical decision-making researchers actually want guidelines for cost-benefit studies rather than CV studies, it would still improve the usefulness of values elicited if methods were common because, relative to another value elicited using the same methods, researchers could infer the degree to which preferences were stronger or weaker. Such an approach need not hinder divergences from the guidelines for further methodological research to be undertaken; guidelines simply impose a constraint to include specific minimum design but do not preclude the use of other approaches or perspectives within the same study.

An alternative viewpoint is that CV is just not up to the job of informing cost-benefit analyses in healthcare (for the reasons mentioned) and therefore should not be used. However, such an opinion might accept the technique as very good at representing the public's intensity of preferences if one accepts the fact that people are familiar with the money metric used. Therefore, while CV should not be used to decide which alternative intervention to provide, it could be used to determine which of the alternative interventions the public really do prefer. Such an approach suggests a very limited and specific role for CV. Proponents of this approach may draw on the fact that results from CV studies are specific to the prevailing income distribution, such that if the current income distribution is not deemed equitable, then the results of CV may well overrepresent the interests of the most affluent in society.

Contingent valuation as applied in health economics is still experimental. Most studies fall far short of the requirements and recommendations in transport and environmental economics, and yet there has been no systematic evaluation of the specific developments that may be required in healthcare to justify such divergences from accepted practice in these other areas. Contingent valuation, and even more so full cost-benefit studies, remain rare in health economics, and their results are not comparable. Without the development of guidelines for the conduct of CV in healthcare, CV holds much unfulfilled promise.

Richard D. Smith and Tracey H. Sach

See also Cost-Benefit Analysis; Cost-Effectiveness Analysis; Cost-Utility Analysis; Discounting; Willingness to Pay

Further Readings

Bateman, I., Carson, R. T., Day, B., Hanemann, W. M., Hanley, N., Hett, T., et al. (2002). *Economic valuation with stated preferences techniques: A manual.* Cheltenham, UK: Edward Elgar.

Drummond, M. F., Sculpher, M. J., Torrance, G. W., O'Brien, B. J., & Stoddart, G. L. (2005). *Methods for the economic evaluation of health care programmes* (3rd ed., chap. 7). Oxford, UK: Oxford University Press.

Olsen, J. A., & Smith, R. D. (2001). Theory versus practice: A review of "willingness-to-pay" in health and health care. *Health Economics, 10,* 39–52.

Sach, T. H., Smith, R. D., & Whynes, D. K. (2007). A "league table" of contingent valuation results for pharmaceutical interventions: A hard pill to swallow? *PharmacoEconomics, 25,* 107–127.

Smith, R. D. (2003). Construction of the contingent valuation market in health care: A critical assessment. *Health Economics, 12,* 609–628.

Cost-Benefit Analysis

Cost-benefit analysis is a form of economic evaluation that can be used to assess the value in terms of money of healthcare interventions. In contrast with cost-effectiveness analysis and cost-utility analysis, which were developed specifically for the healthcare field, cost-benefit analysis has a long history of use in economics and is particularly linked to the theory of welfare economics. Its link with economic theory has led to some favoring this form of evaluation as the "correct" approach to problems of resource allocation in health systems, although it is worthy of note that other commentators have argued that the cost-utility analysis embodies its own theoretical properties and have coined the term *extrawelfarism* to counter the suggestion that only cost-benefit analysis has a grounding in economic theory.

The characterizing feature of cost-benefit analysis is the measurement of costs and benefits in the same units. In practice, this almost always means that the benefits are measured in monetary terms. For many noneconomists, the concept of placing a monetary value on health, and indeed on life itself, has seemed anathema. Indeed, this apparent aversion to monetary quantification of health outcomes explains the relative infrequency of the use of cost-benefit analysis in health economic evaluation, and the relative popularity of alternative evaluative forms such as cost-effectiveness and cost-utility analysis.

Nevertheless, advocates of the cost-benefit approach have continued to develop methods for the monetary valuation of health outcomes. Many early cost-benefit analyses were based on the human capital approach, which takes the (discounted) stream of lifetime earnings for an individual as a valuation of life. However, this approach implies a zero value for individuals outside formal paid employment and has become less used in recent years. More popular are *stated preference* methods that involve subjects responding to questions concerning their willingness to pay for health outcomes. When subjects are asked to reveal their willingness to pay for health outcomes directly, this is known as the *contingent valuation* approach. As with any method of preference elicitation, how such questions are framed can have important consequences for how a subject responds. However, the problems of framing effects and "protest" responses (where a respondent refuses to answer a question or gives a null value) seem particularly acute in contingent valuation of health outcomes. This may explain why much recent research has been based on using a class of methods known as *discrete choice experiments* that estimate preferences for different attributes at

different levels using a series of dichotomous choices across a carefully chosen choice set. When one of the attributes is cost, it is possible to generate indirect estimates of willingness to pay for the other attributes in the experiment. By specifying a profile of levels of the attributes associated with a health state or treatment under consideration it is possible to estimate a monetary value of that health state or treatment.

One of the problems associated with stated preference methods is the danger that respondents overstate their willingness to pay due to the hypothetical nature of the question. That is, if they really had to pay, it is likely that we would observe a lower willingness to pay for the health state or treatment under consideration. In general, *revealed preference,* where willingness to pay is estimated from observed actions in the marketplace, is preferred to stated preference methods. However, the opportunity for revealed preference studies in the healthcare field, where patients rarely pay for their own healthcare, is limited. One example where revealed preference has been used is in studies of behavior regarding radon gas remediation measures taken by households. Radon gas is a naturally occurring phenomenon that is associated with an increased risk of lung cancer and occurs in geographical areas where the geology of the area has a high proportion of granite in the bedrock. Since radon is heavier than air, the simple installation of a sump pump in low-lying areas, such as basements, can reduce the risk of lung cancer. Therefore, the willingness to pay at the household level for such remedial measures can be used to infer the willingness to pay for a reduced risk of lung cancer.

The measurement of both costs and benefits in monetary terms encourages the use of a net-benefit approach to decision making, whereby if a program's benefits exceed its costs, the program should be implemented. Indeed, the ability of cost-benefit analysis to make this comparison is argued by advocates of the approach to be one of its major advantages over other evaluative approaches. Nevertheless, notwithstanding the issues surrounding overestimating in stated preference techniques, many health systems work within a fixed budget for healthcare. In the face of a fixed budgetary constraint, efficient allocation of resources requires the prioritization of programs to be implemented in terms of their cost-benefit ratio rather than simply the condition that benefits exceed costs. From this perspective, the cost-benefit approach to resource allocation is similar to that when *cost-utility analysis* is employed.

Much debate has taken place over whether cost-benefit and cost-utility approaches are formally equivalent, in particular when a monetary value is placed on the quality-adjusted life year (QALY) in cost-utility analysis since this allows net-benefit analysis in monetary terms. The remaining difference between the approaches goes back to the theoretical foundations of cost-benefit analysis in terms of welfare economics. The cost-benefit approach assumes *consumer sovereignty,* that is, the principle that the individual is the best judge of his or her own welfare and it is therefore the individual's values that count. It is reasonable to ask whether this is generally true in healthcare, where there is an asymmetry of information between the physician and the patient regarding the consequences of healthcare intervention. It might be argued, therefore, that cost-benefit analysis in healthcare might work better in those situations where patients have more experience (e.g., visits to the dentist, frequently occurring and more minor problems such as infections and colds, and some chronic conditions such as asthma) and less well for infrequent and more severe problems where patients have little experience (e.g., life-threatening experiences such as cancer treatment).

Andrew H. Briggs

See also Contingent Valuation; Cost-Effectiveness Analysis; Cost-Utility Analysis; Discrete Choice; Economics, Health Economics; Human Capital Approach; Monetary Value; Net Monetary Benefit; Welfare, Welfarism, and Extrawelfarism; Willingness to Pay

Further Readings

Drummond, M. F., Sculpher, M. J., Torrance, G. W., O'Brien, B. J., & Stoddart, G. L. (2005). *Methods for the economic evaluation of health care programmes* (3rd ed.). Oxford, UK: Oxford University Press.

Gold, M. R., Siegel, J. E., Russell, L. B., & Weinstein, M. C. (Eds.). (1996). *Cost-effectiveness in health and medicine.* New York: Oxford University Press.

Cost-Comparison Analysis

A cost-comparison analysis estimates the total costs of two or more interventions, including downstream costs, and the numbers of individuals affected by each intervention but does not estimate cost-effectiveness ratios relative to health outcomes. This approach was developed in the early 1970s as a method of cost accounting with specific applications to ascertaining the lowest-cost methods of pharmacologic dosing and laboratory testing. An assumption that is usually either explicit or implicit in such analyses is that health outcomes are comparable across interventions. Otherwise, the lowest-cost strategy would not necessarily be desirable.

A cost-comparison analysis, which is also commonly referred to as a cost-consequences analysis, is less demanding to perform because it does not require clinical or epidemiologic data on health outcomes, such as long-term morbidity or mortality, although short-term clinical outcomes or healthcare use are typically reported. This approach is attractive in assessing interventions for which it is difficult to ascertain ultimate health outcomes or to calculate summary measures of health that integrate multiple outcomes. The cost-comparison approach is particularly well-suited to assessing screening and diagnostic-testing strategies. It is typical for such analyses to report summary cost ratios, such as cost per individual tested or cost per case detected, for each strategy, as well as incremental cost ratios for pairwise comparisons.

The time horizon, or the period during which healthcare utilization and costs are included in the analysis, is variable for cost-comparison (or cost-consequences) studies. For analyses of pharmacological or surgical interventions, the time horizon that is used is typically quite short, often 12 months to several years from the time of intervention. On the other hand, cost-comparison analyses of genetic testing strategies typically project the costs of monitoring tested individuals over their remaining lifetimes, which can be 40 years or more.

Most published cost-comparison analyses are conducted from the perspective of a healthcare system and only include direct medical costs. However, it is also valuable to calculate cost-comparison analyses from the societal perspective and to include costs occurring outside the healthcare system. Costs of time spent by patients and family members are important to include for interventions requiring substantial time by individuals and relatives. The exclusion of such costs can make such interventions appear more cost-effective than they are. In particular, if one is interested in comparing the actual costs of clinic-based and home-based therapeutic or rehabilitative strategies from a societal perspective, it is essential to include the costs of unpaid or informal caregiving services.

Prior to the mid-1990s, clear distinctions were generally made between cost-comparison, cost-minimization, and cost-consequence analyses. Since then, differences among these methods have become blurred, and articles using them frequently overlap one another. A given analysis that reports or assumes equivalent outcomes of different interventions might be labeled as a cost-comparison analysis, cost-consequence(s) analysis, cost-minimization analysis, or even cost-effectiveness analysis, depending on the preferences of the authors. Consequently, readers should not assume that the terminology used to describe such studies necessarily corresponds to differences in the analytic methods employed. Originally, cost-comparison analyses reported data only on costs, not on outcomes; cost-minimization analyses reported on costs only after ascertaining that health outcomes were equivalent for the interventions being compared; and cost-consequence analyses reported both costs and health outcomes but did not explicitly compare the two in terms of ratios (to let decision makers decide which information is needed to draw inferences).

Cost-comparison analyses differ from a cost-effectiveness or cost-utility analysis because they do not require a summary measure of health such as QALYs or number of symptom-free days to capture health gains. In addition to requiring less data, comparisons that are restricted to financial measures are often easier for healthcare payers and decision makers to understand and appreciate. If the costs included are restricted to short- or medium-term costs incurred within a single healthcare system or paid by a single payer, such a cost-comparison analysis can also be classified as a budget impact analysis, a business case analysis, or a return on investment analysis.

Many studies that report one intervention to be comparably effective but less costly than another

appear to have begun as standard cost-effectiveness or cost-utility analyses. It is likely that after investigators were unable to establish that one intervention was significantly more effective than another in preventing morbidity or mortality, they focused on showing that one particular intervention might be cost-saving. Rather than reflecting an a priori difference in study goals or analytic methods, as is assumed in textbook discussions of cost-consequence analyses, such studies likely indicate an absence of evidence of incremental effectiveness. If one intervention had been found to be more effective, incremental cost-effectiveness ratios would in most cases have been calculated and reported.

A sensitivity analysis allows one to determine the robustness of conclusions with regard to a decision rule. In a cost-comparison analysis that reports that an intervention is cost-saving, a sensitivity analysis can determine the extent to which variation in parameters affects the likelihood of the intervention being cost-saving. Although it is recommended that all economic evaluations include sensitivity analyses, not all cost-comparison analyses do so. This depends on the intended audience and the professional background of the investigators.

Scott D. Grosse

Disclaimer: The findings and conclusions in this entry are those of the author and do not necessarily represent the official position of the Centers for Disease Control and Prevention.

See also Cost-Consequence Analysis; Cost-Minimization Analysis

Further Readings

Bapat, B., Noorani, H., Cohen, Z., Berk, T., Mitri, A., Gallie, B., et al. (1999). Cost comparison of predictive genetic testing versus conventional clinical screening for familial adenomatous polyposis. *Gut, 44,* 698–703.

Koopmanschap, M. A., van Exel, J. N., van den Berg, B., & Brouwer, W. B. (2008). An overview of methods and applications to value informal care in economic evaluations of healthcare. *PharmacoEconomics, 26,* 269–280.

Naslund, M., Eaddy, M. T., Kruep, E. J., & Hogue, S. L. (2008). Cost comparison of finasteride and dutasteride for enlarged prostate in a managed care setting among Medicare-aged men. *American Journal of Managed Care, 14,* S167–S171.

Newman, W. G., Hamilton, S., Ayres, J., Sanghera, N., Smith, A., Gaunt, L., et al. (2007). Array comparative genomic hybridization for diagnosis of developmental delay: An exploratory cost-consequences analysis. *Clinical Genetics, 71,* 254–259.

Papakonstantinou, V. V., Kaitelidou, D., Gkolfinopoulou, K. D., Siskou, O. C., Papapolychroniou, T., Baltopoulos, P., et al. (2008). Extracapsular hip fracture management: Cost-consequences analysis of two alternative operative methods. *International Journal of Technology Assessment in Health Care, 24,* 221–227.

Cost-Consequence Analysis

A cost-consequence analysis (CCA) requires an estimation of the costs as well as the health consequences and other consequences associated with one intervention compared with an alternative intervention for a health condition; these estimates then are presented in a disaggregated tabular or graphical format. This type of analysis has been described in texts on economic evaluation of new healthcare interventions. However, it is generally mentioned only briefly and categorized as either a formal or an informal variant of a cost-effectiveness analysis (CEA).

Types

When a CCA is performed as a variant of a CEA, it takes an incidence-based perspective and estimates the costs and consequences for an individual or disease cohort for as long as the health condition lasts. However, a CCA also can be performed from a prevalence-based perspective, where the costs and consequences of alternative mixes of interventions can be compared over a 1-year time frame for a population with the condition of interest. This type of analysis is an expanded version of a budget impact analysis (BIA). Health and other consequences of the alternative mixes of interventions are presented annually for the population, as are the costs, which are aggregated by cost category.

Since a single overall number is not generated as a result of a CCA, the perspective does not have to

be chosen by the analyst. The perspective of a CCA should be as broad as possible, since the user of the analysis should be able to view a comprehensive listing of the various costs and consequences of alternative interventions. The user then can choose which variables are relevant for their perspective and can ignore the others.

Time Horizon

The time horizon for a CCA should be chosen in the same way as the time horizon for the CEA, the cost-utility analysis (CUA), or the BIA. For an incidence-based CCA, the time horizon will vary, depending on the health condition and the type of intervention, as shown in Figure 1. The duration of the impact of the intervention on the individual with the health condition is the primary determinant of the appropriate time horizon, with acute nonfatal illness requiring a shorter time horizon and chronic or fatal illness requiring up to a lifetime time horizon. Whether or not a healthcare intervention is for prevention or treatment also is a determinant of the appropriate time horizon for the analysis. For a prevalence-based CCA, the chosen time horizon should be relevant to the decision maker. Typically, annualized costs and consequences for 1 to 5 years after a change in treatment patterns is the most relevant time horizon.

Scope

As with other types of economic evaluation, the question of scope of the analysis is important: For example, with interventions that affect life expectancy, should the costs and consequences of alternative interventions include their impact only on condition-related outcomes, or should the impact on healthcare costs and outcomes for other conditions be considered? Generally, costs and consequences for unrelated health conditions are not considered in economic evaluations. Also, which specific costs and consequences should be included is less restricted in a CCA than in a typical CEA or CUA. For a CCA, outcomes may be included that are not typically part of a CEA, a CUA, or a BIA, such as social service costs and dosing convenience. Finally, alternative interventions may have different impacts on different population subsets, and a separate analysis for these different population subsets is important for all types of economic evaluations, including the CCA.

The following are types of costs that can be included in a CCA: direct healthcare costs; other direct costs, including social service costs and transportation costs; indirect costs, including productivity losses and criminal justice costs; and intangible costs, including costs related to the quality-of-life impact of pain and concern about disease prognosis. Since the goal of the CCA is to give the decision maker as broad a view as possible of the costs of alternative healthcare interventions, all costs that are relevant for the condition of interest should be included. Clearly, the types of costs included will vary with the condition: For example, for an acute illness such as influenza, direct healthcare costs and productivity losses are the most important costs to include. For a chronic psychiatric illness such as schizophrenia, social service costs and criminal justice costs also will be important to include. In addition, intangible costs are important in the analysis of all chronic illnesses.

The following are types of consequences that can be included in a CCA:

- Disease symptoms
- Cure rates
- Mortality rates
- Treatment side effects
- Treatment convenience
- Treatment adherence and persistence
- Patient and family quality of life
- Patient and family overall well-being
- Patient and family satisfaction with treatment

Since the goal of the CCA is to give the decision maker as broad a view as possible of the consequences of the alternative interventions, all aspects of the alternative interventions should be included in the analysis, including convenience and patient and family satisfaction with treatment. These types of consequences generally are not included in CEAs and frequently are not included in CUAs. For example, for influenza, the two neuraminidase inhibitors have different dosing modes: via inhalation (zanamivir) and tablets (oseltamivir). For the treatment of a human immunodeficiency virus infection, many combination treatments are now available that have easier dosing regimens for the patient, which may increase adherence and

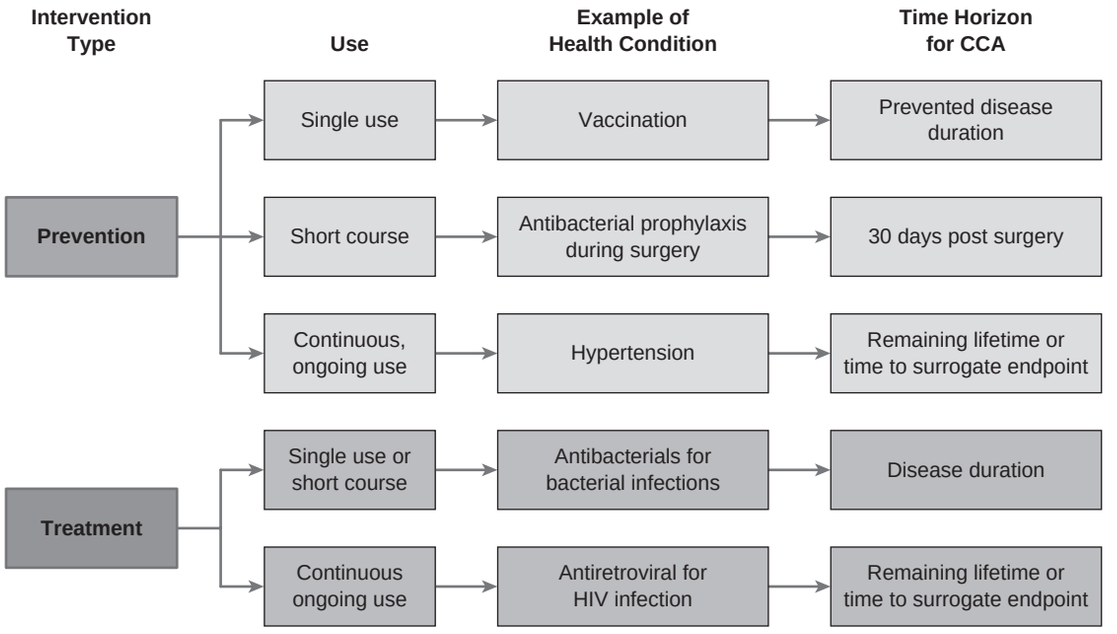


Figure 1 Time horizon for cost-consequence analysis

Source: Adapted from Mauskopf, J. A., Paul, J. E., Grant, D. M., & Stergachis, A. (1998). The role of cost-consequence analysis in healthcare decision making. *PharmacoEconomics, 13,* 277–288.

Note: CCA, cost-consequence analysis; HIV, human immunodeficiency virus.

persistence with treatment and thus increase the effectiveness of the treatment.

As with all economic evaluations, the scope and accuracy of a CCA is limited by the data available. Figure 2 shows the primary data sources for a CCA. Although randomized, controlled clinical trials provide an important data source for CCAs, such trials may have limited external validity because of their generally restrictive inclusion and exclusion criteria. Naturalistic clinical trials or observational data may provide data that more closely approximate the likely costs and consequences in standard clinical practice. Finally, for a chronic illness, the results from a disease progression model can be used to generate estimates of the long-term consequences of alternative interventions when only short-term outcomes data are available.

Sensitivity

Sensitivity analysis is an important component of any economic evaluation because of uncertainty in the input data as well as the modeling assumptions and other assumptions used to estimate the costs and consequences of the intervention. Thus, the sensitivity of the results of the CCA to changes in the input parameter values and all assumptions should be estimated. One possible way to present this component of the analysis is to use estimates of the ranges of different input parameter values (e.g., 95% confidence intervals for data taken from clinical trial data) to estimate a range of values for each of the costs and consequences estimated.

Presentation

The key distinguishing feature of a CCA is the presentation of the results in a simple, disaggregated format. An example of a CCA presentation is given in Table 1. The cost information should be presented in units (e.g., days in the hospital, physician visits) as well as by cost. The costs also should be presented separately for different cost categories as well as in total. Treatment modes and

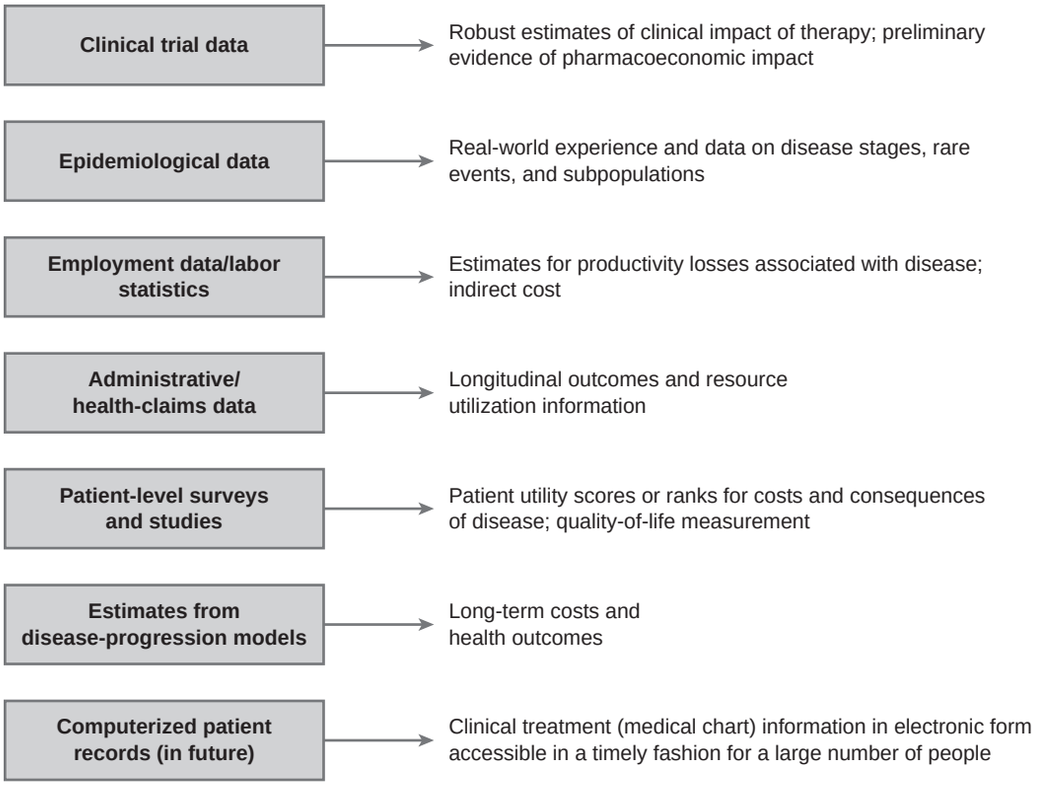


Figure 2 Data sources for cost-consequence analysis

Source: Adapted from Mauskopf, J. A., Paul, J. E., Grant, D. M., & Stergachis, A. (1998). The role of cost-consequence analysis in healthcare decision making. *PharmacoEconomics, 13,* 277–288.

Note: CCA, cost-consequence analysis.

convenience can be included in the tabular listing of the consequences of treatment. In addition to the outcomes for the alternative treatments, a tabular presentation of the results should include two columns showing the difference between the interventions, in units and costs, for each outcome. For an incidence-based CCA, the tabular listing of results applies to one individual or a cohort of individuals over the appropriate time horizon. For a prevalence-based CCA, the tabular listing of results applies to the population of interest to the decision maker and gives annualized results for a 1- to 5-year time horizon.

Advantages and Limitations

There are two types of CCA: An incidence-based CCA can be considered to be a variant of a CEA for a representative individual or for a disease cohort, without the limitation of the consequences to a single outcome and without the calculation of a single ratio of costs to outcomes. The time horizon for an incidence-based CCA is the same as for a CEA, and the data sources will be the same as those for the CEA, with additional sources required for additional cost and consequence measures. A prevalence-based CCA is an extension of a BIA for a prevalent population with the health condition of interest; a prevalence-based CCA includes a broader range of cost categories as well as annualized population estimates of the health and other consequences of a change in the intervention mix.

There are several advantages of a thorough CCA of alternative interventions as an adjunct to other economic value measurements:

- It provides disaggregated information and well-understood measures for a decision maker's review.

Table 1 Example of table of results of cost-consequence analysis for two drugs

Cost Components	*Drug A Units*	*Drug A Costs*	*Drug B Units*	*Drug B Costs*	*Difference (A – B) Units*	*Difference (A – B) Costs*
Direct medical care use and costs						
Drug A or Drug B						
Other drugs						
Physician visits						
Hospital days						
Home care						
Other medical care (e.g., dialysis)						
Direct nonmedical care use and costs						
Transportation						
Social service costs						
Crutches or other equipment						
Paid caregiver time						
Indirect resource use or cost						
Time missed from work for patient						
Time missed from other activities for patient						
Time missed from work for unpaid caregiver						
Time missed from other activities for unpaid caregiver						
Criminal justice costs						
Total direct and indirect costs						
Symptom impact						
Patient distress days						
Patient disability days						
Quality-of-life impact						
Quality-of-life profile scores for patient						
Quality-of-life profile scores for family						
Quality-adjusted life-years decrement for patient						
Quality-adjusted life-years decrement for family						
Patient perception of treatment						
Patient satisfaction scores						
Family satisfaction scores						
Dosing convenience						
Drug adherence						
Drug persistence						

Source: Adapted from Mauskopf, J. A., Paul, J. E., Grant, D. M., & Stergachis, A. (1998). The role of cost-consequence analysis in healthcare decision making. *PharmacoEconomics, 13,* 277–288.

Note: CCA, cost-consequence analysis.

- It allows a decision maker to assign his or her weights to health and other consequences, rather than having an analyst assign weights.
- There is no loss of information when compared with other value measures.
- It can include many consequences that may not be accounted for in other measures, such as dosing, convenience, and patient satisfaction.
- The results can be used as the inputs for a CEA, CUA, or BIA estimate.

There are also some limitations to a CCA:

- Benchmark values and league tables of alternative interventions cannot be developed.
- Direct comparison of value across disease areas is not possible.
- There is no overall quantitative assessment of the value of a new treatment.
- The application of decision-maker weights to the outcome measures may result in decisions based on self-interest rather than on societal value.

Healthcare decision makers need information about the costs and consequences of alternative interventions for different reasons: to determine whether or not to reimburse the different interventions for all the population with the condition of interest or a subset of that population and to determine the extent to which additional healthcare funding will be needed to pay for new interventions for all the population with the condition of interest or a subset of that population. To make these determinations, different national and local healthcare decision makers require information on the costs and consequences of alternative interventions in different formats and with different perspectives, scopes, and time horizons. The CCA can be considered to be a variant of a CEA or an extension of a BIA and can allow the decision maker to choose the combination of costs and consequences that is relevant to him or her and to apply his or her own weights to the consequences.

Because of its limitations in terms of providing overall societal or payer value measures and the associated lack of benchmark values and ability to perform cross-disease comparisons, a CCA will provide the most value when presented together with the results of a CEA, a CUA, and a BIA. Such a package of information will provide a comprehensive assessment of the economic value that can meet all the information requirements of local or national healthcare decision makers.

Josephine Mauskopf

See also Cost-Effectiveness Analysis; Cost-Utility Analysis

Further Readings

Drummond, M. F., O'Brien, B. J., Stoddart, G. L., & Torrance, G. W. (1997). *Methods for the economic evaluation of health care programmes* (2nd ed.). Oxford, UK: Oxford Medical Publications.

Gold, M. R., Siegel, J. E., Russell, L. B., & Weinstein, M. C. (1996). *Cost-effectiveness in health and medicine.* New York: Oxford University Press.

Grant, D. M., Mauskopf, J. A., Bell, L., & Austin, R. (1997). Comparison of valaciclovir and acyclovir for the treatment of herpes zoster in immunocompetent patients over 50 years of age: A cost-consequence model. *Pharmacotherapy, 17,* 333–341.

Kernick, D. (2002). *Getting health economics into practice.* Abingdon, UK: Radcliffe.

Mauskopf, J. A., Paul, J. E., Grant, D. M., & Stergachis, A. (1998). The role of cost-consequence analysis in healthcare decision making. *PharmacoEconomics, 13,* 277–288.

McMurray, J. J. V., Andersson, F. L., Stewart, S., Svensson, K., Solal, A. C., Dietz, R., et al. (2006). Resource utilization and costs in the Candesartan in Heart Failure: Assessment of Reduction in Mortality and Morbidity (CHARM) programme. *European Heart Journal, 27,* 1447–1458.

PausJenssen, A. M., Singer, P. A., & Detsky, A. S. (2003). Ontario's formulary committee: How recommendations are made. *PharmacoEconomics, 21,* 285–294.

Rosner, A. J., Becker, D. L., Wong, A. H., Miller, E., & Conly, J. M. (2004). The costs and consequences of methicillin-resistant *Staphylococcus aureus* infection treatments in Canada. *Canadian Journal of Infectious Diseases and Medical Microbiology, 15,* 213–220.

Straka, R. J., Mamdani, M., Damen, J., Kuntze, C. E. E., Liu, L. Z., Botteman, M. F., et al. (2007). Economic impacts attributable to the early clinical benefit of atorvastatin therapy: A US managed care perspective. *Current Medical Research Opinion, 23,* 1517–1529.

Wang, Z., Salmon, J. W., & Walton, S. M. (2004). Cost-effectiveness analysis and the formulary decision-making process. *Journal of Managed Care Pharmacy, 10,* 48–59.

Cost-Effectiveness Analysis

Cost-effectiveness analysis involves comparison of the additional costs and health benefits of an intervention with those of the available alternative(s). The aim of such an analysis is to determine the value in terms of money of the intervention(s). Within a cost-effectiveness analysis, the health benefits associated with the various interventions are measured in terms of natural units (e.g., survival, life years gained, the number of clinical events avoided). This entry introduces the concept of cost-effectiveness analysis and reviews the key elements, including the incremental cost-effectiveness ratio (ICER), the cost-effectiveness plane, the cost-effectiveness threshold, and the cost-effectiveness frontier.

Concept

The objective of economic evaluation of healthcare interventions is to inform resource allocation decisions in the healthcare sector, through determining whether a proposed intervention is a "good" use of scarce resources. This is assessed through comparison of the additional resources consumed (costs) for the improvement in health benefits generated (e.g., life years gained) associated with one health intervention compared with another. Cost-effectiveness analysis, where the health benefits are measured in terms of a single dimension represented by natural units, is just one type of economic evaluation. It is used to determine which of the alternative interventions provides the most efficient method to achieve a particular outcome (technical efficiency). As such, the units chosen to represent the effect in a cost-effectiveness analysis should be deemed worthwhile (to society or the policy maker), appropriate for measuring the key impact of the intervention, and common across the alternatives to be compared. For example, the cost-effectiveness of a screening test may be established in terms of the cost per case detected, the cost per percent survival at 5 years, the cost per life saved, or the cost per life year gained. Ideally, the measure of effect chosen will relate to a final outcome (e.g., life years gained), but where this is not possible, there should be a way to link it to final effect (e.g., symptom days averted), or it should be deemed to have value in itself (e.g., cancers detected). Alternative methods for economic evaluation include cost-benefit analysis (where health benefits are measured and valued in monetary terms) and cost-utility analysis (where quality of life is considered alongside quantity of life and health benefits are valued according to patient preferences to construct a composite measure of health outcome, e.g., the quality-adjusted life year or QALY). It should be noted, however, that sometimes the term *cost-effectiveness* is used to cover any of these methods of economic evaluation, where the comparison need not be measured in natural units.

Perspective

The perspective of the analysis determines the extent of the costs and health benefits measured and incorporated. Taking a societal perspective, as advocated by economists, requires the measurement and valuation of all the effects of the intervention(s) irrespective of where, or whom, they affect, including all healthcare costs, all non-healthcare costs, and all costs to the patient, his or her family, and carers. Narrower perspectives restrict the impacts that are included within the analysis, making them more manageable. For example, adopting the commonly used third-party payer perspective for costs would restrict measurement to the costs that fall on the payer (e.g., health insurance company) but would exclude any costs which fall directly on the patient or his or her family and carers. Restricting the perspective for health benefits to the patient would exclude any health benefits received by his or her family, friends, or carers or an altruistic society.

Incremental Cost-Effectiveness Ratio

Cost-effectiveness is assessed by relating the additional costs incurred to provide an intervention to the additional health benefits/effects received as a result of the intervention compared with the available alternative(s). This information is generally reported as an incremental cost-effectiveness ratio (ICER)—a measure of the additional cost per unit of health gain:

$$\text{ICER} = \frac{\text{Cost}_{\text{new intervention}}}{\text{Effect}_{\text{new intervention}}} - \frac{\text{Cost}_{\text{current intervention}}}{\text{Effect}_{\text{current intervention}}}.$$

Incremental cost-effectiveness ratios are only calculated between interventions that address the same patient group with the aim of identifying and selecting the most efficient of these competing (mutually exclusive) interventions. For example, different methods of managing adult women with symptoms of urinary tract infection are mutually exclusive and can be compared within a cost-effectiveness analysis. Cost-effectiveness ratios are not calculated between interventions that address distinct (independent) patient groups. This is because both, or all, the independent interventions may be selected as cost-effective. For example, methods for managing children with symptoms of urinary tract infection should not be compared within a cost-effectiveness analysis with methods for managing men or women. However, once the analysis is done and the ICERs are calculated, independent interventions can (and should) be compared with each other to determine which are funded in a resource-constrained system. This is only plausible where the units of outcomes are measured on the same scale (e.g., life years) for the various interventions or where there is a known common trade-off between the various outcomes.

When determining ICERs for a set of mutually exclusive interventions, the interventions should be ranked in ascending order of effect (or cost) and a ratio calculated for each intervention relative to the next best (more costly) viable intervention by dividing the additional cost by the additional health benefit involved.

Interventions that are both less effective and more costly than other interventions are deemed "dominated" and are not considered viable (Step 2 below). This is because a decision maker should never select an intervention that is both more costly and less effective than an alternative. Interventions that involve larger ICERs than other, more effective, alternatives are deemed "extended dominated" and are also not considered viable (Step 5 below). This is because the intervention would be "dominated" by a program that consisted of a mixture of the next most effective and the next less effective interventions and therefore should not be selected (see Figure 3).

Calculating the ICER: An Example

Consider a situation where there are six mutually exclusive interventions (A to F) that could be adopted. These interventions are characterized by the costs and effects given in the table below.

	Effects	*Costs ($)*
B	2	211,500
A	10	41,868
D	38	256,731
C	48	879,500
E	68	1,138,000
F	73	1,601,500

(1) Step 1: Rearrange in order of ascending effect.

(2) Step 2: Exclude any interventions where the cost is higher than for an alternative intervention with a greater effect (dominated).

(3) Step 3: Calculate the incremental effect and incremental cost of each intervention in comparison with the prior (less effective) intervention.

	Effects	*Costs ($)*	*Inc. Effect*	*Inc. Cost ($)*
B	2	211,500	Dominated by A	
A	10	41,868	—	—
D	38	256,731	28	214,863
C	48	879,500	10	622,769
E	68	1,138,000	20	258,500
F	73	1,601,500	5	463,500

(4) Step 4: Calculate the incremental cost-effectiveness ratio for each successively more effective intervention, compared with the previous intervention in the list.

	Effects	*Costs ($)*	*Inc. Effect*	*Inc. Cost ($)*	*ICER*
B	2	211,500	Dominated by A		
A	10	41,868	—	—	—

D	38	256,731	28	214,863	$7,674
C[a]	48	879,500	10	622,769	$62,277[a]
E	68	1,138,000	20	258,500	$12,925
F	73	1,601,500	5	463,500	$92,700

a. Extended dominated.

(5) Step 5: Identify and exclude any interventions that have a higher ICER than more effective interventions (extended dominated), and recalculate the ICERs.

	Effects	*Costs*	*Inc. Effect*	*Inc. Cost ($)*	*ICER Recalculated*
B	2	211,500	Dominated by A		
A	10	41,868	—	—	—
D	38	256,731	28	214,863	$7,674
C	48	879,500	Extended dominated		
E	68	1,138,000	30	881,269	$29,376
F	73	1,601,500	5	463,500	$92,700

Repeat Step 5 until all dominated interventions are removed and ICERs have been calculated for all nondominated interventions.

Cost-Effectiveness Plane

A cost-effectiveness (CE) plane can be used to provide a visual representation of the results of a cost-effectiveness analysis by plotting the costs against the effects for the various interventions. When comparing just two mutually exclusive interventions, the incremental cost-effectiveness (ICE) plane can be presented as in Figure 1. Here the figure shows a plot of the additional (or incremental) costs and effects of the intervention compared with the alternative (represented by the origin).

The horizontal axis divides the plane according to incremental cost (positive above, negative below), and the vertical axis divides the plane according to incremental effect (positive to the right, negative to the left). This divides the incremental cost-effectiveness plane into four quadrants through the origin. These four quadrants are commonly referenced according to the compass points. The northwest

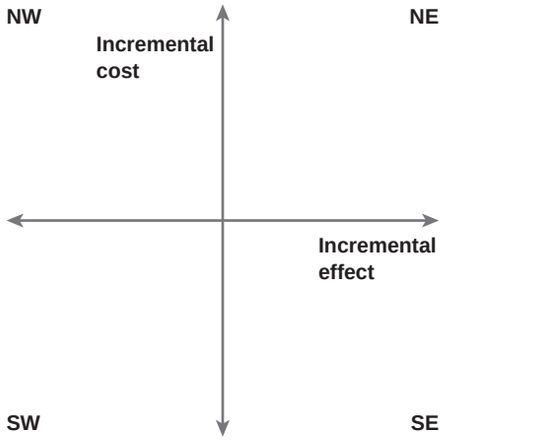


Figure 1 Incremental cost-effectiveness plane

(NW) quadrant involves negative incremental effect but positive incremental cost, as such an intervention falling in this quadrant would be "dominated" by the alternative and therefore not be considered cost-effective. The southeast (SE) quadrant involves negative incremental cost but positive incremental effect; an intervention falling in this quadrant would dominate the alternative and therefore be deemed cost-effective. The northeast (NE) quadrant involves positive incremental cost and positive incremental effect, while the southwest (SW) quadrant involves negative incremental cost and negative incremental effect. An intervention falling into either of these quadrants *may* be deemed cost-effective compared with the alternative, depending on the trade-off between costs and effects. Note that the incremental cost-effectiveness ratio associated with an intervention in either the NE or SW quadrant is given by the slope of a line connecting the intervention to the origin.

When comparing more than two mutually exclusive interventions, a cost-effectiveness plane can be plotted, where all interventions appear within the cost and effect space. Alternatively, and more commonly, the interventions can be plotted relative to the least costly, least effective alternative (represented by the origin) on the incremental cost-effectiveness plane (see Figure 2). Note that this requires calculation of the additional costs and effects of each alternative with respect to the same least costly, least effective comparator. In this case, identifying dominant or dominated interventions

can be done by systematically dropping a horizontal line and a vertical line through the point representing each comparator. This essentially replicates the process undertaken above for two interventions by making each point in turn the origin. The incremental cost-effectiveness ratio associated with an intervention is determined by the slope of a line connecting it with the next less effective, nondominated, alternative. Steeper slopes represent larger incremental cost-effectiveness ratios.

Plotting an ICE Plane for Multiple Interventions: An Example

Figure 2 illustrates the incremental cost-effectiveness plane for the six mutually exclusive interventions (A to F) under consideration.

The figure clearly indicates that Intervention B is dominated by Intervention A, which involves greater health benefits for a lower cost. In addition, the figure indicates that Intervention C is extended dominated as it involves a higher ICER than a more effective intervention, E (the slope of the line joining D and C is greater than the slope of the line joining C and E). Figure 3 illustrates that a mixed strategy involving programs D and E, represented by any point between M^1 (a strategy with identical health benefits to C that can be achieved at cheaper cost) and M^2 (a strategy involving identical costs to C that involves greater health benefits) dominates Intervention C.

Cost-Effectiveness Frontier

On the cost-effectiveness plane, the cost-effectiveness frontier is established by connecting together progressively more effective, nondominated interventions. This frontier has a gradually increasing slope (ICER), representing the increased price that must be paid for additional effects. The cost-effectiveness frontier for Figure 3 is represented by the line ADEF.

Cost-Effectiveness Threshold: Identifying the Cost-Effective Intervention

Once the dominated interventions have been excluded, the ICERs calculated, and the cost-effectiveness frontier established, one of the remaining (viable) interventions is identified as cost-effective and providing value for the money. Traditionally, the cost-effective intervention is identified as the one associated with the largest ICER that falls below a specified monetary threshold (often denoted by λ). This externally set cost-effectiveness threshold represents the maximum amount that the decision or

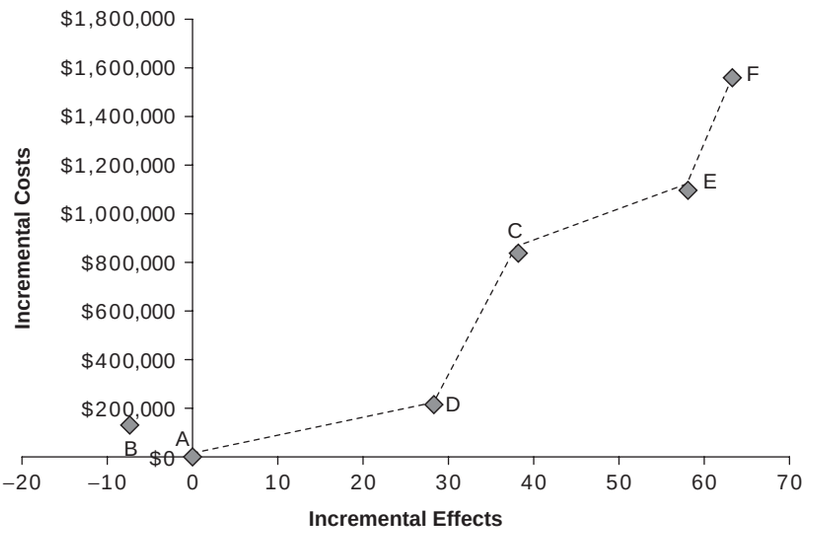


Figure 2 Interventions on the incremental cost-effectiveness plane

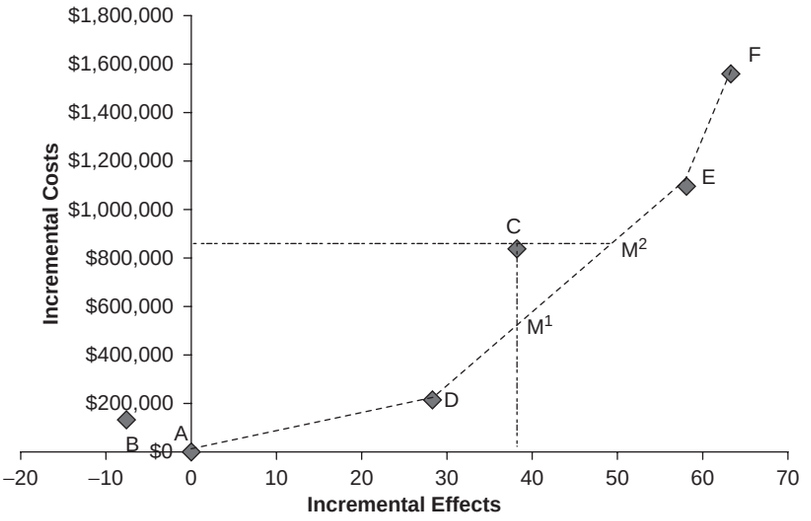


Figure 3 Cost-effectiveness frontier

policy maker is willing to pay for health effects. The threshold can be derived in two ways. The first method involves establishing and fixing the threshold at the maximum price that society is willing to pay for health benefits. The second approach involves deriving the shadow price of health benefits purchased from a fixed budget, that is, the amount by which the health benefits achievable will be improved by relaxing the fixed budget by a small amount. This approach, while theoretically correct, has an enormous informational requirement. Initially it involves selecting interventions onto a list, lowest ICER first, until the budget is expended. During the process, independent interventions are added to the budget while mutually exclusive interventions with higher ICERs replace those with lower ICERs that were included earlier. The following hypothetical example considers a situation where there are three independent programs (cancer screening, management for diabetes, and treatment for heart failure) and each program involves a choice between four viable mutually exclusive interventions (1 to 4) that could be considered cost-effective. These interventions are characterized by the costs and effects given in the table below.

Assuming a budget of 100,000, the initial budget determination would be as follows:

(6) Step 1: Implement b1 based on lowest ratio of cost to effect.

$$b1 \rightarrow \text{Budget used} = 1{,}500$$
$$\rightarrow \text{Shadow price} = 1{,}250$$

(7) Step 2: Add a1 based on ratio of cost to effect.

$$a1 + b1 \rightarrow \text{Budget used} = 6{,}000$$
$$\rightarrow \text{Shadow price} = 2{,}813$$

(8) Step 3: Replace b1 with b2 based on ratio of incremental cost to incremental effect.

$$a1 + b2 \rightarrow \text{Budget used} = 6{,}500$$
$$\rightarrow \text{Shadow price} = 3{,}846$$

	Cancer Screening (a)			*Management of Diabetes (b)*			*Treatment for Heart Failure (c)*		
	Costs	Effects	ICER	Costs	Effects	ICER	Costs	Effects	ICER
1	4,500	1.60	—	1,500	1.20	—	26,000	3.00	—
2	18,000	2.30	19,286	2,000	1.33	3,846	43,000	3.50	34,000
3	27,000	2.69	23,077	7,100	1.67	15,000	65,900	4.10	38,167
4	47,000	3.15	43,478	14,800	1.80	59,231	178,000	5.30	93,417

(9) Step 4: Add c1 based on ratio of cost to effect.

$$a1 + b2 + c1 \rightarrow \text{Budget used} = 32{,}500$$
$$\rightarrow \text{Shadow price} = 8{,}667$$

(10) Step 5: Replace b2 with b3 based on ratio of incremental cost to incremental effect.

$$a1 + b3 + c1 \rightarrow \text{Budget used} = 37{,}600$$
$$\rightarrow \text{Shadow price} = 15{,}000$$

(11) Step 6: Replace a1 with a2 based on ratio of incremental cost to incremental effect.

$$a2 + b3 + c1 \rightarrow \text{Budget used} = 51{,}100$$
$$\rightarrow \text{Shadow price} = 19{,}286$$

(12) Step 7: Replace a2 with a3 based on ratio of incremental cost to incremental effect.

$$a3 + b3 + c1 \rightarrow \text{Budget used} = 60{,}100$$
$$\rightarrow \text{Shadow price} = 23{,}077$$

(13) Step 8: Replace c1 with c2 based on ratio of incremental cost to incremental effect.

$$a3 + b3 + c2 \rightarrow \text{Budget used} = 77{,}100$$
$$\rightarrow \text{Shadow price} = 34{,}000$$

(14) Step 9: Replace c2 with c3 based on ratio of incremental cost to incremental effect.

$$a3 + b3 + c3 \rightarrow \text{Budget used} = 100{,}000$$
$$\rightarrow \text{Shadow price} = 38{,}167$$

Following the initial budget determination, including new interventions (whether independent or mutually exclusive) will involve displacing intervention(s) already included on the list. Thus, when considering a new intervention, the associated ICER is compared with that of the last intervention(s) included in the list that will be displaced by the new program (in the example, this is 38,167). It is the ICER of the displaced intervention that implicitly provides the shadow price for health benefits.

Identifying the Cost-Effective Intervention: An Example

Returning to the example, once the ICERs have been calculated for the nondominated interventions, they should be compared with the cost-effectiveness threshold to establish which intervention provides value for the money.

Assuming a cost-effectiveness threshold of $50,000, Intervention E would be identified as cost-effective as this provides the largest effect at an acceptable "price" (i.e., largest ICER below the cost-effectiveness threshold).

	Effects	*Costs ($)*	*Inc. Effect*	*Inc. Cost ($)*	*ICER Recalculated*
B	2	211,500	Dominated by A		
A	10	41,868	—	—	—
D	38	256,731	28	214,863	$7,674
C	48	879,500	Extended dominated		
E	68	1,138,000	30	881,269	$29,376
F	73	1,601,500	5	463,500	$92,700

A threshold of $20,000 would mean Intervention D was cost-effective, while a threshold of $100,000 would mean Intervention F was cost-effective.

Elisabeth Fenwick

See also Cost-Benefit Analysis; Cost-Utility Analysis; Dominance; Marginal or Incremental Analysis, Cost-Effectiveness Ratio

Further Readings

Black, W. C. (1990). The CE plane: A graphic representation of cost-effectiveness. *Medical Decision Making, 10,* 212–214.

Cantor, S. B. (1994). Cost-effectiveness analysis, extended dominance, and ethics: A quantitative assessment. *Medical Decision Making, 14*(3), 259–265.

Drummond, M. F., O'Brien, B. J., Stoddart, G. L., & Torrance, G. W. (1997). *Methods for the economic evaluation of health care programmes* (2nd ed.). New York: Oxford University Press.

Johannesson, M., & Meltzer, D. (1998). Some reflections on cost-effectiveness analysis. *Health Economics, 7,* 1–7.

Karlsson, G., & Johannesson, M. (1996). The decision rules of cost-effectiveness analysis. *Pharmaco-Economics, 9,* 113–120.

Weinstein, M. C., & Stason, W. B. (1977). Foundations of cost-effectiveness analysis for health and medical practices. *New England Journal of Medicine, 296*(13), 716–721.

Cost-Identification Analysis

Cost-identification analysis is the assignment of a value to healthcare use. The costs of healthcare encounters, treatment episodes, or healthcare interventions are found to consider the economic impact of medical decisions. Cost identification is part of budget impact analysis, cost-minimization analysis, cost-comparison analysis, cost-consequences analysis, cost-effectiveness analysis, cost-utility analysis, and cost-benefit analysis.

Cost-identification analysis is affected by the choice of analytic perspective and time horizon. This choice depends on the type of application and its intended audience. Cost-identification analysis is also affected by simplifying assumptions that may sacrifice comprehensiveness or precision to save research expense.

Standardized Methods

Medical decision models are commonly used to assess the cost-effectiveness of new healthcare interventions. Guidelines for cost-effectiveness analysis (CEA) have been developed so that the cost-effectiveness ratios of different interventions conducted by different analysts may be compared without concern that differences are methodological artifacts. Standardization also enhances the generalizability of study findings, allowing them to be applied to new settings.

Although there are a number of different guidelines for CEA, they largely agree on the principles of cost identification. These guidelines recommend that all relevant costs be included, that resources be valued at their opportunity cost, and that cost be estimated from the societal perspective using a long-term time horizon.

Standards have also been developed for budget impact analysis (BIA). This type of study provides healthcare plans or healthcare providers with information on the total cost of implementation. BIA generally uses a short-term horizon and the perspective of a particular health plan or provider.

Perspective of Analysis

Most studies consider all costs incurred in the healthcare system. Adoption of the societal perspective requires inclusion of costs incurred by patients and their families. These include cost of unpaid caregivers, cost of travel to medical care providers, and the value of time seeking care. A cost-identification analysis sometimes includes the value of wages lost due to illness. In practice, many studies ignore costs incurred by patients and their families. This may result in analyses that are biased in favor of interventions that shift costs from health system to patient.

Time Horizon

The time horizon is the period over which costs are identified. CEA guidelines recommend a long-term perspective, one that includes lifetime costs and outcomes. The use of a short time horizon may result in bias. A short-term horizon may favor an intervention that defers costs to the future or disadvantage one in which benefits are realized after significant delay. A short-term horizon may be appropriate to the immediate concerns of a BIA.

Cost-identification analysis ordinarily expresses the cost of care that spans more than 1 year in real (inflation-adjusted) terms. Future costs are discounted (expressed as the present value) to reflect the lower burden imposed by healthcare costs that will not be incurred until the future. Inflation adjustment and discounting are separate adjustments; both adjustments are needed.

The time horizon has an additional effect on cost-identification analysis; it determines whether fixed costs and development costs are included. In the short run, the decision to provide an additional health service does not increase institutional overhead (e.g., the cost of nonpatient care hospital departments such as human resources, finance, administration, and environmental services). These costs are fixed in the short run. In economic terms, the short-run marginal cost is the cost directly attributable to producing an extra unit of output and does not include the fixed costs of the enterprise.

In the long run, the institution must adjust the size of overhead departments to provide the right amount of services needed by its patient care departments. Additional health services increase institutional overhead over the long run. In economic terms, the long-run marginal cost is equal to the average cost. In other words, the long-run cost of producing an extra unit of output includes the

variable cost associated with that output, and a share of the fixed costs of the enterprise.

The difference in time horizon means that BIA ordinarily involves marginal cost and excludes facility overhead. Since guidelines recommend a long-term time horizon, institutional overhead is included in CEA.

The time horizon may also determine whether the cost of developing a new intervention is included. In the short run, the decision maker may regard these as sunk costs, an expenditure that has already been made and is not relevant to subsequent decisions. The long-run horizon requires inclusion of development costs.

The market price of pharmaceuticals must result in sufficient revenue so that over the long run the manufacturer can recoup development costs and earn a return on investment. Managerial and behavioral interventions are often developed as part of a research study, and their development costs are often ignored by analysts. Consistency requires inclusion of the cost of developing these interventions. This cost should be amortized over the expected size of the population of beneficiaries. Failure to include development costs may bias analyses against interventions such as pharmaceuticals and devices, which include development cost as part of their market price.

Methods of Determining Cost

Cost should be representative of the healthcare system where the study will be applied. Cost-identification methods include gross costing, use of data from claims and cost allocation systems, and microcosting. The choice between these methods represents a trade-off between precision and expense. Microcosting is the most accurate method, but it is labor-intensive. Gross costing is less accurate but much easier to employ. Each method has its limits and appropriate use. Multiple methods may be needed within a single study.

Gross Costing

Gross costing requires information on the quantity of each type of health service used and information on unit costs. A count of the resources employed in a particular healthcare strategy may be based on a hypothetical model or expert opinion. Alternatively, actual use may be recorded during the course of a clinical trial. Gathering service use from a study participant involves a trade-off between accuracy and expense. Accuracy can be improved by more frequent surveys and by employing logs and other memory aids. Counts of resources may also be obtained from administrative data of providers or health plans.

The cost of hospitalization may be estimated with different unit costs, including an average daily rate, a specialty-specific daily rate, or a diagnosis-weighted rate. Use of an average daily rate makes the assumption that all days of hospitalization have the same cost. Daily costs vary markedly by diagnosis, however. The accuracy of cost estimates is enhanced if they reflect the effect of diagnosis and the use of surgery and intensive care. Separate rates should be used to estimate the cost of hospitalization in psychiatric and long-term care facilities.

The cost of ambulatory care can be estimated by multiplying a count of visits by a unit cost. Not all ambulatory care visits have the same cost. The accuracy of cost estimates can be improved if they reflect differences in care, such as a hospital clinic or other facility, medical and surgical procedures, emergency room care, or a visit to a specialist or office-based care physician.

Estimates of pharmacy use are often based on patient self-report. The average wholesale price should not be used as the unit cost for pharmacy as healthcare payers receive substantial discounts from this price. Unit cost should also reflect the dispensing fee paid to pharmacies.

It is not desirable to estimate unit costs based on the fee schedule or cost data from a single provider as they may not be representative. Gross costing is not appropriate if the intervention affects the resources employed in care without affecting the units chosen to measure cost.

In the United States, Medicare is the predominant payer, and its payment schedule is often used for unit costs. Physician fee schedules are also available in countries outside the United States. A set of standard unit cost estimates have been offered as part of CEA guidelines used in the Netherlands and Australia. These unit costs have helped standardize estimates of health services costs in CEA studies of new pharmaceuticals. Such standard estimates must be used with care. If a

standard estimate for an ambulatory visit includes the cost of associated laboratory tests, it will not capture the incremental effect of an intervention that generates additional laboratory orders.

Gross costing is an important method appropriate for many studies, but the analyst should avoid any analytic assumption that interferes with identification of the effect of intervention on resource use.

Cost Estimates Based on Claims Data

Charges, cost-adjusted charges, and reimbursements from administrative data are widely applied by economic analysts based in the U.S. healthcare system. Administrative data are much less freely available outside the United States. Even within the United States, claims data may not always be available. Managed care organizations are reimbursed according to the number of patients served, and not for the type or quantity of services they provide. As a result, they may not prepare a claim, or they may not be required to provide claims data to the healthcare sponsor.

Claims data provide information on cost from the point of view of the healthcare payer or provider, and this is often the economic cost. Raw charges should not be used as an estimate of the cost of care as they greatly exceed the economic cost.

Charges are cost-adjusted by multiplying by a ratio of cost to charges. This ratio may be determined from data in publicly available cost reports that U.S. hospitals submit to Medicare. Use of cost-adjusted charges makes the strong assumption that the charge for a specific service is proportionate to its economic cost. This assumption is not always warranted. Hospitals may set their charges without knowing the relative cost of different services. There are strategic reasons to overcharge for some services and undercharge for others.

Some analysts have found costing to be more accurate if cost adjustment is done at the department level. A ratio of cost to charges is found for each department in the hospital and applied to the charges incurred in that department. It may be difficult to obtain charges at the department level. Departments may be defined differently in cost reporting and billing systems, making department-level adjustment problematic.

U.S. hospital bills exclude physician charges for inpatient services, and these must be estimated separately. When ambulatory care is provided by a facility, the facility and physician bill separately, and neither cost should be ignored.

Cost information is rarely available to adjust charges for physician services. When charges cannot be cost-adjusted, reimbursement may be a more appropriate estimate of cost. It should include any co-payment made by the patient.

An important limitation to administrative data is their coverage. The analyst must take care not to ignore significant costs not recorded in administrative data.

Activity-Based Costing (ABC) Systems

Activity-based costing (ABC) systems are used in some hospitals in the United States, Taiwan, and Canada. ABC systems are more complex than the cost reports U.S. hospitals submit to the Medicare program to determine reimbursement rates. Costs, services, and products are identified at a much finer level of detail. ABC systems extract databases to determine the quantity of all different services provided. The costs of staff time, supplies, and equipment are assigned to departments. Overhead expenses are distributed to patient care departments. A schedule of relative values is used to find the cost of specific products. The cost of these products is assigned to specific stays or encounters according to products used in providing care.

An important limitation of ABC systems is that they have not been widely adopted. Hospitals may regard ABC estimates as confidential information needed to negotiate contracts. The analyst must consider that hospitals using ABC systems may not have typical costs.

Microcosting

Microcosting is the direct measurement of cost by observation and survey. It is needed when no unit cost is available from fee schedules or claims systems. A common application is to estimate the cost of a novel intervention. Microcosting may be needed when claims data are not sensitive to the effect of an intervention. Since microcosting is too labor-intensive to use for all healthcare, its use must be limited to activities most likely to be affected by the intervention under study.

To find the cost of a treatment innovation, all intervention-related activities must be identified.

When patients must be screened to determine if they are eligible for treatment, this cost is not a research cost but a cost that should be included as it will be incurred when the intervention is replicated in clinical practice.

The cost of labor should not be limited to wage costs. It should also include the employer's share of taxes and benefits. Labor cost is often estimated by determining the number of minutes each worker spends in direct activities involved in providing the service. This effort is measured by direct observation, staff activity logs, supervisor report, or other methods. When the long-run perspective is used, labor costs should include nonpatient care activities: training to maintain credentials, answering the phone, meeting with colleagues, taking vacations, and going on sick leave.

For hospitals and other large institutions, it is not feasible to use microcosting to determine overhead. One approach is to apply the ratio of overhead to direct expense for a similar department in the hospital cost report.

Choice of Cost Method

Guidelines agree that more exact methods should be used to determine the cost of services most affected by the intervention under study. Simpler methods may be employed to avoid spending scarce resources on precise measurement of unimportant services. Each method involves assumptions, and the analyst must review whether these assumptions are appropriate. An important additional concern for CEA studies is the wider applicability of cost estimates and resulting study findings to other providers, health plans, or countries.

Reporting Cost-Identification Analysis

The quality of economic evaluations of healthcare has been studied in a large number of reviews. Some reviews have found modest improvements in the quality of economic analyses since the promulgation of CEA guidelines. Reviews have noted problems with the CEA studies in general and cost-determination methods in particular.

Studies should separately identify the cost of the intervention being evaluated and include all relevant costs. The analyst should identify the cost determination method used, source of cost data, time horizon, analytic perspective, price index used to adjust for inflation, and discount rate used to express costs in their present value.

Paul G. Barnett

See also Cost-Comparison Analysis; Cost-Consequence Analysis; Cost-Effectiveness Analysis; Cost Measurement Methods; Cost-Minimization Analysis; Costs, Direct Versus Indirect; Costs, Fixed Versus Variable; Costs, Out-of-Pocket; Cost-Utility Analysis; Marginal or Incremental Analysis, Cost-Effectiveness Ratio

Further Readings

Dranove, D. (1995). Measuring costs. In F. A. Sloan (Ed.), *Valuing health care: Costs, benefits, and effectiveness of pharmaceuticals and other medical technologies* (pp. 61–75). Cambridge, UK: Cambridge University Press.

Drummond, M. F., Sculpher, M. J., Torrance, G. W., O'Brien, B. J., & Stoddart, G. L. (2005). *Methods for the economic evaluation of health care programmes* (3rd ed.). Oxford, UK: Oxford University Press.

Luce, B., Manning, W., Siegel, J., & Lipscomb, J. (1996). Estimating costs in cost-effectiveness analysis. In M. R. Gold, J. E. Siegel, L. B. Russell, & M. C. Weinstein (Eds.), *Cost-effectiveness in health and medicine*. New York: Oxford University Press.

Mauskopf, J. A., Sullivan, S. D., Annemans, L., Caro, J., Mullins, C. D., Nuijten, M., et al. (2007). Principles of good practice for budget impact analysis: Report of the ISPOR Task Force on good research practices—budget impact analysis. *Value Health, 10*(5), 336–347.

Ramsey, S., Willke, R., Briggs, A., Brown, R., Buxton, M., Chawla, A., et al. (2005). Good research practices for cost-effectiveness analysis alongside clinical trials: The ISPOR RCT-CEA Task Force report. *Value Health, 8*(5), 521–533.

Smith, M. W., & Barnett, P. G. (2003). Direct measurement of health care costs. *Medical Care Research Review*, *60*(3 Suppl.), 74S–91S.

Cost Measurement Methods

Cost measurement is fundamental to all economic evaluations for healthcare, which have become increasingly important throughout the world in policy decision making concerning new medical

interventions. Appropriate cost measurement can contribute to the efficient allocation of resources within the health system. The goal of cost measurement is to assess the costs that are needed to produce or are consequent to the outcomes of the intervention of interest, relative to an alterative intervention such as standard of care.

Cost measurement involves identifying, measuring, and valuing all relevant resource uses that are attributable to the medical interventions, including the resources needed for implementing the interventions and those that are associated with medical and nonmedical outcomes of the interventions. The costs related to a healthcare intervention and its outcomes can include direct medical and nonmedical costs, as well as indirect costs (e.g., work productivity loss).

Generally, cost measurement methods include the following steps: (a) specifying the perspective of the study, (b) identifying relevant resources used, (c) determining the quantity of resources, and (d) valuing these resource items (services, goods, time).

Specification of Study Perspective

Economic evaluation studies can frame decision problems from different perspectives, which will lead to different costs and even the final decision. Therefore, specifying the perspective of the cost analysis plays a crucial role in determining the relevant resources and how they should be measured and valued.

Societal perspective is a standard for cost-effectiveness analysis. From a societal perspective, all resources and their net costs to the society should be taken into account, including patient and unpaid caregivers' time, as well as work productivity loss. A government purchaser may only bear the costs incurred to the government; thus the patient and unpaid caregivers' time would not be included from a government payer perspective. A commercial insurer may only be concerned with the direct medical costs; thus direct medical costs will be included. Therefore, the perspective of the study will determine the relevance of the resources, their quantity, and their costs.

Identification of Relevant Resources

A healthcare intervention has various and far-reaching effects with economic implications. Ideally, any use of resources that are associated with the alternative healthcare interventions and their effects on health outcomes should be identified. Depending on the study perspective, the nature of the intervention, and health outcomes, many or all of the following resources should be considered in the process of cost measurement: the acquisition and administration of the healthcare intervention (e.g., drug, provider service for the intervention, patient's time involved in the intervention), additional services (e.g., follow-up lab tests) associated with the intervention, change in healthcare resource uses associated with change in health status and outcomes, and change in non-healthcare resource uses associated with change in health status, such as improved work productivity and reduced unpaid caregivers' time. The study perspective should be considered in determining whether each component should be finally included.

The timelines during which these resources have implications should be considered, which determines the appropriate time horizon of the economic study. An appropriate time horizon should allow inclusion of the full consequences of the intervention.

Though theoretically all relevant resources and costs should be included, the availability of information, resources, and research time are often limited. Some resource items are likely to form the largest components (i.e., cost drivers) of the total and incremental costs. They often involve only a few resource items. These cost drivers should be considered first, especially those resources on which the intervention has a measurable impact. Therefore, the process of resource identification and costing requires scientific rigor as well as researchers' discretion because costing is a methodology for practical purposes.

Measurement of Resource Use

Depending on the list of identified resource uses, the data sources to quantify these resources can include randomized clinical trials (RCTs), an administrative and accounting database, observational studies such as a large national survey and patient registry, the published literature, clinical practice guidelines, and expert opinions. It is common for a variety of data sources to be used in economic modeling studies. The quantification of

resources in measurement units also depends on the costing method to be used, such as gross costing or microcosting methods.

It is an increasingly common practice to conduct economic evaluation alongside an RCT, which is often termed a *piggyback evaluation*. Such an evaluation has the advantage of leveraging randomization (thus with good internal validity for incremental costs), and availability of individual patients' data that provide variation and distribution for costs estimation. Some resource use data may be available among those collected for the purposes of the clinical trial. For example, an occurrence of hospitalization is almost always recorded as a serious adverse event because it is a necessary component for reporting in RCTs. Additional resource uses often need to be collected. Patient medical records or patient diaries/interviews can be used. Medical records provide accurate patient-level resource use information without an additional burden on the patients enrolled in the trial, but the records may not include all relevant resources, and the recording method may not be standard across different centers (especially for internal studies). Patient diaries or interviews allow the recording of nonmedical resource uses, such as time (e.g., transportation), that are incurred with the intervention.

In piggyback evaluation, special issues should be noted in resource quantification. In RCTs, extra resources could be consumed due to more frequent lab monitoring, additional scheduled follow-up visits, and better compliance. Such protocol-driven resource use should be excluded to approximate the costs incurred in the real-world practice better. In addition, RCTs often have a short follow-up period, and therefore not all resource use differences between two treatment arms are fully realized. Additional research effort could be expended to overcome such limitations, including conducting an open-label extension study if possible; extrapolating from final endpoints observed in the trial using modeling techniques; and predicting final outcomes from intermediate outcomes based on established models.

Resource Valuation

The assignment of costs to resources, or costing, can be performed from an aggregate (gross) level to a more detailed microlevel. Three basic costing methods are gross costing (top-down costing), unit costing, and microcosting (bottom-up costing).

In gross costing, health services or healthcare interventions are broken down into large components, and these large cost items have to be identified. As a result, gross costing can be simple and transparent. Gross costing estimates an event or diagnosis as a whole. National tariffs are preferably used whenever available, such as diagnosis-related group (DRG) payments in the United States and Australia and health resource group (HRG) payments in Great Britain. These rates are often reliable and standard and allow international comparison.

Unit costing applies costs to each type of resource consumed, such as emergency room visits, inpatient hospitalization stays, physician visits, lab tests and procedures performed, and drugs administered. Unit costs can be obtained from the national payment schedule (e.g., Medicare reimbursement rates), administrative claims database, and published literature.

The microcosting (bottom-up costing) method establishes a very detailed service delivery process (inventory) and identifies the relevant resource items and measures them separately. It is based on direct observation, on an item-by-item basis; thus it could be expensive and time-consuming. Microcosting methods include time-and-motion studies, activity logs, and surveys of patients, providers, and managers. In a piggyback evaluation, microcosting can also be conducted by reviewing medical bills of patients in trials. In the United States, a common method for estimating the economic cost of medical services is to adjust the charges through the use of cost-to-charge ratios to reflect their true economic costs.

A patient can incur loss of time due to time spent in seeking treatment, impaired productivity while at work, and short- or long-term absences from work associated with poor health status. Two methods that are generally used to measure work loss are the human capital method and friction cost method. The human capital method estimates the production cost during the employment period that is lost due to illness. However, the friction method restricts the period of the productivity loss to the period needed to replace the sick employee. So the productivity loss to society is limited to the time before the sick person is replaced.

Special attention should be given to determining drug cost, especially costs of patented drugs. Though the price of the brand drug is often used in economic evaluation, this can overestimate drug cost because the price of the brand drug is not the true market price of the drug with a complex rebate and co-payment system. From a payer perspective, the net drug payment incurred by the payer should be used, which is net of all rebates, co-payments, and other adjustments. From a societal perspective, because the cost transfer from one party to another within the society should be excluded from the costs, the drug price should be greatly discounted for economic evaluation. This is because a portion of the cost is transferred to a pharmaceutical company for rewarding innovation in drugs.

In perfectly competitive markets, the prices of inputs are equal to opportunity costs, but this does not hold for many components in healthcare. Consequently, tariffs and other prices in the healthcare sector should be applied with care, and often other valuation methods are used instead. Ideally, a resource used should be valued at its opportunity cost, that is, the value of its best alternative use. The concept of opportunity costs can help determine the value of those resources.

Incremental costs, instead of total costs, are of central interest because often two or more treatments are compared and evaluated during the decision-making process. Therefore, the cost measurement should focus on the difference in costs between treatments, and common costs that are invariant to treatments should be excluded.

Because costs can be measured in different ways, the choice of cost measurement should depend on the purpose of the study, and it has consequences for the identification of resource items and the measurement of resource use. Some general elements should be clarified in the cost measurement, including the perspective, the list of assumptions, the role of prices, the time horizon, and allocation of overhead costs. The choice of costing method in practice will be highly conditional on the information available, the limited resources available to undertake the analysis, and whether the study involves multiple countries.

Andrew Peng Yu

See also Cost-Comparison Analysis; Cost-Effectiveness Analysis; Cost-Identification Analysis; Costs, Direct Versus Indirect; Costs, Opportunity; Marginal or Incremental Analysis, Cost-Effectiveness Ratio; Time Horizon

Further Readings

Brouwer, W., Rutten, F., & Koopmanschap, M. (2001). Costing in economic evaluations. In M. Drummond & A. McGuire (Eds.), *Economic evaluation in health care: Merging theory with practice* (pp. 68–93). Oxford, UK: Oxford University Press.

Glick, H. A., Doshi, J. A., Sonnad, S. S., & Polsky, D. (2007). *Economic evaluation in clinical trials*. Oxford, UK: Oxford University Press.

Luce, B., Manning, W. G., Siegel, J. E., & Lipscomb, J. (1996). Estimating costs in cost-effectiveness analysis. In M. R. Gold, J. E. Siegel, L. B. Russell, & M. C. Weinstein (Eds.), *Cost-effectiveness in health and medicine* (pp. 200–203). New York: Oxford University Press.

Oostenbrink, J. B., Koopmanschap, M. A., & Rutten, F. F. (2002). Standardisation of costs: The Dutch Manual for Costing in economic evaluations. *PharmacoEconomics, 20,* 443–454.

O'Sullivan, A. K., Thompson, D., & Drummond, M. F. (2005). Collection of health-economic data alongside clinical trials: Is there a future for piggyback evaluations? *Value in Health, 8,* 67–79.

Cost-Minimization Analysis

Cost-minimization analysis is a special form of cost-effectiveness analysis where the health outcomes can be considered to be equivalent between two treatment alternatives and therefore the interest is only on which of the two strategies has the lower cost. Cost-minimization analysis appears to have much to commend it: in particular, it embodies an apparently simplified approach to decision making by looking at only the cost side of the equation. However, there are a number of potential pitfalls that exist in terms of the practical use of cost-minimization analysis.

The first of these represents a problem of definition. Many apparent examples of cost-minimization studies fail to present any justification of the

equivalence of health outcomes between two treatments and are therefore more accurately described as cost analysis. A simple cost analysis should not be considered a true cost-minimization study without some form of evidence for the equivalence of health outcomes being presented. Note that these cost analyses are also often incorrectly described as "cost-benefit analyses" due to the net-benefit approach to decision making, particularly in the early health economic evaluation literature.

More recently, as economic evaluation alongside clinical trials has become more common, the problem has become one of interpretation. It is all too common to see "cost-minimization analyses" presented that turn out to be based on the interpretation of lack of significance of an effect measure in a clinical trial as evidence of equivalence. In the clinical trial field, there is a well-known adage that "absence of evidence is not evidence of absence." To interpret the lack of a significance as evidence of no effect is to place the importance of the Type I error (concluding a difference exists when the null hypothesis of no difference is true) above that of the Type II error (concluding that no difference exists when in fact the alternative hypothesis of a difference is true). To properly show that two treatments are no different (within a small margin of error) requires an appropriately designed equivalence study that typically requires a greater sample size to reliably demonstrate equivalence than is recruited to many superiority (difference) trials.

Furthermore, clinical trials typically are powered to detect differences in only a single effect measure (primary trial endpoint). In contrast, health economic analyses are multidimensional, often trading off different effects (risks and benefits) to obtain a composite measure of outcome. It would be very rare indeed for two treatments to be truly equivalent on all measures of outcome and rarer for a clinical trial to be adequately powered to demonstrate such a multidimensional equivalence.

As a consequence of these difficulties, examples of true cost-minimization studies are rare. One of the most popularly cited (though rather old) examples relates to a cost-minimization study of alternative oxygen delivery methods, with the underlying assumption that the treatment (oxygen) is truly equivalent between alternative delivery systems. It is worthy of note that the original analysis (in common with the healthcare perspective of many economic studies) did not include any convenience to the patient in the analysis.

Although conceptually appealing, due to the simplified approach to decision making, the practical problems associated with cost-minimization analysis have led some commentators to argue the "(near) death of cost-minimization analysis." The appropriate framework for analysis of most studies will be the estimation of cost-effectiveness. It is clear that the use of separate and sequential tests of hypothesis of cost and effect based on superior study designs does not constitute appropriate grounds for using cost-minimization as a decision-making tool.

Andrew H. Briggs

See also Cost-Benefit Analysis; Cost-Effectiveness Analysis; Marginal or Incremental Analysis, Cost-Effectiveness Ratio

Further Readings

Briggs, A. H., & O'Brien, B. J. (2001). The death of cost-minimization analysis? *Health Economics, 10,* 179–184.

Drummond, M. F., Sculpher, M. J., Torrance, G. W., O'Brien, B., & Stoddart, G. L. (2005). *Methods for the economic evaluation of health care programmes* (3rd ed.). Oxford, UK: Oxford University Press.

Gold, M. R., Siegel, J. E., Russell, L. B., & Weinstein, M. C. (1996). *Cost-effectiveness in health and medicine.* New York: Oxford University Press.

Costs, Direct Versus Indirect

Within economic evaluation, the analysis of costs is meant to provide a valuation of the resources consumed as a result of an intervention. Such an analysis, like that involved in the valuation of outcomes, would result in different answers depending on the perspective of the analyst. The perspective adopted is, in turn, determined by the policy question that the evaluation is seeking to answer. For instance, the health sector perspective generally includes costs of treatment and cost offsets, that is, costs and cost savings to the health sector through, say, differences in hospitalizations associated with differences in outcomes between intervention

alternatives. Such a perspective typically includes out-of-pocket payments incurred by patients and charges on other funders of healthcare, including government and health insurers. A narrower perspective on costs might be justified if the evaluation is to address specific funding questions, for instance, to an individual insurer where costs incurred beyond the organization are deemed not to be relevant. Alternatively, a broader societal perspective might be relevant in instances where it is of interest to compare the intervention with options outside the health sector or if the policy in question is concerned with the potential economic impact on patients and their households.

The adoption of the societal perspective generally means the inclusion of indirect costs. These refer to resources incurred outside the health sector, including costs to patients, carers, and firms. Direct costs, in contrast, pertain to the specific resources involved in the delivery of a health intervention, for example, costs of medications, medical consultations, and equipment used in treatment. These terms have very specific definitions in economic evaluation that may differ from their meanings in common parlance. For instance, the term *indirect costs* is sometimes used to refer to the cost of infrastructure such as building and core administrative staff, particularly in the context of university funding.

For purposes here, indirect costs are potentially factored into an economic evaluation in two ways:

1. Time inputs into an intervention such as waiting, treatment, and travel time. Such costs may be incurred by patients, their household, or other parties such as firms that employ patients. These enter into the evaluation specifically as costs.

2. Production gains to the economy resulting from improvements in health to patients. Confusingly, these are generally treated as benefits within an evaluation although it could be argued that such benefits are savings in disease costs that have been brought about by health gains. Again, such benefits are deemed to be to society rather than to any specific party. This approach to valuation is used in cost-benefit analyses and is typically labeled the "human capital approach."

Both aspects of evaluation have in common the problem of how to value a unit of time, whether it is time spent in accessing and receiving treatment or time gained as a result of improved health (through improved survival or improved functioning translating into increased work or leisure time). The issues considered in this entry are thus generally relevant to both aspects of evaluation although the focus will be on Item 1, given that the primary interest is in the assessment of costs.

The Valuation of Time Inputs

Economic theory suggests that any such measure should reflect the opportunity cost of the time input. In practice, this entails first identifying the nature of the displaced activity. Based on such a perspective, it is relevant to consider whether such activity is work or nonwork.

The valuation of the opportunity cost of work time is dependent first on whether output is replaced. In instances where it is not, the opportunity cost is set at the value of the marginal product of labor (e.g., if Fred takes a day off work, then the opportunity cost of that would be measured by his productivity of the previous workday). The value of the marginal product of labor is in turn dependent on a number of macroeconomic variables such as the level of competition in product markets and the presence of income and sales taxes. In general, the full wage rate, which is the gross wage plus other costs to the employer, is a good benchmark estimate for the marginal product of labor since employers will only incur such a cost if the value of additional output exceeds this cost. However, if there is any error caused by imperfect product markets and the presence of taxes, this benchmark will be rendered an underestimate.

Where production is replaced, then the opportunity cost of this time input is best estimated by the marginal cost of labor. This is proxied by the net wage rate (an individual's wage after taxes) and can be seen as an individual's reservation price for selling his or her labor, reflecting the marginal utility (or satisfaction) gained from leisure time balanced against the marginal (dis)amenity of work (the more unpleasant one finds work, the greater the take-home pay an individual would need to be remunerated to forgo leisure). The presence of involuntary unemployment will introduce

some error in this estimate, causing the reservation wage to be an overestimate of opportunity cost.

To determine the opportunity cost of nonwork time, a distinction needs to be made first between whether the individual is currently in paid employment or not. For those currently in paid employment, a proxy for this time input is the net wage (in spite of some possible error) since it reflects the marginal valuation of time for that person (for the same reason as stated in the previous paragraph).

For those not in employment and where activity, say housework, is not replaced (for instance, if it means there are certain household chores that ultimately do not get done), then the average wage of a housekeeper is a suitable proxy (the value of the foregone housework). Where it is replaced, that is, done at a later date, then the average net wage across all occupations would be a useful proxy recognizing that there will be some error depending on whether the individual is voluntarily or involuntarily unemployed. Table 1 summarizes the various proxies that are available for valuing the opportunity cost of time.

The friction cost method is an alternative to the valuation of the opportunity cost of time, although its relevance seems only to apply to work time. It essentially values the lost production from time off work by assuming that firms are able to make certain adjustments to absences, in both the short term and the long term, which will to some extent offset potential production losses. In the short term, factors such as the spare capacity within firms and the possibility of workers making up for lost production mean that the value of lost production from short-term absence is less than the marginal cost of labor as reflected in the full wage rate. This contrasts with the usual approach, which values lost production at the full cost to the firm, that is, the wage paid for the entire period of absence. Those advocating the friction cost approach have estimated that in the short term, the friction cost represents 80% of the full cost of labor. In the longer term, beyond what is known as the friction period, which is based on the average amount of time a particular labor market is able to fill vacancies caused by illness, the costs are deemed to be zero. One criticism of the friction cost approach is that by ignoring leisure time, it implicitly values it at zero. Its advocates have argued that the value of leisure time is instead factored into health outcomes of an economic evaluation (such as quality-adjusted life years) and therefore need not be taken into account in the analysis of costs.

Issues Around the Inclusion of Indirect Costs

The inclusion of indirect costs into an economic evaluation enables a societal perspective, thereby allowing for a complete picture of the resource implications associated with an intervention. A societal perspective is consistent with the notion of social welfare maximization underlying cost-benefit analysis, that is, the maximization of the well-being of individuals within society.

A limitation of narrower perspectives to evaluation is that by definition they do not account for costs that fall outside the organizations or health systems on which they are based and thereby implicitly encourage cost-shifting. For instance, a regulatory agency that evaluates new technologies for public subsidy based strictly on a health sector perspective tends to favor technologies that cost-shift onto households (such as to carers) and other sectors of the economy. Aside from the equity implications of potentially adding to hardships experienced by households already faced with illness, these narrower perspectives can fail to distinguish between new technologies that are genuinely cost-effective and those that simply shift costs away from the health sector.

Table 1 Summary of proxy measures for measuring opportunity cost of time

	Paid Work Time	*Nonpaid Work Time*
When outputs/activities are replaced	Net wage	Net wage for *employed* individuals Average net wage for *unemployed individuals*
When outputs/activities are not replaced	Gross wage	Wage of housekeeper

The argument against adopting a societal perspective that will enable the inclusion of indirect costs is that such a perspective is often not relevant to decision making in the health sector. Decisions are generally made by organizations geared toward specific secular interests, and therefore evaluation is required to reflect these. For instance, a health sector perspective is often adopted simply because ministries of health are generally not accountable for the downstream implications of healthcare on employment, social services, schooling, and so on. Based on this line of argument, the merits of an evaluation tool ultimately lie simply in the ability of such a tool to match the objectives of those making decisions rather than its comprehensiveness.

One of the difficulties of including indirect costs is the complexities inherent in their measurement and valuation. As highlighted above, the value of time inputs needs to reflect the opportunity cost of that time. The fundamental problem with the empirical analysis of opportunity cost is that it is not directly observable. Ultimately, it needs to be implied from a number of indicators such as employment status of the individual and the nature of the product and employment markets in which the evaluation is taking place. In practice, this opens up certain ambiguities in the methods for estimating time costs, in measuring both the time inputs (e.g., where there is joint production) and their subsequent valuation.

There are also strong equity implications in the way in which indirect costs are generally measured. Because wage rates are used as markers of opportunity cost, the time costs of high-income earners are generally valued more highly than the time costs of low-income earners. This means technologies that benefit the wealthy tend to be favored over those that benefit the poor. This has been used as a further argument for their exclusion from economic evaluation.

The Decision-Making Context and the Inclusion of Indirect Costs

This entry explores the arguments around the inclusion of indirect costs in economic evaluation. Pivotal is the adoption of a societal perspective consistent with the welfare principles underlying cost-benefit analysis. Nevertheless, in healthcare, such costs often tend to be excluded from analysis, probably because of their lack of relevance to the perspective taken in evaluation and also the equity implications around their valuation. Ultimately, the merits of whether to factor in such costs in evaluation need to be judged pragmatically and based on whether such an approach is consistent with the specific policy questions under consideration.

Stephen Jan

See also Cost-Benefit Analysis; Costs, Opportunity; Costs, Out-of-Pocket

Further Readings

Brouwer, W. B., & Koopmanschap, M. A. (2005). The friction-cost method: Replacement for nothing and leisure for free? *PharmacoEconomics, 23*, 105–111.

Johannesson, M., & Karlsson, G. (1997). The friction cost method: A comment. *Journal of Health Economics, 16,* 249–255 (Discussion pp. 257–259).

Koopmanschap, M. A., Rutten, F. F., van Ineveld, B. M., & van Roijen, L. (1995). The friction cost method for measuring indirect costs of disease. *Journal of Health Economics, 14,* 171–189.

Liljas, B. (1998). How to calculate indirect costs in economic evaluations. *PharmacoEconomics, 13,* 1–7.

Olsen, J. A., & Richardson, J. (1999). Production gains from health care: What should be included in cost-effectiveness analyses? *Social Science and Medicine, 49,* 17–26.

Posnett, J., & Jan, S. (1996). Indirect cost in economic evaluation: The opportunity cost of unpaid inputs. *Health Economics, 5,* 13–23.

Sugden, R., & Williams, A. (1978). *The principles of practical cost-benefit analysis*. Oxford, UK: Oxford University Press.

Costs, Fixed Versus Variable

Costs refer to the economic input required to achieve a certain outcome, that is, the amount one spends to produce a service or a product, or the value imputed to a resource. Costs are distinguished from *charges,* which are the prices of services and do not reflect the actual costs of all

inputs. Costs are usually divided into fixed and variable costs. With regard to healthcare, *fixed costs* are expenses that do not vary with physician care decisions or treatment—such as rent, salaries, mortgage payments, and fire insurance—and that do not vary with the level of patient activity, or products, and once sunk, they cannot be easily recovered. They are also called *sunk costs* because they are beyond the control of the entrepreneur. Other types of costs such as wages of production workers or doctors, medical supplies, drugs, electric power to run machines, and bed-days change with the number of patient visits or products offered for sale. These are called *variable costs*.

In a world of limited healthcare resources, medical decision makers must make challenging management decisions. Without a systematic evaluation of benefits of health interventions or programs in relation to their costs, it is difficult to make rational and sound judgments. This entry reviews key elements related to identification, measurement, and valuation of costs.

Identification

By identifying and controlling all relevant costs, healthcare managers are better able to earn a profit and be successful. Fixed costs are those that generally do not vary between payment intervals. Generally, these costs cannot be altered on a short-term basis because of contractual agreements. Variable costs are those that increase with increasing units of service. For example, an increase in the number of patient visits would result in the use of additional materials, extra labor, and wages. One way to determine fixed costs is to consider the expenses that would continue to be incurred if a healthcare facility were to be temporarily closed and no patients were to be treated. In this case, rent, fees, and loan payments would still be due. They generally do not change with increases or decreases in facility activity. It is important to note that fixed costs are unvarying only within a certain range of facility activity. For example, if the facility activity grows enough to require additional space or additional employees, the fixed costs associated with rent or salaries will change as well. Variable costs are those that change as the level of facility activity changes. Examples of the variable costs within a healthcare facility would be supplies used for each patient visit, and wages for hourly, part-time employees. These costs are driven primarily by the facility's activity and would stop only if the facility were to close for a period of time, such as a month. Once the difference between fixed and variable costs is understood, it is important to know how to distinguish one from the other. For instance, consider a clinic that has fixed costs of $3,800 and variable costs of $7 per patient. To cover its monthly expenses, the clinic would have to earn $3,800 in fees plus $7 per patient treated. If the clinic had only one patient visit per month, it would have to charge $3,807 for that one treatment to cover its fixed and variable costs! If the practice had 1,000 patient visits during the month, its total costs would be $10,800 ($3,800 in fixed costs plus 1,000 patient visits at $7 each). Therefore, this clinic would only have to charge $10.80 per patient visit to cover its fixed and variable costs. This example illustrates that the amount of fixed costs that each patient visit must cover depends on the total number of patient visits across which these fixed costs are to be spread.

For example, for cost control purposes, it is possible to determine a flexible budget using a formula expressed as a linear equation in which the slope is the variable cost per unit (or per direct labor hour). Graphically, this would appear as shown in Figure 1.

By definition, fixed costs do not change with the level of activity. As a result, the budget for cost control purposes would be displayed graphically as shown in Figure 2.

It is important to note that the costs to be included depend on whose perspective is being used and on the question of *whose costs matter?* The view can be that of the healthcare facility, the insurance company, the patient, or society. They are not interchangeable. An action that reduces facility cost, such as early discharge, may increase the cost to the patient or insurance company by, for example, the need to pay for home healthcare or a stay at an extended-care facility. If the societal perspective is adopted, then all costs must be considered. If the perspective is that of the facility, costs such as patient and caregiver time would be excluded since they are not part of the facility's financial responsibility.

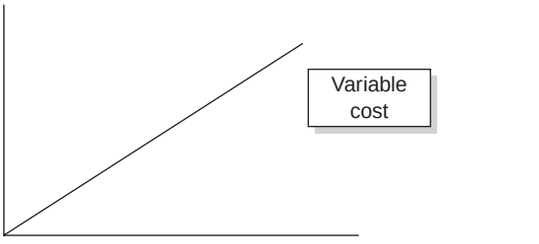


Figure 1 Variable costs

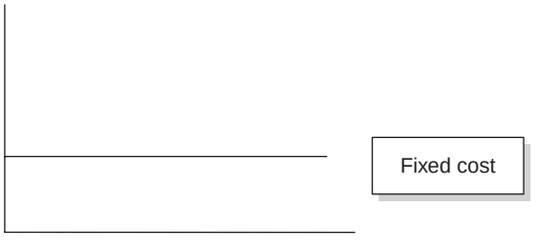


Figure 2 Fixed costs

Measurement

The measurement of costs is similar regardless of the type of analysis being undertaken. Measurement refers to the resource changes included in the analysis. Resources consumed can be divided in a number of different ways. Typically, these will be amounts of labor inputs or outputs but may also include patients' time.

Valuation

The most accurate method of cost estimation is that known as *microcosting,* in which every resource use is identified, measured, and quantified into a unit cost. Microcosting refers to detailed analysis of the changes in resource use due to a particular intervention, like time-and-motion studies. Although many analysts favor microcosting, it tends to be costly. *Gross* or *top-down costing* allocates a total budget to specific services such as hospital stays or doctors' visits. The simplicity of top-down costing may be offset by a lack of sensitivity, which in turn depends on the type of routine data available. The choice between microcosting and gross costing depends on the needs of the analysis. It is common to use substitute proxies for cost, such as Medicare or Medicaid reimbursement. This method has the advantage of using a nationally relevant estimate as opposed to a single facility's cost. Another popular technique is to start with a facility's charges and then multiply them by an adjustment called the cost-to-charge ratio. Although the cost-to-charge ratio is convenient, it is usually available only for a facility and not for an intervention or diagnosis.

There are two main limitations to conducting a cost study:

1. Costs may vary from one facility to another. They have different purchasing contracts for goods and services. Different staffing levels affect marginal costs and the labor component of variable costs. These may affect the generalizability of the results and may need confirmation before each facility implements changes.
2. The facility may have old costs listed by the accounting system that have not been updated to reflect current market conditions, leading to inaccurate results.

Economic Evaluation

To carry out an economic analysis alongside a study, a researcher can do the following:

- Collect information on the costs and the effectiveness of the alternative interventions from patients in all arms of the trial
- Identify and measure resource volumes, for example, drug quantities for every individual trial patient
- Attach unit costs to each resource item to obtain a mean cost per patient per arm of the trial
- Combine mean patient costs with mean effectiveness measures from the trial to establish the cost-effectiveness of each alternative

Finally, cost studies are typically divided into cost minimization, cost benefit, cost utility, and cost-effectiveness. Cost-minimization studies compare at least two equally effective therapies to find the least expensive. Cost-benefit studies call for converting all outcomes (pain, emesis, renal failure, myocardial infarction, death, etc.) to a

monetary value. Cost-utility studies establish the price of a utility metric for each quality-adjusted year of survival. Cost-effectiveness studies decide the cost of avoiding undesirable outcomes (death, ventilation-associated pneumonia, etc.). Suggestions on carrying out cost-effectiveness studies have been disseminated by the U.S. Public Health Service and the European Society of Intensive Care Medicine.

Catherine Kastanioti

See also Cost-Benefit Analysis; Cost-Effectiveness Analysis; Cost-Minimization Analysis; Costs, Direct Versus Indirect; Costs, Semifixed Versus Semivariable; Cost-Utility Analysis

Further Readings

Armstrong, R. A., Brickley, M. R., Shepherd, J. P., & Kay, E. J. (1995). Healthy decision-making: A new approach in health promotion using health state utilities. *Community Dental Health, 12,* 8–11.

Donaldson, C. (1990). The state of the art of costing health care for economic evaluation. *Community Health Studies, 14,* 341–356.

Drummond, M. F., & Davies, L. (1991). Economic analysis alongside clinical trials: Revisiting the methodological issues. *International Journal of Technology Assessment in Health Care, 7,* 561–573.

Drummond, M., & McGuire, A. (2001). *An economic evaluation in health care: Merging theory with practice.* Oxford, UK: Oxford University Press.

Drummond, M., O'Brien, B., Stoddart, G., & Torrance, G. (1997). *Methods for the economic evaluation of health care programmes* (2nd ed.). Oxford, UK: Oxford University Press.

Engoren, M. (2004). Is a charge a cost if nobody pays it? *Chest, 126*(3), 662–664.

Gold, M. R., Gold, S. R., & Weinstein, M. C. (1996). *Cost-effectiveness in health and medicine.* Oxford, UK: Oxford University Press.

Hunink, M., & Glasziou, P. (2001). *Decision making in health and medicine: Integrating evidence and values.* Cambridge, UK: Cambridge University Press.

Meltzer, M. I. (2001). Introduction to health economics for physicians. *Lancet, 358,* 993–998.

Robinson, R. (1993). Costs and cost-minimisation analysis. *British Medical Journal, 307,* 726–728.

Robinson, R. (1993). Economic analysis and health care. What does it mean? *British Medical Journal, 307,* 670–673.

Costs, Incremental

See Marginal or Incremental Analysis, Cost-Effectiveness Ratio

Costs, Opportunity

The notion of opportunity cost is one of the fundamental concepts of economics. If resources are limited, then there is a choice to be made between desirable, yet mutually exclusive, results. The true or opportunity cost of one alternative is the benefit foregone from not being able to have the next best alternative. The concept has been encapsulated in the truism that "there's no such thing as a free lunch," meaning that things that appear free are always paid for in some way.

The estimation of the opportunity cost of a policy will almost certainly vary depending on the person or persons who are doing the assessing. For example, if a health authority is considering building and staffing a new hospital from public sector funds, the opportunity cost might be the benefit that could have been obtained by increasing or improving facilities and staffing at neighbouring healthcare providers. From this point of view, the choices might be represented as between different policies with the aim of maximizing health, given the funds available to the health authority. However, from the point of view of the public sector as a whole, the opportunity cost might be that this money could have been used to improve the criminal justice system. From this point of view, the choice or trade-off is between improved health and improved criminal justice. From a wider, societal perspective, the evaluation might consider that investment by public services might in some circumstances displace investment by the private sector.

Opportunity cost is a wider concept than accounting or monetary cost. Accounting cost attempts to value the outcomes and resources used in a program or policy at their monetary cost or price. However, not all the outcomes and resources of the policy may have a monetary cost, or there may be no market in the good or service with which to value that outcome, or the price may be

thought to omit some important aspect of the benefits or costs of the good or service. If these benefits and costs fall on third parties, they are known as externalities. As an example, consider the use of an intensive care bed after a surgical operation. The monetary value of an hour of care in that bed that is charged to the patient's health insurance or health authority might be calculated as the sum of the hourly salaries of the medical and nursing staff attending the patient, the use of consumables and drugs, and the overheads of the hospital. However, in many hospitals, intensive care facilities are very scarce. The use of this facility by a patient might mean that another patient's planned operation must be postponed until a bed becomes spare, in case it is needed. In these circumstances, the opportunity cost of use of the bed by one patient might be considered in terms of the inconvenience, risks, and costs of cancelling another person's operation. Irrespective of whether resources are allocated to healthcare services by a market mechanism or by a government ministry, they must be valued at their true or opportunity cost if society is to invest its resources efficiently.

One important opportunity cost that is often omitted from decisions about resource allocation in healthcare is the cost of capital. In some countries, public-sector hospitals are owned by one organization, such as a municipality or local government, and managed by another one, such as a health authority. The owner of the facility may consider the cost of the land and building a *sunk cost,* that is, one that has been made in the past and cannot be recovered. This can result in inefficient use of resources (e.g., low bed occupancy rates or underused wards) if the management of the hospital does not pay a rent that appropriately takes account of the alternative use of the land, building, and working capital tied up in the hospital, and this ensures that these costs are reflected in the prices charged to the healthcare purchaser or third-party payer.

People's preference for benefits now rather than in the future is another form of opportunity cost. In this case, the discount rate is a means of adjusting future benefits and costs to current values. Many policies that have an impact on health, especially preventive policies that aim to reduce the risk of future illness, require an immediate investment but might not generate benefits for many years. In these cases, the choice of discount rate can be highly influential in determining whether present-value benefits exceed costs.

The evaluation of the opportunity cost of a policy necessarily implies that all the outcomes of all the feasible alternative policies can be assessed on some common scale. The concept of utility is convenient to measure the relative satisfaction from or desirability of different goods, services, or outcomes. If more than one person is affected by a policy, the utilities of all those persons must be aggregated and compared somehow. An evaluation of alternative health service policies might take into account the effect on health, on work, and on leisure, or the quality of the care provided. These examples illustrate that opportunity cost is a normative concept, that is, it requires an element of subjectivity to decide which benefits and costs to value and the weight that should be given to each type of benefit. A number of conceptual frameworks have been developed to maintain scientific rigour in an evaluation of health technologies. Cost-benefit analysis aims to evaluate all benefits and costs in monetary terms, making use of methods such as contingent valuation or hedonistic pricing to identify people's willingness to pay for each type of benefit, including health. Cost-utility analysis attempts to cut through the debate surrounding the difficulty of valuing different kinds of benefits on a common scale by assuming that health is the only benefit to be valued and that the health of the population is simply the sum of the health of the individuals in it. The Panel on Cost-Effectiveness in Health and Medicine (Gold Report) recommended that all health technology assessments should include a reference case to ensure as far as possible that evaluations by different authors include a common set of outcomes valued by comparable techniques.

The concept of opportunity cost relies on the idea that benefits in one dimension can only be obtained by sacrificing other desirable outcomes. This trade-off implies that the economy is at a point of productive efficiency, that is, it is only possible to produce more of one type of good by diverting resources from the production of another good. However, a principle of Keynesian macroeconomics is that, in some circumstances, there can be underemployed resources. The United

Nations Commission on Macroeconomics and Health assembled considerable evidence that lack of health and education are both a cause and a consequence of enduring poverty. This suggests that policies that tackle the health and education of the poor may be an important lever with which to increase productivity and generate economic growth with benefits for society as a whole.

David Epstein

See also Contingent Valuation; Cost-Benefit Analysis; Cost-Utility Analysis; Discounting; Efficient Frontier; Reference Case; Utility Assessment Techniques; Willingness to Pay

Further Readings

Gold, M. R., Siegel, J. E., Russell, L. B., & Weinstein, M. C. (1996). *Cost effectiveness in health and medicine*. Oxford, UK: Oxford University Press.

WHO Commission on Macroeconomics and Health. (2001). *Macroeconomics and health: Investing in health for economic development*. Geneva: World Health Organization.

Costs, Out-of-Pocket

For most goods and services, the full price is borne by consumers. However, for healthcare services, third-party payers (e.g., government programs or private insurers) typically make partial or full payments on the consumer's behalf. Therefore, the amount paid out-of-pocket (OOP) by consumers represents only a fraction of the full payment received by the providers of services.

The fundamental purpose of imposing cost-sharing requirements on consumers is to control moral hazard (use of services beyond the quantity at which marginal benefit equals marginal cost). Although a risk-averse consumer would prefer full coverage (no OOP obligations) in the first best situation, the first best is generally not attainable because fully insured individuals have an incentive to use care until marginal benefit is zero. These low-benefit services will increase the cost of insurance or the burden on public finance without creating sufficient value to justify the extra cost. Imposing OOP obligations on consumers trades some risk spreading for the preservation of a partial incentive for the consumer to consider the cost of the chosen services relative to their expected value.

Determinants of Out-of-Pocket Prices

The gap between the total price paid for a service and the OOP price faced by the consumer is a function of the basic provisions with respect to patient obligations contained in the public payment policy or the private health insurance contract under which third-party payments are made. Such provisions are often complex, including deductibles (consumer is fully responsible for the first specified amount of spending during a time period), co-payments (consumer is responsible for a fixed payment for each unit of service received once the deductible has been satisfied) or co-insurance (consumer is responsible for a fixed percentage of the price of each unit of service received once the deductible has been satisfied), and stop-loss (consumer is fully insured for additional services once a prespecified, maximum OOP expenditure has been exceeded during a time period).

In addition to these basic policy provisions, the OOP price to the consumer can also be modified by a number of other factors. Third-party payers often place a variety of restrictions on coverage that can directly or indirectly change consumers' OOP obligations. These include service-specific limits (e.g., maximum number of visits allowed to a certain type of provider during a time period) and overall limits (e.g., lifetime maximum expenditures), after which the consumer will face the full price of additional services. In addition, coverage for some services may be denied if specific requirements are not met (e.g., approval of the service by a "gatekeeper" physician or by a third-party payer's pre-authorization or use review process). Third-party payers also often specify whether or not providers can engage in *balance billing*. Suppose the third-party payment plus the patient's contractual obligation as determined by the provisions discussed above (e.g., co-payments) falls short of the provider's charges. If the provider is allowed to balance bill, the consumer's OOP obligation would increase by the excess of the provider's charges above the amount paid by the third-party payer and the consumer's co-payment obligations. Finally, third-party payers

may distinguish between preferred or nonpreferred providers (in-network vs. out-of-network) or treatments (e.g., generic vs. brand name pharmaceuticals) by obligating consumers who elect nonpreferred providers or treatments to pay a higher OOP price. The recent trend toward value-based insurance design operates analogously, identifying classes of patients who may be exempted from OOP obligations for specified services deemed clinically valuable (e.g., diabetes patients may be exempted from insulin co-payments).

Because some of these provisions are based on the use of services over a period of time (e.g., deductibles, service limits), consumer decision making also has an important dynamic aspect. Using an annual deductible for purposes of illustration, the consumer's actual OOP price for any given service can deviate from the rationally anticipated OOP price. Suppose a consumer with a chronic illness knows that he or she has a very high probability of exceeding his or her deductible during the year. Consuming an extra service early in the year will cause the consumer to satisfy his or her deductible sooner, thereby creating an implicit "discount" on a service that will be consumed later in the year. Conversely, a consumer who has not satisfied the deductible late in the year would be unlikely to obtain such an implicit discount by consuming an additional service. More generally, the anticipated OOP price today depends on use earlier in the year, and the anticipated OOP later in the year depends on the use decisions made today.

Richard A. Hirth

See also Dynamic Decision Making; Economics, Health Economics; Pharmacoeconomics; Value-Based Insurance Design; Willingness to Pay

Further Readings

Hirth, R. A., Greer, S. L., Albert, J. M., Young, E. W., & Piette, J. D. (2008). Out-of-pocket spending and medication adherence among dialysis patients in twelve countries. *Health Affairs, 27*(1), 89–102.

Manning, W. G., Newhouse, J. P., Duan, N., Keeler, E. B., Leibowitz, A., & Marquis, M. S. (1987). Health insurance and the demand for medical care. *American Economic Review, 77*(3), 251–277.

Costs, Semifixed Versus Semivariable

Consideration of costs is an important factor in medical decisions, including budgeting and planning, pricing for healthcare products or services, operational control, and selection of therapeutic options. Costs may be viewed in different ways. One approach to describe costs is a cost behavior pattern in which a cost is analyzed by its reactions to different levels of activity. Understanding the cost behavior patterns will facilitate medical decision making.

Two common types of cost behavior patterns are fixed and variable costs. Fixed costs remain constant over different levels of activity (e.g., volume, workload). Variable costs vary with changed levels of activity, such as costs for medications and medical supplies, which represent a major part in healthcare. In some cases, neither fixed costs nor variable costs alone can fully describe cost behavior patterns. Semifixed or semivariable costs are conceptually used as other types of cost behavior patterns. Semifixed or semivariable costs contain a portion of fixed costs and another portion of variable costs. Eventually, all costs can be properly explained by different combinations of fixed costs and variable costs. Semifixed and semivariable costs are explained as follows.

Semifixed Costs

Semifixed costs are also called stepped, stepped-fixed, step-variable, step-fixed, or step-function costs. This type of cost remains a constant within a particular range of activity and sharply changes after exceeding the threshold of this range, and then again remains constant during another range of activity. In other words, semifixed costs could be viewed as a combination of multiple fixed costs in which each has a much narrower relevant range. If semifixed costs are plotted against levels of activity, the pattern of semifixed costs looks like steps. Figure 1 illustrates a semifixed cost that increases with increased level of activity. The activity range may be different within each step, and the overall change varies with increased level of activity.

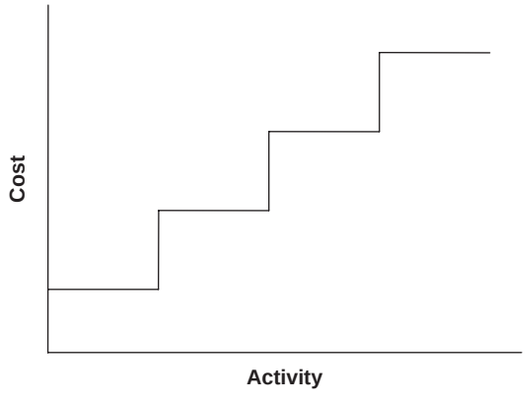


Figure 1 An example of semifixed costs

An example of semifixed costs is the total staff cost for pharmacists. Suppose a pharmacist can handle a maximum of 50 prescriptions per day. Accordingly, a pharmacy needs 10 pharmacists to deal with 451 to 500 prescriptions or 11 pharmacists for 501 to 550 prescriptions. Similar examples include medical or nursing staff costs, administration costs, information technology costs, and equipment maintenance costs.

Semivariable Costs

Semivariable costs are sometimes called mixed costs. This type of cost contains a portion of fixed costs, and the remaining portion varies with an increased level of activity. Semivariable costs can be further classified as linear or nonlinear semivariable costs, and the classification of patterns depends on the relation of the variable portion to the change of activity. Typical figures of semivariable cost patterns are shown in Figure 2. Note that the total cost line does not pass through the origin because there is a fixed cost component.

An example of linear semivariable costs is laboratory costs. For a diagnosis test, the device cost and the annual maintenance cost are fixed, and the total cost of test strips varies with an increased number of tests. A utility cost may be an example of a nonlinear semivariable cost. A basic monthly fee is charged regardless of the amount of the utility used, and an additional charge increases with increased use of the utility; however, the rates may vary with an increased amount of use. Other examples of semivariable costs include car rental (fixed vehicle

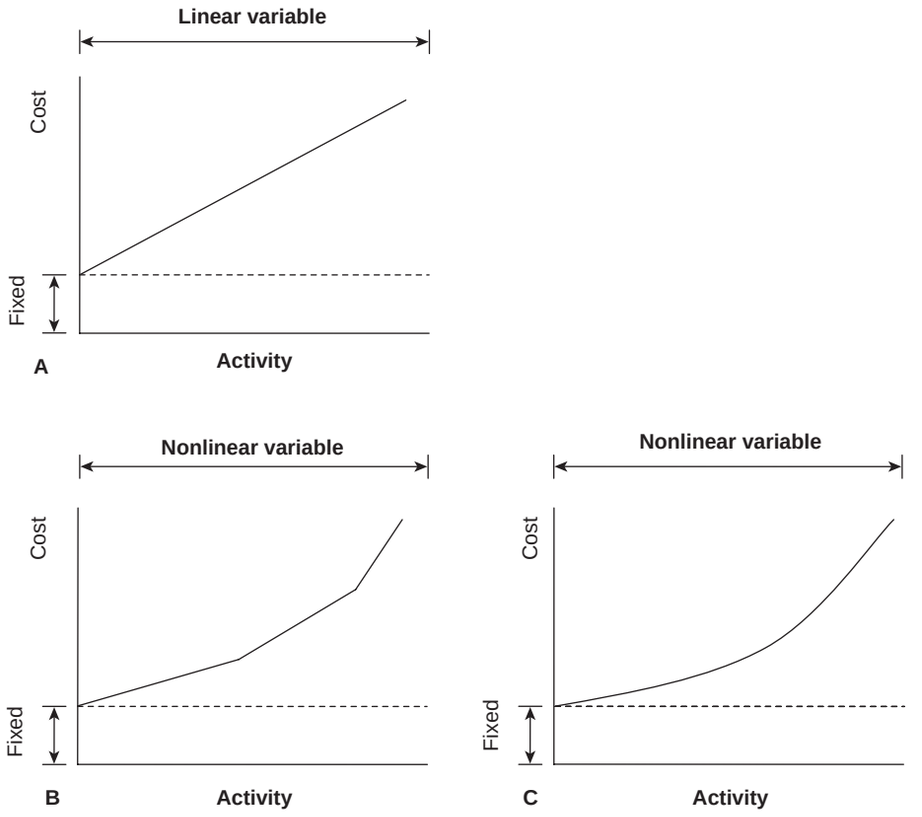


Figure 2 Examples of semivariable costs: linear (A) and nonlinear (B, C)

rental fee plus variable costs for fuel and mileage) and facility costs (fixed rental or maintenance costs plus various total utility costs in general).

Costing Methods

Costs related to medical decisions may evolve in many types of cost behavior patterns, and costs usually require further analysis to provide information for medical decisions. There are at least three methods to analyze cost behavior patterns: top-down, bottom-up, and graph analysis. The top-down method breaks down aggregated cost data into smaller pieces from higher levels (e.g., hospital costs) to lower levels (e.g., departmental costs) based on a principle of allocation. This method is often used for retrospective data, and it is often difficult to break down data to the individual level (e.g., patients). The bottom-up method uses cost data from the individual level and then adds up all costs to the total costs. This method can use either retrospective or prospective data, and patient-level data of use can be further analyzed. Therefore, the bottom-up method is frequently used in economic analyses. However, this method bears several limitations, including difficulty in obtaining sensitive personal data (e.g., payment), so a proxy (e.g., hospital charge) is often used. The graph method is to plot costs against activity levels, as shown in the figures. This method allows investigators to evaluate or present cost behavior patterns in summarized data, but detailed cost information will not be available.

Jun-Yen Yeh

See also Costs, Direct Versus Indirect; Costs, Fixed Versus Variable; Costs, Opportunity; Sunk Costs

Further Readings

Cleverly, W. O. (1997). Cost concepts and decision making. *Essentials of health care finance* (4th ed., pp. 223–227). Gaithersburg, MD: Aspen.

Gyldmark, M. (1995). A review of cost studies of intensive care units: Problems with the cost concept. *Critical Care Medicine, 23*(5), 964–972.

Jegers, M., Edbrooke, D. L., Hibbert, C. L., Chalfin, D. B., & Burchardi, H. (2002). Definitions and methods of cost assessment: An intensivist's guide. *Intensive Care Medicine, 28*(6), 680–685.

Lere, J. C. (2000). Activity-based costing: A powerful tool for pricing. *Journal of Business & Industrial Marketing, 15*(1), 23–33.

Lubasky, D. A. (1995). Understanding cost analyses: Part 1. A practitioner's guide to cost behavior. *Journal of Clinical Anesthesia, 7*(6), 519–521.

Costs, Spillover

When producing, selling, buying, or consuming a good or service affects people other than those directly involved in the market exchange—for example, when a factory emits smoke that pollutes the air breathed by those in the vicinity—the economic activity is said to "spill over" and impose costs (or confer benefits) on people other than those directly involved in the transaction. Economists call spillover effects *externalities*. The extent of spillover effects in healthcare is one of several features that contribute to the failure of private markets to achieve efficient results and health-related outcomes relative to their costs. Externalities serve as one rationale for public-sector involvement in healthcare.

Spillover effects must be taken into account when evaluating the impact of healthcare services, their financing, and their delivery in cost-benefit or cost-effectiveness analyses and when making decisions about how healthcare resources should be invested. The level and distribution of health status and longevity within a population can also have economic impacts (both financial and in terms of well-being) beyond the individual level.

Classic examples of spillover effects in public health and health services are the transmission and control of communicable diseases and immunization, cases for which untreated disease or services to the individual have costs or benefits for others. Because communicable diseases impose costs (spread of disease to others) beyond those borne by the individual infected, the willingness of the individual to pay for the disease's prevention or treatment may be less than the total value to the community of taking action to prevent the spread of the disease. This circumstance justifies public provision or subsidy of services to prevent disease transmission.

Another spillover effect in healthcare stems from the value individuals place on others' access

to and use of needed healthcare. In this case, the rationale for public financing or provision of healthcare is that the welfare of individuals who are taxed or otherwise support health services provided to unrelated others is increased because this subsidy satisfies a moral sentiment of altruism or justice. The provision of healthcare or health coverage to others in a community (local or national) is deemed a "merit good."

The ability to identify and measure spillover effects of an activity or policy depends on what is encompassed by an analysis and the perspective that is adopted in conducting it. A study of the overall use of medical services by Medicare end-stage renal disease (ESRD) patients exemplifies this point. Providers of routine dialysis services receive a capitation payment from Medicare for outpatient services only. The cost of each dialysis treatment depends on its intensity, measured in terms of the rate of urea removal during the procedure. Avi Dor found that less intensive dialysis treatments resulted in higher rates of hospital admissions among Medicare ESRD patients. From the perspective of the outpatient dialysis provider, the cost of lower intensity of outpatient treatments on hospitalization rates accumulates—this cost is manifested by the outpatient provider and is most pronounced in the worse health of patients receiving the less intensive treatment and higher overall Medicare costs for ESRD care. The full impact of the Medicare outpatient ESRD capitation rates and provider treatment decisions can be captured only when a broader analysis is undertaken.

An ethical conundrum that clinicians face is whether they should take costs into account when making treatment choices and recommendations for their patients, given that spending on the care of those with coverage may indirectly make it more difficult politically and fiscally to extend healthcare coverage to those without it. This constitutes a spillover cost of individual doctor-patient transactions. Christine Cassel and Troyen Brennan argue that physicians share a "medical commons" and that they should be accountable for how resources devoted to health services are managed. In prepaid group practices and in fixed-budget national healthcare systems, physicians' ethical duties to individual patients are linked with a shared responsibility for a community's resources. For the larger U.S. healthcare enterprise, however, the full cost implications of individual treatment choices are not internalized.

Just as the costs of care of people with coverage affect those who are uninsured, uninsurance in a community can affect those who have coverage. The first systematic look at the spillover costs of uninsurance was undertaken by the Institute of Medicine in a study published in 2003. Although healthcare access problems related to lack of coverage are most severe for people who are uninsured, other vulnerable population groups (Medicaid enrollees, low-income inner-city residents, members of racial and ethnic minority groups) who tend to rely on the same care providers (e.g., public clinics, hospital outpatient departments) experience reduced access to care in communities with high uninsurance rates due to crowding and provider instability because of high uncompensated care burdens. Furthermore, communities with higher-than-average uninsurance rates tend to have fewer specialized hospital services such as trauma, psychiatric, or burn units than communities with relatively low rates of uninsurance. This reduced access to a variety of health services experienced across a community is a spillover cost of high uninsurance rates.

As noted by Jeremiah Hurley in his overview of the economics of the health sector, spillover effects have been the subject of much theoretical discussion and far less empirical analysis in the field. Doing a better job of capturing spillover effects will require both a wider-angle lens when focusing on a subject and ingenuity and persistence in acquiring the kinds of data that reveal these effects. The consequence of measuring spillover impacts will be more complete information for policy choices.

Wilhelmine Miller

See also Cost-Benefit Analysis; Cost-Effectiveness Analysis; Economics, Health Economics

Further Readings

Cassel, C. K., & Brennan, T. E. (2007). Managing medical resources: Return to the commons? *Journal of the American Medical Association, 297,* 2518–2521.

Dor, A. (2004). Optimal price rules, administered prices and suboptimal prevention: Evidence from a Medicare program. *Journal of Regulatory Economics, 25,* 81–104.

Hurley, J. (2000). An overview of the normative economics of the health sector. In A. J. Culyer & J. P. Newhouse (Eds.), *Handbook of health economics* (pp. 56–118). Amsterdam: Elsevier.

Institute of Medicine Committee on the Consequences of Uninsurance. (2003). *A shared destiny: Community effects of uninsurance*. Washington, DC: National Academies Press.

Rice, T. (1998). *The economics of health reconsidered*. Chicago: Health Administration Press.

Cost-Utility Analysis

Cost-utility analysis is a special form of cost-effectiveness analysis where the health outcomes are measured in terms of a preference-based *utility* measure. Like cost-effectiveness analysis, this yields an outcome of the evaluation that is expressed in terms of a cost per unit effect. However, in contrast to cost-effectiveness analysis, provided this measure is considered to be a generic measure of health, then cost-utility analysis is sufficient to efficiently allocate resources from a fixed healthcare budget in terms of maximizing the health achievable from those fixed resources. Of crucial importance is the validity of the preference-based utility measure as a generic measure of health outcome that can be used to compare the allocation of resources across disease areas.

A number of different candidate utility measures have been proposed. Most popular are the disability-adjusted life year (DALY), which has been used extensively by the World Health Organisation (WHO) to compare the burden of disease between countries (particularly in the developing world), and the quality-adjusted life year (QALY), which is widely used in developed countries, such as Australia, Canada, the United Kingdom, and the United States. Other measures, such as the healthy year equivalent (HYE), have not gained widespread acceptance despite apparently addressing some of the acknowledged problems in the other measures.

All the measures have the same fundamental goal—to represent the two dimensions of health, morbidity and mortality, in a single measure that represents the value of the underlying health state in a way that can be validly compared across disease areas. The QALY does this by weighting length of life by a health-related quality-of-life measure. The QALY is simply the area under this quality-adjusted survival curve, and the QALY gained from a treatment under evaluation is estimated as the difference between two quality-adjusted survival profiles representing the treatment under evaluation and the relevant alternative treatment.

The accurate measurement of mortality presents few challenges due to the definitive nature of the health outcome. However, the measurement of health-related quality of life is far more controversial. The health-related quality-of-life measure, to be suitable for quality adjusting life years, must represent a preference for health on a cardinal ratio scale (such that an improvement of 0.2 is twice as good as an improvement of 0.1) that is anchored at the top end by the value of 1 for perfect health and where 0 represents death. Negative values are allowed and represent health states worse than death.

The accurate assessment of health-related quality-of-life utility has become a major research area. Direct utility assessment methods involve asking patients or lay populations to provide a value for a specific health state—often presented to the respondent in the form of a vignette. Popular utility elicitation instruments include the standard gamble, time trade-off, and person trade-off techniques. In recent years, much debate has centered on whether it is patients or lay populations who should form the respondent base for utility assessments. Advocates of the patient-based approach cite the experience of patients as the principal advantage, while advocates of asking lay populations cite the role of the layperson as taxpayer and potential patient in publicly funded systems and suggest that patients may provide strategic responses if they realize that their values are being used to allocate resources. A popular compromise in recent years has been the use of health-related quality-of-life instruments such as the EQ-5D (EuroQol) and Health Utility Index, which are generic descriptive systems for health. These are suitable for use with patients to map into the descriptive system with tariff utility values assigned from large-scale population surveys.

The avoidance of placing a monetary value on health, as is required in cost-benefit analysis, is seen as a practical advantage by many for whom

monetary valuation of health is seen as distasteful. Proponents of cost-benefit analysis typically criticize the lack of theoretical foundation for cost-utility analysis, whereas proponents of the approach have claimed it embodies its own justification on the grounds of equitable treatment of health outcomes across individuals, coining the term *extrawelfarism* to describe the ethic embodied in the approach. Nevertheless, monetary valuation in cost-utility analysis cannot be avoided. In the face of a fixed budget constraint, utility maximization requires ranking of interventions by cost-utility ratio with healthcare interventions adopted in order of ascending cost-utility ratio until the budget is exhausted. The cost-utility ratio of the last program funded "reveals" the (shadow) price (willingness to pay) for a unit of health outcome implied by the budget constraint.

A more practical approach to allocating resources has been to consider a "threshold" value of a unit of health output above which a program or treatment would not be funded. Although such an approach has been criticized for failing to recognize the budget constraint (and therefore encouraging uncontrolled healthcare expenditure), the use of arbitrary threshold values as a decision-making rule of thumb is widespread.

The concept of a decision-making threshold has had an important influence on the analysis of cost-utility (and cost-effectiveness) studies by encouraging the use of the net-benefit approach to decision making. By translating health outcome into a monetary value, it is possible to analyze the overall net benefit of a program or intervention conditional on the threshold value. This has led some commentators to question whether there is a practical difference between cost-benefit and cost-effectiveness/utility analyses. Nevertheless, it is important to recognize that in the presence of a fixed budget constraint, net-benefit decision making will not necessarily lead to optimal allocation of resources. This is because the budget may not allow all "net-beneficial" programs to be provided. Where this is the case, the health benefit is only maximized if programs are implemented in order of increasing cost-utility ratio, emphasizing the importance of the continued presentation of the cost-utility ratio.

Andrew H. Briggs

See also Cost-Benefit Analysis; Cost-Effectiveness Analysis; Disability-Adjusted Life Years (DALYs); EuroQol (EQ-5D); Health Utilities Index Mark 2 and 3 (HUI2, HUI3); Healthy Years Equivalents; Net Monetary Benefit; Quality-Adjusted Life Years (QALYs); Utility Assessment Techniques

Further Readings

Drummond, M. F., Sculpher, M. J., Torrance, G. W., O'Brien, B., & Stoddart, G. L. (2005). *Methods for the economic evaluation of health care programmes* (3rd ed.). Oxford, UK: Oxford University Press.

Gold, M. R., Siegel, J. E., Russell, L. B., & Weinstein, M. C. (Eds.). (1996). *Cost-effectiveness in health and medicine*. New York: Oxford University Press.

Counterfactual Thinking

Counterfactual thinking in judgment and decision making occurs when the decision maker considers or imagines outcomes of a decision that could have occurred but did not. For example, a patient who experiences a surgical complication that results in disability might easily imagine counterfactual worlds in which his outcome was different. A great deal of theoretical and empirical work on counterfactual thinking in decision making has its genesis in the seminal work of Kahneman and Miller on norm theory.

Types

In most decisions, there are many counterfactual outcomes and several different ways that counterfactual outcomes can be imagined to occur. First, in decisions under uncertainty, chance factors (the "state of the world") could be imagined to have been different. For example, the surgical patient might imagine that his surgery had proceeded without the complication. Second, decisions taken by others could have been different. For example, the surgical patient might imagine that his surgeon had chosen a different procedure that could not lead to the complication. Third, the decision of the decision maker could have been different. For example, the surgical patient might imagine that he had chosen a medical treatment (with a successful outcome)

instead. Fourth, the decision maker could imagine himself or herself to be a different person, a so-called social counterfactual. For example, the surgical patient might imagine other people he knows with different health problems.

Each counterfactual can potentially result in a comparison between the actual outcome and the counterfactual outcome. The surgical patient might compare his new life with disability to (a) how he imagines his life might have been if the surgery had been uncomplicated, (b) how he imagines his life might have been if the surgeon had chosen a different surgery, (c) how he imagines his life might have been if he had chosen a medical treatment, or (d) how he imagines the lives of his peers (with different health problems) might compare with his new life.

Ease of Imagining Counterfactuals

Although multiple counterfactuals are nearly always available, the ease with which a particular counterfactual outcome is generated or used in comparisons varies. The psychological literature uses the term *mutability* to refer to the aspects of reality that are most amenable to yielding counterfactuals. For example, exceptional events are more mutable than normal events (so people are more likely to imagine what would have happened if an exception had not occurred than to imagine what would have happened if an exception had occurred). Events under the decision maker's control are typically more mutable than uncontrollable events, actions are more mutable than inactions or omissions, repeatable events are more mutable than one-time events, and effects are more mutable than causes.

Direction and Impact on Postdecision Emotion

Counterfactuals are also referred to by their direction or valence. Upward counterfactuals are alternative outcomes that the decision maker considers superior to the actual outcome. Downward counterfactuals are alternative outcomes that the decision maker considers inferior to the actual outcome. Counterfactual comparisons reliably change the way decision makers feel about their actual decision outcomes (their postdecision affect). Research on counterfactual comparisons has demonstrated that upward counterfactual comparisons, which typically result in lower postdecision satisfaction and more negative postdecision affect, are more common and carry more weight than downward comparisons, which typically result in greater postdecision satisfaction and more positive postdecision affect. In addition, surprising outcomes, which more easily evoke counterfactual alternatives, typically result in more extreme postdecision affect. For example, a rare and surprising recovery is experienced with greater elation than a common and expected return to health.

Functional Impacts

Counterfactual thinking may serve functional purposes. Upward counterfactuals may direct the decision maker to reflect on aspects of the decision process that may have led to a poor outcome and could have been undertaken differently. This reflection may result in an improved decision process if the decision maker is again faced with the same or a similar decision. Downward counterfactuals may reduce postdecision regret by providing the decision maker with a comparison in which the decision outcome can be cast as superior to the counterfactuals.

To the degree to which decision makers actively seek to consider potential alternative outcomes prospectively, they may also anticipate the counterfactual comparisons that are likely to co-occur with particular outcomes. Such anticipated counterfactuals may form the basis for decision-making strategies that seek to, for example, minimize expected regret.

Alan Schwartz

See also Emotion and Choice; Regret

Further Readings

Kahneman, D., & Miller, D. T. (1986). Norm theory: Comparing reality to its alternatives. *Psychological Review, 93,* 136–153.

Mandel, D. R., Hilton, D. J., & Catellani, P. (2005). *The psychology of counterfactual thinking*. New York: Routledge.

Roese, N. J., & Olson, J. M. (Eds.). (1995). *What might have been: The social psychology of counterfactual thinking*. Mahwah, NJ: Lawrence Erlbaum.

Cox Proportional Hazards Regression

In the analysis of survival data, researchers want to ascertain characteristics of the patient that influence patient survival time. The relationship between a single response variable (survival time) and covariates (patient/disease characteristics) is often inferred through the use of a regression model. Typical regression models, such as linear or logistic regression, do not work when the response variable is survival time, since the time to death may not be recorded for all patients at the time of analysis. If a patient is still alive at the time of analysis or has been lost to follow-up, the patient survival time is said to have been right censored (or simply censored) at the time of the last observed follow-up. If the patient has been lost to follow-up, an important assumption in many survival analytic methods is that the reason a patient is lost to follow-up is unrelated to the risk of death.

In survival analysis, when the survival time, T, is possibly right censored, the Cox proportional hazards model is the predominant regression model. The proportional hazards model is written as

$$h(t|X) = h_0(t)\ \exp[b^{\mathrm{T}}X], \tag{1}$$

where $h(t|X)$ is the hazard function conditional on a set of patient-specific covariates, which is denoted by the vector X, and b represents the vector of regression coefficients that determines the relationship between the covariates and the risk of death. Covariates in the Cox model are handled using standard regression techniques. Thus, categorical factors may be entered into the model using dummy variables, and interactions may be introduced through the multiplication of two covariates. However, due to complications that stem from censored observations, additional methodology, based on what is called the partial likelihood, is needed for estimation of the regression coefficients b.

The conditional hazard $h(t|\mathrm{X})$ provides the patient-specific risk of death over time. The proportional hazards specification, Equation 1, divides the conditional hazard into two components, a baseline hazard function $h_0(t)$ independent of the patient characteristic vector and the patient relative risk function, $\exp[b^{\mathrm{T}}X]$, independent of time; the relationship between the two components is multiplicative. The baseline hazard function is left unspecified but governs how the patient-specific hazard varies over time. Heuristically, the hazard function is proportional to the probability of death by time t, given the patient has not died prior to time t. For any two patients with characteristics X_1 and X_2, the ratio of their conditional hazards,

$$h(t|X_1)/h(t|X_2) = \exp[b^{\mathrm{T}}(X_1 - X_2)],$$

is independent of time. The term *proportional hazards* refers to the fact that the two conditional hazards are proportional to each other, with the proportionality constant equal to $\exp[b^{\mathrm{T}}(X_1 - X_2)]$.

The widespread popularity of the proportional hazards methodology stems from the interpretation of the regression parameter, b, as a relative risk parameter constant with respect to time, the accuracy of the estimate of the relative risk parameter in the presence of censored data, the development of inferential procedures that are easy to implement with available software, and the efficiency of the regression parameters for a wide range of baseline hazard functions.

An alternative specification of the proportional hazards regression model is through the patient specific (conditional) survival function,

$$S(t|X) = S_0(t)^{\exp[bX]}, \tag{2}$$

where the term $S(t|X)$ represents the probability that a patient with characteristics denoted by the covariate vector X survives beyond time t. This specification of the proportional hazards model enables the regression model to be used for prediction. For example, using Equation 2, the analyst can predict the patient-specific probability of survival beyond 5 years or the median survival time for a given set of patient characteristics. Thus, the proportional hazards model enables a refinement of the Kaplan-Meier estimate of a survival probability by providing an estimate for the probability of survival beyond t years for a patient with characteristics represented by the covariate vector X.

In addition to ascertaining the risk profile of a patient, the proportional hazards model is used to adjust for patient risk in testing the equality of the survival distributions between exposure and treatment groups. This application, often termed the

analysis of covariance, has historically been used in the analysis of observational studies. For this application, the proportional hazards model may be written as

$$h(t|Z, X) = \exp[aZ + b^{\mathrm{T}}X],$$

where Z represents the treatment group classification, X represents the vector of potential confounding factors, and the parameter of interest, a, represents the treatment effect on survival time. The analysis of covariance in the setting of survival analysis would test whether $a = 0$, that is, whether there is a treatment effect, after adjusting for potential confounding factors.

An interesting generalization of the proportional hazards model is the incorporation of time-dependent covariates,

$$h(t|X) = \exp[b^{\mathrm{T}}X(t)].$$

Under this generalization, the proportional hazards specification no longer holds, as the relative risk,

$$h(t|X_1(t))/h(t|X_2(t)) = \exp[b^{\mathrm{T}}(X_1(t) - X_2(t))],$$

now changes over time. As a result, the model is often referred to as the time-dependent Cox model rather than the proportional hazard model. The time-dependent Cox model is useful when disease-related patient characteristics change over the course of follow-up. For example, a prostate cancer patient who experiences a prostate-specific antigen (PSA) relapse after receiving a course of therapy is at greater risk of death after relapse than before it. The time-dependent covariate Cox model enables the analyst to recalibrate the risk of death at the point of time during follow-up that the patient experiences the PSA relapse.

Although the proportional hazards model is robust, its application is not universal. The proportional hazards model is termed a semiparametric model because the baseline hazard function, $h_0(t)$, and the baseline survival function, $S_0(t)$, are not specified for the purpose of estimating the relative risk coefficient, b. This provides a robustness quality to this regression model, enabling the proportional hazards regression model to be applied to a wide array of survival data. There are characteristics of the data, however, which need to be compatible with the assumptions implicit in the proportional hazards model, in order for the results of the analysis to be meaningful. For example, in a simplified version of the proportional hazards model with a single binary treatment covariate, the Cox model implies that the survival probability for patients on one treatment dominates the survival probability for the cohort of patients on the other treatment over the entire patient follow-up. If, however, the survival curves cross over time, the proportional hazards assumption does not hold, and the proportional hazards model is not appropriate for data summarization. The validity of the proportional hazards specification is more difficult to diagnose if there are many (possibly continuous) covariates under consideration.

In general, if the proportional hazards assumption is incorrect, application of this model is likely to lead to incorrect conclusions regarding the relationship between the covariates and survival time. In this circumstance, it would behoove the data analyst to consider alternative regression models for survival data. The most common alternative to the proportional hazards model is the accelerated failure time model

$$\log t_i = b^{\mathrm{T}}X_i + e_i,$$

where the e_i represent stochastic errors generated independently from a common but unknown distribution, the vector X denotes the patient-specific covariates, and b is the regression coefficient vector.

An additional assumption in the proportional hazards model is that the relative risk is monotonically increasing or decreasing in the covariates. If some of the important covariates in a particular data set are continuous, such as age or white blood cell count, it is important to assess whether this specification is correct. For example, it is plausible that a patient with either a low or a high white blood count (WBC) is at greater risk of death than a patient with a WBC in the normal range, and thus Equation 1 is inappropriate. An approach to generalizing Equation 1 is based on nonparametric estimation methods, such as spline or kernel estimation, which provide a more flexible approach to specifying the relative risk function.

Finally, like uncensored regression models, individual observations may either provide a poor fit for the model or have undue influence of the

estimated regression coefficients. In classical statistical terms, these would be defined as data points with large residuals or high leverage. These data values should be monitored and either downweighted or removed during the course of the data analysis.

Glenn Heller

See also Analysis of Covariance (ANCOVA); Hazard Ratio; Logistic Regression; Nomograms; Prediction Rules and Modeling; Survival Analysis

Further Readings

Collett, D. (1994). *Modelling survival data in medical research*. London: Chapman & Hall.

Cox, D. R. (1972). Regression models and life-tables (with discussion). *Journal of the Royal Statistical Society, Series B, 34,* 187–220.

Fisher, L. D., & Lin, D. Y. (1999). Time dependent covariates in the Cox Proportional Hazards Regression Model. *Annual Review of Public Health, 20,* 145–157.

Harrell, F. E. (2001). *Regression modeling strategies: With applications to linear models, logistic regression, and survival analysis*. New York: Springer-Verlag.

Hosmer, D. W., & Lemeshow, S. (1999). *Applied survival analysis: Regression modeling of time to event data*. New York: Wiley.

Cues

The term *cue* in decision making is a broad one denoting every piece of information outside the decision maker that may help in a decision or judgment under uncertainty. Other personal information such as goals or preferences also influence decisions, but these pieces of information are not called cues. Many ways of integrating cue information exist, called decision rules. Their accuracies crucially depend on the structure of the decision environment, and therefore, statistical models of the decision domain are necessary to derive prescriptions of a good decision strategy.

Cue Values

Cues are variables that can be used to judge, infer, or predict the value of an unknown criterion variable of interest. In a specific decision situation, a cue may take on a certain cue value that is indicative of the value of the to-be-inferred criterion. In medicine, for example, symptoms and laboratory results are cues that are used to infer the underlying disease. Likewise, medical parameters such as blood pressure, smoking habits, and symptom severity may serve as cues to predict the survival time of a patient. Hence, the term *cue* is neutral as to whether it is a cause or an effect of the variable which is inferred. Even a merely statistical relation between cue and criterion (without causation) can render the cue useful for inferences. The inference or prediction can be a classification (categorical variable, e.g., disease), a continuous judgment of a quantity (e.g., expected survival time), or a comparative judgment concerning several options (e.g., which treatment will be most successful?). To be useful for inferences, cues must have a high predictive power or correlation with the criterion variable, called the *ecological cue validity*.

Cue Validity

Like the criterion, cues can be continuous variables (e.g., blood pressure) or categorical variables (e.g., symptoms). Depending on the nature of the cues and criterion, different measures of cue validity may be useful. If cue and criterion are continuous variables, Pearson correlations or partial correlations (if a whole set of correlated cues is used for prediction) measure the predictive power. Likewise, (point-)biserial correlations or different contingency coefficients can be used to express the degree of the statistical relationship between the cues and criterion if one or both variables are categorical or binary. In pairwise comparisons (e.g., "Who of two patients has better survival chances when treated first in the emergency room?") with binary cues (e.g., Symptom X present vs. absent), the validity is often defined as the conditional probability of deciding correctly, given that the cue discriminates between the options. A cue discriminates if it takes on different values for the compared objects. Hence, besides validity, the discrimination rate of a cue is another important aspect of its usefulness for decisions because a cue is only helpful if the values differ between options.

In principle, in a set of statistically related variables, any of these variables can serve as cues for

predicting one of the other variables. However, a high cue validity in one inference direction does not imply high validity in the other direction. For instance, there may be a high conditional probability of a symptom given a disease (e.g., fever given pneumonia), whereas the reverse is not necessarily true if the symptom is not specific for the disease. Hence, for using cues in a systematic fashion, their relation to the criterion must be known. If only one valid cue is available for a decision, matters are quite easy since the best bet is to go with the cue. Typically, however, multiple (and potentially contradicting) cues have to be integrated into one judgment or decision that requires conflict resolution and information integration via decision rules. The success of a decision rule depends on its fit to the statistical structure of the environment. For example, if the available cues are highly correlated, it may be worthwhile and time-saving to consider only a small subset of cues because the other information is redundant.

Models of the Environment

The psychologist Egon Brunswik introduced the idea of the lens model, which is an attempt to model the environment, the decision process, and their mutual fit simultaneously. The cues are the "lens" through which the distal criterion variable can only indirectly be perceived. Further developments of the lens model in social judgment theory use a multiple linear regression to predict the criterion on the basis of the cues. This regression informs the investigator how predictable the criterion is given the cues and provides beta weights that measure the contribution of each single cue to a weighted linear prediction of the criterion, hence its ecological validity. On the other side of the lens, one can perform a regression of actual judgments on the cue values as predictors. This can be seen as a model of the decision maker (called policy capturing), and the regression weights measure the influence of the cues on judgments, or cue use. Both regressions can be compared to see if the judgmental cue weighting matches the optimal weighting, that is, if use coefficients match the ecological validities. The use of linear regressions has dominated research for decades, but the idea of analyzing the environmental structure and its match with psychological processes can be applied more generally, to include nonlinear cue-criterion relationships (e.g., U-shaped or exponential) or nonlinear cue combinations. Optimal decision algorithms can also be identified using machine learning approaches or Bayesian networks, which need a large amount of training in huge databases.

However, the "optimal" rules often need extensive computation to combine cues in sophisticated ways, for example, in Bayesian networks or a weighted additive integration. In many instances, the accuracy of such complex rules can be approximated by simpler algorithms or so-called heuristics. For instance, extensive simulations have shown that linear models often have a *flat maximum*, which means that nonoptimal weighting of cues does not hurt the predictive accuracy very much as long as the direction of the cue-criterion correlation is correctly specified. This is especially the case in environments with many cues that do not differ too extremely regarding their validities. In environments with few available cues of very different predictive power, simple noncompensatory rules such as truncated decision trees or lexicographic rules can approximate the performance of optimal models. In a noncompensatory rule, a bad (or good) value on one cue cannot be compensated for by other cues. For example, if a disease has an obligatory symptom, the missing of this symptom rules out the disease regardless of other symptoms that may be present and fit the diagnosis of the disease. A lexicographic choice rule, for example, would look up the options' values on the most valid cue and ignore other cue information unless the best cue does not discriminate. In this case, the second best cue is searched and so on. The rule is also "noncompensatory" because a choice determined by a better cue cannot be revised by less valid cues. For this simplified rule to work well, one needs an accurate knowledge of the validity hierarchy of cues, that is, which cue is best, second best, and so on.

Models of the Decision Maker

In a research tradition called multiple cue probability learning, psychologists have investigated people's ability to abstract information about cue-criterion correlations from feedback and to use them for prediction. Typically, learning from feedback is not overly successful unless there are only very few cues with simple linear relationships to

the criterion. If the situation gets more complex, cognitive feedback or causal models help. Cognitive feedback not only provides outcome feedback after a choice but also gives further information about the direction and the amount of the deviation from the correct judgment or even points to explicit cue-criterion relationships. However, people are generally not successful if cue-criterion relations are nonlinear or cue interactions occur. The judgments can often nevertheless be described by a weighted linear model. In experimental situations with novel tasks and explicit cues, participants sometimes use simplifying noncompensatory strategies, especially under time pressure or when cue acquisition is costly.

On the other hand, experts who have had extensive training and feedback often show remarkable decision accuracy in their domain. For example, weather forecasters are very well-calibrated in predicting the probability of precipitation. Also, pathologists may be very accurate in judging tissue samples as malign or benign although they cannot verbalize how they do it. It is obvious that these experts use effective cues, but neither all the cues used nor the decision rule are accessible to verbalization. In this case, the researcher's challenge is to identify the cues and strategies these experts use. It must be acknowledged, however, that judgments of experienced experts based on multiple explicit verbal cues (e.g., clinical judgments based on personality profiles or symptom patterns) are often outperformed by relatively simple statistical models of the environment.

Arndt Bröder

See also Cognitive Psychology and Processes; Decision Rules; Decision Tree: Introduction; Information Integration Theory; Lens Model; Ordinary Least Squares Regression; Social Judgment Theory

Further Readings

Brehmer, B. (1994). The psychology of linear judgement models. *Acta Psychologica, 87*(2/3), 137–154.

Dawes, R. M. (1979). The robust beauty of improper linear models in decision making. *American Psychologist, 34*(7), 571–582.

Gigerenzer, G., Todd, P. M., & ABC Research Group. (1999). *Simple heuristics that make us smart*. New York: Oxford University Press.

Hammond, K. R., & Stewart, T. R. (Eds.). (2001). *The essential Brunswik*. New York: Oxford University Press.

Martignon, L., & Hoffrage, U. (2002). Fast, frugal, and fit: Simple heuristics for paired comparison. *Theory and Decision, 52,* 29–71.

Cultural Issues

Culture encompasses the acquired knowledge, beliefs, values, and behavior patterns shared by the members of a particular group of people. Common elements of cultures include language, diet, dress, and religion, among others. In the past, the typical person in any given culture had little if any contact with individuals from other cultures. But marked shifts in economic, social, and political arrangements, including unprecedented worldwide immigration flows, have ended such isolation. Thus, people are constantly interacting with others who embrace customs markedly different from their own; they might even live right next door to them. This creates the real possibility that a person faced with a significant medical issue will be dealing with a healthcare provider from another culture. Such cross-cultural encounters pose challenges to how and how well the required healthcare decisions are made. This entry describes and analyzes some of the most important of those challenges. It also outlines approaches to meeting them.

The Patient–Provider Relationship

The first key challenges bear on the personal relationship between the patient and the provider in a cross-cultural interaction. Specifically, they concern confidence, comfort, and trust.

Confidence

Healthcare providers can serve several distinct decision-making roles vis-à-vis their clients. First, they can be *agents*, making decisions on the patient's behalf, as when the patient says (explicitly or merely implicitly), "I realize that the decision is mine legally, but would you please decide for me? After all, you're the expert, and besides, I'm just too upset by this horrible news to make the decision

myself." Second, they can be *co-deciders*, in the spirit of the shared decision-making paradigm. That is, the provider and patient work toward mutual agreement about, say, a workable hypertension management regimen for the patient to follow. Third, providers can be *consultants*, such that the patient reserves the right to decide personally but seeks the provider's opinion as input to the decision process, for instance, in the form of a prostate cancer prognosis or a recommendation for radiation therapy versus radical prostatectomy. Finally, providers can be (and invariably are) *decision managers*, deliberately or inadvertently exerting influence over how the patient chooses by, say, providing literature that favors radiation rather than surgical treatment for the patient's condition. Whether and how the patient allows the provider to assume these roles depends directly on the client's confidence in the provider's competence or expertise. And that confidence can easily be affected by cultural differences.

Impressions of expertise generally, and of decision-making expertise in particular, rest on several considerations. One consideration is *acclamation*, consensus among people one already respects. Such consensus in turn depends on factors such as the person's visibility and regard by one's peers. In the health arena, credentials and accomplishments in contemporary science-based medicine undoubtedly carry great weight even with people who normally have little to do with modern societies. Yet, all else being the same, a provider from a culture different from the patient's own almost necessarily is less well known in that patient's social circles than a provider who shares the patient's own culture and therefore suffers a perceived competence liability. Another consideration is *style of speaking*. People recognized as experts tend to speak with precision and confidence. A provider from a culture different from the patient's is unlikely to be fluent in the patient's native language and thus is incapable of exhibiting the linguistic trappings of expertise. Yet another consideration is *factual knowledge*. People expect true experts to be able to recite extensive facts about the domain in question. And they certainly expect experts to know virtually all the facts they know themselves, and more. As discussed below, cultures often differ in terms of disease prevalence rates as well as common treatments, including folk remedies. A provider from a different culture might well be ignorant of these facts that the patient knows personally, thereby suffering damage to his or her credibility in the patient's eyes.

Comfort

Extensive research (e.g., on the "mere exposure effect") has shown that familiarity generally does not breed contempt but instead fosters at least mild liking. Thus, when a healthcare provider shares cultural customs with a patient, such as language, memories of the same kinds of schools, and similar tastes in entertainment, the patient feels at ease. This comfort can be highly beneficial for managing the stress that naturally accompanies health crises. It also undoubtedly contributes to the success of clinics that cater to immigrant and expatriate communities and which emphasize elements of the pertinent cultures, such as their languages and religious sensitivities. All the common provider decision-making roles—agent, co-decider, consultant, and decision manager—can be enacted more smoothly.

The flip side of the coin is where the greatest challenges lie. When there are significant differences in the cultures of a patient and a provider, those differences often constitute barriers that must be overcome. Studies on decision "bolstering" have shown that, when a person chooses X over Y, in that person's estimation, the appeal of X increases and that of Y diminishes. That person wonders, "How could I have ever even considered Y?" It only stands to reason that, all else being equivalent, people who like and choose Y will be seen as having tastes that are not merely different but in some sense inferior. Thus, in a broader context, it should not be surprising if, when unchecked, other cultures' "choices" are initially regarded somewhat negatively; they are not our own. Their music sounds like "noise," their food tastes too bland or too spicy, and some of their worldviews seem unreasonable. Unaddressed, the resulting discomfort can stand between the patient and provider, even when, ostensibly, the cultural differences in question have nothing to do with medicine.

Trust

Cultural variations are often accompanied by economic and political rivalries. Consider, for

instance, the Fleming and Walloon cultures in Belgium or the Malay and Chinese cultures in Malaysia. The resulting antagonisms can fan distrust in all manner of cross-cultural encounters, including medical ones. Initially, at least, patients therefore might easily say things such as the following to themselves concerning a provider from a rival culture: "They usually don't like or respect us, so I wonder if she's that way, too. Will she really give me her best efforts, just like she would one of her own?" Unless and until such fears are dispelled, patients are hesitant about having providers make medical decisions on their behalf.

To address threats to effective cross-cultural patient-provider relationships, particularly concerning comfort and trust, those who seek to promote cultural competency in healthcare offer several recommendations for providers:

- Undertake exercises intended to uncover one's personal feelings about various cultural variations. As suggested previously, these feelings are virtually guaranteed to exist.
- Cultivate respect for customs different from one's own. At minimum, be nonjudgmental about them.
- Avoid stereotyping. A provider should anticipate and prepare for customs that are especially common in a given patient's culture. At the same time, the provider should not lose sight of the extensive individual differences that invariably exist among the members of that culture. Proceeding as if every member of the group is the same invites anger and resentment.

And when the concern is nurturing patient confidence in a provider's competence, there is no substitute for establishing and publicizing a solid track record among patients in a given cultural community. That includes learning more about that culture, especially facts pertaining to medical conditions that are particularly problematic in that community and health practices that are distinctive for the people concerned.

Participation

In many societies, particularly in North America and Western Europe, decisions about medical questions are personal and private affairs, solely between the patient and the physician. Expectations and reality can be markedly different in other societies, particularly ones where collectivism rather than individualism holds sway, for instance, in much of Asia, Africa, and Latin America. By definition, the term *collectivism* refers to an outlook that emphasizes the interdependence of people and the importance of collectives to which they belong. In contrast, *individualism* highlights independence and the relative significance of people's personal interests.

In collectivistic cultures, participation in the medical decision process tends to be broader than in individualistic cultures, with the family often assuming especially prominent roles. For instance, for a long time in Japan, the norm has been not the patient autonomy that is ascendant in the United States but, instead, reliance on the beneficence of the patient's family and physician. Thus, in this alternative arrangement, a physician could disclose to the family that a patient has cancer and the family might choose to withhold that diagnosis from the patient. Prominent roles for families in medical decision making in collectivistic societies are consistent with the high degree of interdependence characteristic of those societies. In those contexts, the reality is that events involving any one member of a family (e.g., a new, high-paying job or a serious illness) often have a much greater impact on the other members than would be the case in an individualistic society where independence is prized. So when one family member becomes sick, prescribed changes in diet for managing chronic conditions such as diabetes must take into account the impacts for numerous individuals besides the person who is ill.

Beyond the family, the traditions in some cultures reserve decision-making roles for others, too. Most notably in some Asian, African, and Native American cultures, these might be people with religious responsibilities, including ones some would call shamans. Other participants might be nonreligious traditional or alternative healers. Patients from cultures that maintain roles for these additional parties sometimes attempt to follow the guidance of these authorities as well as the instructions of their doctors practicing contemporary scientific medicine. Since patients might be reluctant to volunteer such information, it is wise for providers to inquire sensitively about the possibility. When other parties are involved, the provider

must, in effect, negotiate a hybrid treatment plan. To do otherwise runs the risk of treatment incompatibilities, perhaps tragic ones.

Language

The complications of patients and providers speaking, reading, and writing different formal languages are apparent. For instance, the provider's misunderstanding of a patient's symptom descriptions could lead to a misdiagnosis and then an ineffective, even harmful, treatment choice. This implicates the need for the services of skilled professional interpreters. (Reliance on bilingual family members and friends is often discouraged because they tend to censor remarks on both sides of a conversation.) But even when a patient and a provider in a cross-cultural encounter use a common formal language, there remain significant risks. Because their life experiences might be so different, so too might be the assumptions they make in a given exchange. That is why it is often advised that clinicians depend especially heavily on open-ended questions and requests (e.g., "Would you please tell me what *you* think led to this?"). Communications can also be compromised by cultural differences in traditions of directness, politeness, and deference to authority or status. Thus, "Yes" in response to a request might not mean "Yes" literally, expressing an intention to comply with that request. The speaker's true meaning might have to be inferred from other aspects of the situation, including nonverbal signals such as facial expressions. Becoming skilled in nonverbal communication is therefore essential although difficult to achieve, especially since the same signal (e.g., a smile) can sometimes carry opposite meanings in different cultures.

Decision Problem Deliberation

The specific decision problems that patients and providers must confront concern acute and chronic conditions as well as health maintenance measures. Effective decision making requires that deciders successfully address certain recurring issues. These include anticipating or recognizing problems that demand decisions (e.g., slowly developing cancers), judging the chances of pertinent events (e.g., that a treatment would work), determining values (e.g., the patient's true feelings about possible outcomes and side effects), and creating or identifying viable options (e.g., crafting effective, doable treatment plans). Specific considerations are likely to affect precisely how these issues are resolved in cross-cultural encounters.

Correlated Health Facts

An especially significant reality of cultural variations is that some of them are correlated with important health facts. Incidence rates are one instance. For genetic reasons, certain diseases are more common in some cultural groups than in others. Sickle cell disease, with relatively high incidence rates among African Americans, provides a ready illustration. The disease has sometimes been misdiagnosed in African American patients because it simply never occurred to their physicians as a possible explanation for their signs and symptoms. That has happened because those doctors had little experience with non-Caucasian patients or, perhaps in a spirit of fair-mindedness, they simply assumed that, physiologically, "people are people." Other incidence rate differences are tied more directly to true cultural variations in behavior. Such is the case for cultural differences in obesity and hypertension traced to customary diets heavy in fats or salt. More generally, if a provider is ignorant of incidence rates that are distinctive for a cultural group, this foreshadows diagnostic errors as well as cases of *blindsiding*. These are cases such as those in which a serious illness is inadvertently allowed to progress to an untreatable state because its actual high-probability presence was never even imagined.

Differential efficacy rates are another health fact sometimes correlated with culture. Treatments that are effective for some cultural groups are less useful for others. For instance, studies have found especially high rates of adverse reactions among East Asian patients for certain antihypertension drugs. The implications of culture-specific efficacy rates for the wisdom of treatment choices are readily apparent. Clearly, when a clinician is called on to serve patients from an unfamiliar culture, a first order of business must be to actively seek out known health fact correlations involving that group.

Explanation and Belief

Cultures sometimes differ in how people explain what they observe. Related to this, cultures can also

differ in how people arrive at what they believe or expect to be true. These differences can have implications for the treatment options that occur to patients and providers in cross-cultural encounters as well as for their expectations about the effectiveness of those treatments, if they were to be chosen. For example, in numerous African and Asian cultures, people sometimes believe that a person's illness is punishment for transgressions that the person has committed. Or they suspect that the sick person is the victim of malevolent enemies with spiritual powers that enable them to cast hexes on whomever they wish. These explanations differ sharply from the accounts for sickness that underlie current scientific medicine. And they rationalize radically different treatment alternatives. Whereas scientific explanations point toward treatments that address factors such as pathogens, spiritual or moral explanations implicate actions such as restitution or prayer. A provider who ignores (or worse, belittles) a patient's beliefs in alternative explanations for illness is unlikely to succeed in achieving the patient's cooperation in implementing a purely science-based treatment plan.

In order for a clinician to persuade a patient to undergo a particular treatment, regardless of its character, the clinician must somehow get the patient to conclude that the chances of good outcomes are high. This challenge is conditioned by culture, too. Studies have documented reliable cultural variations in people's probability judgments. Surprisingly for many people, one of the most consistent differences is that the overconfidence implicit in such judgments is stronger among Chinese than among Americans. Further studies have demonstrated that these differences are not a reflection of "ego." Instead, they seem to result from culturally distinct customs for reasoning while arriving at one's conclusions.

Treatment Options and Expectations

If a treatment option goes unrecognized, then neither the patient nor the provider can choose it. On the other hand, they could not disagree and argue about it either. These truisms highlight the importance of cultural variations in the options that come to mind when a medical issue arises. There is good evidence that cultural differences often broaden the pool of alternatives that surface, even beyond the kinds of scientific and spiritually inspired options suggested previously. For some rural Mexican patients, medical conditions are classified as either "hot" or "cold." Furthermore, to reestablish balance, "hot" conditions are thought to require "cold" treatments, and vice versa. Pregnancy is regarded as a "hot" condition, and thus "hot" treatments would not be on the list of options considered legitimate for a pregnant woman. But vitamins are a hot treatment. And therein lies a conflict that must be worked through since a practitioner of contemporary scientific medicine almost certainly would recommend the regular intake of vitamins to assure the health of the child and the mother. Or consider Japanese family medicine customs. Over time, Japanese patients have developed an expectation that an end result of virtually every visit to the doctor should be a prescription for medicine. There is also an expectation that the patient will be seen again soon, say, in a month. Naturally, then, the treatment plans that experienced, wise, and highly rated physicians craft for patients' consideration conform to these expectations.

The advice implicit in these kinds of scenarios is similar to that articulated earlier with respect to culture-correlated health facts. A provider who anticipates working with patients from a new, unfamiliar culture can at least prepare for the challenges of broader collections of patient-preferred treatment options by studying what the common expectations are within that culture.

Value

The final class of cultural variations that have special significance in deliberations bear on what is perhaps the most fundamental defining characteristic of decision making generally—value. Decision problems are special largely because their solutions are not unique. Since people's ways of valuing things tend to differ, outcomes that are highly pleasing for one patient (e.g., the ebullient personality of the physician's assistant assigned to the patient) can easily be unbearably annoying for another. Cultural variations concerning value are particularly important in the provider's roles as an agent, making decisions on the patient's behalf, and as a co-decider, seeking to reach agreement with the patient about how to proceed in dealing with a medical situation.

There is evidence that physicians can err substantially in their expectations about the values their patients attach to various aspects of medical situations, such as cancer treatment side effects. This should mean that, if a physician is making medical choices for a patient, at least sometimes these value judgment errors should result in the patient being stuck with alternatives that are worse than they could be, in terms of what the patient truly desires. Importantly, these errors are not haphazard; there is evidence for *false consensus*. That is, people tend to believe that others' values are closer to their own than they really are. Suppose that a provider attempts to answer the question, "How would my patient feel about the limited degree of mobility likely to result from this treatment?" The conclusion is essentially, "Pretty much the way I would," and more so than would actually be the case.

There is reason to expect these value assessment errors to be especially large in cross-cultural patient-provider encounters. That is because the values of the patient and provider should differ more than in instances where the parties share a common culture. It is easy to appreciate sizable cross-cultural value differences when the focus is something like tastes in food or music. But the same underlying principles, such as historical isolation of groups of people from one another, should yield similar strong value differences for aspects of health situations. British versus American differences in the appeal of aggressive as opposed to conservative treatments for cancer provide a good example. Relative to Americans, Britons have been more likely to regard the side effects of aggressive treatments to be unacceptably harsh. Heroic and invasive efforts to extend for a few hours the lives of terminally ill patients in Japan constitute another illustration. The justification for some such actions is to provide time for all close relatives to be at their loved one's side at the moment of death, which is extremely important in Japanese society, much more so than elsewhere. Thus, as co-deciders, there would be little disagreement between a Japanese family and their Japanese doctor about undertaking the requested life-extending measures. That might not be so if the physician were non-Japanese or at least ignorant of Japanese traditions and values.

J. Frank Yates and Laith Alattar

See also Informed Consent; International Differences in Healthcare Systems; Models of Physician-Patient Relationship; Moral Choice and Public Policy; Religious Factors; Worldviews

Further Readings

Fetters, M. D. (1998). The family in medical decision making: Japanese perspectives. *Journal of Clinical Ethics, 9*(2), 132–146.

Hammoud, M. M., White, C. B., & Fetters, M. D. (2005). Opening cultural doors: Providing culturally sensitive healthcare to Arab American and American Muslim patients. *Obstetrics & Gynecology, 193,* 1307–1311.

Juckett, G. (2005). Cross-cultural medicine. *American Family Physician, 72*(11), 2267–2274.

Management Sciences for Health. (n.d.). *The provider's guide to quality and culture.* Retrieved June 7, 2008, from http://erc.msh.org/mainpage.cfm?file=1.0.htm&module=provider&language=English

McDowell, S. E., Coleman, J. J., & Ferner, R. E. (2006). Systematic review and meta-analysis of ethnic differences in risks of adverse reactions to drugs used in cardiovascular medicine. *British Medical Journal, 332,* 1177–1181.

Sequist, T. D., Fitzmaurice, G. M., Marshall, R., Shaykevich, S., Safran, D. G., & Ayanian, J. Z. (2008). Physician performance and racial disparities in diabetes mellitus care. *Archives of Internal Medicine, 168*(11), 1145–1151.

Triandis, H. C., Bontempo, R., Villareal, M. J., Asai, M., & Lucca, N. (1988). Individualism and collectivism: Cross-cultural perspectives on self-ingroup relationships. *Journal of Personality and Social Psychology, 54*(2), 323–338.

Yates, J. F., Lee, J.-W., Sieck, W. R., Choi, I., & Price, P. C. (2002). Probability judgment across cultures. In T. Gilovich, D. Griffin, & D. Kahneman (Eds.), *Heuristics and biases: The psychology of intuitive judgment* (pp. 271–291). New York: Cambridge University Press.

D

Data Quality

Quality data are at the heart of quality healthcare. It is well known that poor data can lead to incorrect diagnoses, prescription errors, or surgical errors with tragic consequences. Similarly, the day-in, day-out consequences of poor data are enormous as well, leading to added time and expense throughout the system. In short, improving data quality is essential.

There are many approaches to defining data. The one that is often used for data quality recognizes that data consist of two interrelated components: data models and data values. *Data models* define *entities*, which are real-world objects or concepts, *attributes*, which are characteristics associated with entities, and *relationships* among them. As an example, each reader is an entity, and his or her employer is interested in attributes such as name, date of birth, and specialty. Relationships may include report manager and subordinates. A *data value* is the specific realization of an attribute/relationship for a specified entity. For example, a member of a medical research team may be assigned the specialty "statistician." Clearly, data, per se, are abstract. *Data records* are the physical manifestations of data in paper files, forms, spreadsheets, databases, and so forth.

Physicians are uniquely positioned to initiate data quality efforts and have much to gain by doing so. However, most are unfamiliar with the thinking that underlies data quality management: As in healthcare, the steps one takes to improve data quality are rooted in the scientific method. Thus, this entry focuses on physicians. The first part summarizes three key principles of data quality management and the second part offers eight simple prescriptions that physicians can follow to make immediate improvements. These will not, of course, address all the data quality issues that currently afflict healthcare. But they form a solid beginning: the data quality equivalents of the age-old dictum, "First, do no harm."

Principles of Data Quality Management

The "muscle and bone" of data quality are measurement and control. One simply must have the facts and work through the laborious process of formulating and testing hypotheses to search for and eliminate root causes of error. In these ways, data quality management most resembles the scientific method.

If measurement and control are the muscle and bone, then three simple management principles form the head and eyes. The first principle is that data quality is defined not in some strict, technical sense but by customers such as patients, doctors, insurance companies, and billing departments. Specifically, data are of high quality if they meet customers' needs. This is an especially demanding approach because each customer may have different needs and uses for the data. As a consequence, they may rate the quality of data provided differently. For instance, while one patient may understand his or her diagnoses perfectly and take appropriate steps, another may

misinterpret the same data and do just the opposite. According to this principle, the same data were of high quality in the first case and of poor quality in the second.

The second principle is that those who create data must be held accountable for its quality. Practically, everyone agrees with this principle in theory, but implementing it is far from trivial. What nurse wants to tell a chief of staff that she cannot read his orders? But experience shows that finding and correcting errors downstream is unreliable, expensive, and time-consuming.

The third principle is that customers and data sources must be tightly coupled if high-quality data are to result. The customer-supplier (C-S) model, depicted in Figure 1, has proven an excellent means of enabling the required communications.

The C-S model features three entities: Customers, as described above, are on the right; suppliers (or data sources) such as laboratories, admissions, and other doctors on the left; and the physician and his or her work processes in the middle. Physicians use data provided by their suppliers to do their work, create new data, and pass relevant data onto their customers. This data flow is illustrated by the left to right arrows in the figure.

More important, the C-S model features four communications channels in the opposite direction from the data flow. These channels help ensure that data needs (i.e., requirements) and feedback, both good and bad, are provided to data sources from customers. Unless physicians seek to actively construct and maintain these channels, they become blocked or noisy. In sum, data sources simply cannot be expected to provide high-quality data without knowing what is expected and understanding how well they are performing.

Prescriptions for Physicians

Assuring data quality can be enormously complex. But data quality can also be quite straightforward. The following are eight prescriptions that physicians can follow to initiate data quality improvement.

1. Treat patients like customers when explaining their diagnoses, courses of treatment, and prognoses. Too often, patients simply do not understand what a physician tells them. Of course, they may be scared and nervous and may not listen well. And they come from a variety of backgrounds. But the physician must make himself or herself understood, whatever it takes. Use simple words for some people, explain in more complicated terms to others, and draw pictures for still others. Make sure they understand exactly how they can best contribute to their care (e.g., taking their prescriptions). Encourage them to ask questions. And, perhaps most important, say "I don't know" when you don't know the answer.

2. Treat the "next person in the process," whoever it is, as a customer. Patients, and their data surrogates, often get lost in the system. Indeed, this is a systemic problem and cannot be solved by one individual. But treating the next doctor (or clinic) the patient will visit, the technician who must follow the orders, the hospital administrator, and the insurance companies who pay the bills as customers is an effective way to promote a culture that values high-quality data. Talk and listen well to one customer every month (say) and ask them what data they really need from you, what data they actually get, what you're doing that helps, and what you do that slows them down; then make necessary improvements.

3. Become intolerant of simple data errors made by others. One reason for data errors in healthcare is that people tolerate them, even accommodate them, in their work processes. For example, the triage nurse in the emergency room, caregivers, technicians, indeed everyone who sees a patient asks the same questions. This increases the chances of error and is considered by some to be bad practice. Furthermore, when you spot errors, provide feedback to the source of the error, as quickly as you can. Don't blame individuals—most of them are doing their best within an imperfect system. Instead, reach out to managers and ask that they find and eliminate root causes of error. Finally, keep a log of the errors you find. Revisit it from time to time to look for patterns.

4. Become extremely intolerant of simple data errors made by yourself or your team. It only stands to reason that you have to be even more demanding of yourself than you are of others. One way an orthopedist could make sure that he or she was operating on the correct limb would be to ask

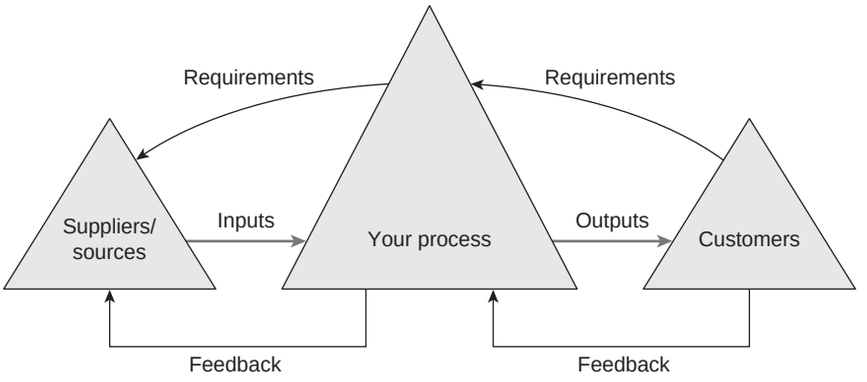


Figure 1 The customer-supplier model

all in the operating theater to concur before starting. Another way could be to write "Not This One" on the left knee when the right is the correct one. Both methods help foolproof the orthopedist's work and that of his or her team. It is an example of a good, proactive way to prevent data errors. Equally important is acknowledging errors when they occur, learning from them, and fixing the root cause to prevent future errors.

5. Make handwriting legible. Much has been written about the frequency and dangers of misread handwriting leading to wrongly filled prescriptions and other errors. As a physician, do whatever you need to do to prevent such error, even typing or learning to print legibly.

6. Put patient records into a computer. So far, healthcare has not yielded much to full automation of patient records. It is a complex technical problem made more complex by not yet fully understood issues such as patient privacy. And automation, in and of itself, is no panacea. But the computer facilitates better record keeping, and better record keeping means better data for making diagnoses, deeper analyses that will improve healthcare, and smoother pathways between all involved. Eventually, all important healthcare data will be digitized and all players interconnected. Physicians can aid the evolution, even if only within their practices and clinics.

7. Learn to distinguish common causes from special causes. If an overtired laboratory technician makes a simple (even if dire) mistake, the solution may be to instruct that person on his or her responsibilities and the importance of coming to work fully rested (importantly, the root cause of the error may be something else, such as a miscalibrated measurement device). But if seven laboratory technicians make the same mistake over a 6-month period, then a root cause analysis must be conducted—even if each admits that he or she was tired! Perhaps the lab is understaffed, perhaps shifts are too long, perhaps a particular piece of equipment becomes erratic as it heats up late in the day. The prescription is to distinguish "common causes" from "special causes." The analogy is not perfect, but common causes are like chronic conditions. They always exist and are inherent to the process or system. Special causes are like acute conditions. They need to be addressed individually and in different ways. Distinguishing common causes from special causes is not easy. In the example above, the bad lab tests come up one at a time, are discovered by different people, and each may be addressed before the next occurs. So spotting them requires a certain aptitude. But there is no substitute. Telling technicians to get more sleep will simply not cool down an overheating piece of equipment.

8. Lead one improvement project every year. Organizations that put forth reasonably diligent efforts often reap order-of-magnitude improvements to data quality. And they've done so without special investment. The secret is completing improvement projects on a regular basis. Frequently, eliminating a relatively few root causes produces dramatic improvement within a department. So define a problem, assemble a team (and personally

lead it), uncover the root cause, and figure out how to make it go away permanently.

Frank M. Guess, Thomas C. Redman, and Mahender P. Singh

See also Constraint Theory

Further Readings

Berwick, D. M. (2003). *Escape fire: Designs for the future of health care*. San Francisco: Jossey-Bass.

Berwick, D. M., Godfrey, A. B., & Roessner, J. (2002). *Curing health care: New strategies for quality improvement*. New York: Wiley. (Other publishers have translated earlier versions into Japanese and into Portuguese.)

Halamka, J. (2008). *Vision for hospital's future HIT*. Retrieved May 29, 2008, from http://www.thehealthcareblog.com/the_health_care_blog/2008/05/vision-for-hosp.html

Redman, T. (2008). *Data driven: Profiting from your most important business asset* (Chapter 3). Boston: Harvard Business School Press.

Decisional Conflict

Every day, people face healthcare decisions involving trade-offs between potential benefits and risks. Which birth control method should I use? Are my symptoms (acne, attention deficit/hyperactivity disorder, hot flashes, chronic pain) bad enough to warrant stronger medication with potentially more serious side effects? Should I have surgery for poorly controlled benign uterine bleeding, back pain, benign prostatic hyperplasia, obesity, osteoarthritis? Should my relative receive care for dementia or terminal illness at home or at a care facility?

Decision making is the process of choosing between alternative courses of action (including inaction). Generally, people choose the option that they perceive will be effective in achieving valued outcomes and in avoiding undesirable outcomes. However, many decisions are *choice dilemmas* or *conflicted decisions*. No alternative will satisfy all personal objectives and none is without its risk of undesirable outcomes.

Among the 2,500 healthcare interventions evaluated by the *Clinical Evidence* group, 13% were classified as "beneficial," 23% as "probably beneficial," 8% as "need to weigh benefits versus risks," 6% as "probably nonbeneficial," 4% as "probably useless or dangerous," and 46%, the largest number, as having insufficient evidence of usefulness. Consequently, patients need help in resolving uncertainty when facing clinical decisions. They may express uncertainty or difficulty in identifying the best alternative due to the risk or uncertainty of outcomes, the need to make value judgments about potential gains versus potential losses, and anticipated regret over the positive aspects of rejected options.

The aim of this entry is to briefly review what has been learned on how patients make difficult decisions by highlighting the value of screening for decisional conflict. The first section summarizes research on patient decisional conflict. It also reviews tools for assessing and addressing decisional needs. The second section reports on the effects of decision support interventions on decisional conflict. The last section highlights the gaps in knowledge and areas needing further research.

Research

Definition of Decisional Conflict

Psychologists Janis and Mann describe decisional conflict as the concurrent opposing tendencies within a person to accept and decline an option. The North American Nursing Diagnosis Association (NANDA) defines decisional conflict as personal uncertainty about which course of action to take when the choice among competing actions involves risk, loss, regret, or challenge to personal life values. Decisional conflict is an intrapersonal psychological construct that is felt by individuals. In lay terms, it refers to one's level of comfort when facing and making a health-related decision.

How Much Do Patients Experience Decisional Conflict?

NANDA defines verbalized uncertainty as the hallmark of decisional conflict (e.g., "I'm not sure which option to choose"). In three large surveys that have been conducted, about half the respondents reported feeling uncertainty about their best

course of action. The first is a Canadian national telephone survey in which 59% of respondents reported feeling unsure about what to choose when facing complex decisions regarding medical or surgical treatments or birth control. In the second case, Légaré measured decisional conflict in 923 patients after they were counseled about options in five family practices; 52% of patients had personal uncertainty about common treatment options. In the third case, Bunn and colleagues conducted a household survey of impoverished women in Santiago, Chile, and found that 54% reported personal uncertainty, commonly about decisions around navigating the healthcare system (where, when, and from whom to seek care).

NANDA describes other manifestations of decisional conflict. The aforementioned Canadian survey reported their prevalence as follows: 77% of respondents questioned their personal values, 61% verbalized concern about undesired outcomes, 40% were preoccupied with the decision, 27% wanted to delay the decision, 27% had signs and symptoms of stress or tension, and 26% wavered between choices.

Contributing Factors

Nonmodifiable Factors

The type of decision can influence decisional conflict. In the Canadian survey, higher rates of physical stress were reported by those who had made decisions about placing a relative in an institution (54%) or medical treatment (46%) as compared with those pondering birth control decisions (23%). Decision delay was more common among those deciding about institutionalization (50%), as compared with those making surgical decisions (20%).

Personal characteristics also influence personal uncertainty. In two studies, which controlled for other potential factors, women reported higher decisional conflict than men. A clinical study of patients considering warfarin therapy found that older people had higher decisional conflict scores. In contrast, the Canadian survey found that younger people had higher decisional conflict scores.

Modifiable Factors

According to NANDA, modifiable factors influencing decisional conflict include deficits in (a) knowledge and expectations (condition, options, benefits, risks, probabilities); (b) clarity of values or priorities (personal desirability or importance of benefits vs. harms); and (c) support and resources (access to advice, support, pressure from others involved in the decision, personal skills, self-confidence, resources). The Canadian survey examined these modifiable factors when controlling for the inherent factors such as type of decision and personal characteristics. More manifestations of decisional conflict were observed with those who had deficits in knowledge as well as support and resources (pressured to select one particular option and unready or unskilled in decision making). When the hallmark of decisional conflict (personal uncertainty about the best course of action) was analyzed separately, those reporting feeling uncertain were also more likely to report problems with the NANDA modifiable factors as compared with those who did not experience uncertainty.

Measuring Decisional Conflict and Modifiable Factors

The Decisional Conflict Scale (DCS) has been developed for research and clinical assessment purposes. It measures personal uncertainty in patients and its modifiable factors such as feeling informed, clear about values, and supported in decision making. This reliable and valid measure shows that greater decisional conflict occurs in those who delay decisions, score lower on knowledge tests, are in the early phases of decision making, and/or have not yet received decision support. High decisional conflict after decision support predicts downstream delay or discontinuance of the chosen option, regret, and the tendency to blame the practitioner for bad outcomes. More recently, the DCS has been adapted for measuring personal uncertainty in health professionals as well.

Decision Support Interventions

Although there are several conceptual frameworks of shared decision making, the Ottawa Decision Support Framework specifically addresses decisional conflict using conceptual definitions and theories from NANDA as well as psychology, social psychology, economics, and social support. The Ottawa framework applies to all participants

involved in decision making, including the individual, couple, or family and their health practitioner. The focus here is on patients' needs and the role of the practitioner in supporting them. As illustrated in Figure 1, the framework has three key elements: (1) decisional needs, (2) decision quality, and (3) decision support. The framework asserts that unresolved decisional needs will have adverse effects on decision quality. However, decision support can improve decision quality by addressing unresolved needs with clinical counseling, decision tools, and coaching.

Decisional Needs

Unresolved decisional needs that adversely affect decision quality include the following: decisional conflict (uncertainty), inadequate knowledge and unrealistic expectations, unclear values, inadequate support or resources, complex decision type, urgent timing, unreceptive stage of decision making, polarized leaning toward an option, and participants' characteristics (e.g., patients' cognitive limitations, poverty, limited education, or physical incapacitation). Therefore, practitioners should be skilled at assessing decision needs by first screening for decisional conflict. A shorter clinical version of the DCS is currently being tested and may hold promising in this regard.

Decision Quality

To help resolve decisional needs, it is important to describe the goal of an intervention. Generally,

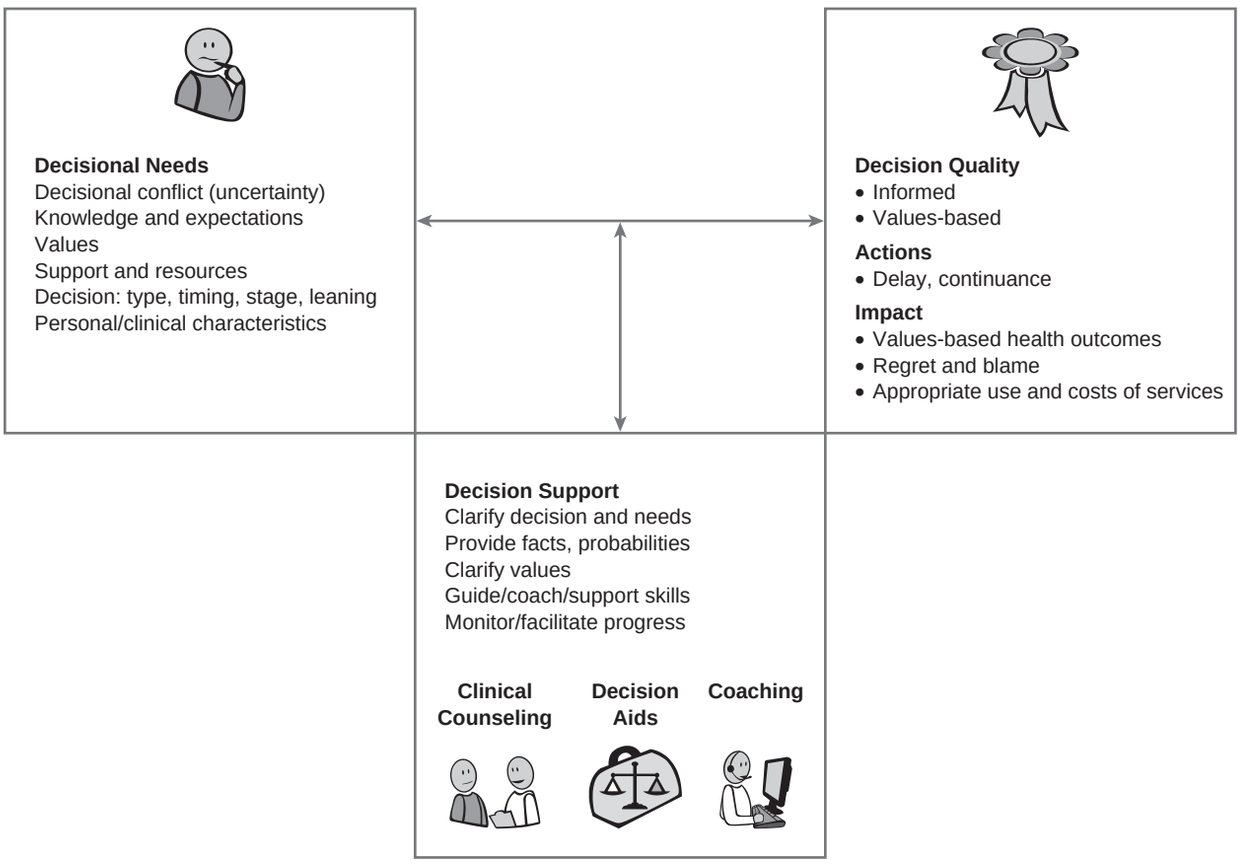


Figure 1 Ottawa decision support framework

Source: O'Connor, A. M. *Ottawa decision support framework to address decisional conflict*. © 2006. Available from http://www.ohri.ca/decisionaid.

medical professionals wish to help people make a "good" decision. However, what is a good decision when there is more than one medically reasonable option and the best choice depends on how a person weighs the known benefits versus harms as well as the scientific uncertainties? Although the issue is not completely resolved, there is emerging consensus that good decisions are ones that are informed and consistent with personal values. Does the person understand the key facts about their condition, options, benefits, and harms? Does the person have realistic expectations (perceptions of chances of benefits and harms)? Is there a match between the option that is chosen and the features of options that matter most to the informed person?

The consequences of decisions are of interest to different groups. For example, behavioral evaluators are often interested in the impact of the decision on behavior. Did a person delay or make a decision? Did the person continue with his or her chosen option? On the other hand, clinicians are interested in the impact of the decision on health outcomes. It is important to note that the types of decisions that create decisional conflict often have no clear best option that has a positive effect on health outcomes. The question the clinician may have to ask is "Did the informed patient achieve the good outcome and avoid the bad outcome that mattered most to him or her?" Health psychologists may find effects on emotions such as regret or blame as most interesting. Health service evaluators and economists often focus on the use of health services and costs.

Decision Support

Decision support is aimed to address a patient's unresolved needs through clinical counseling, decision aids, and coaching. Decision support involves the following: (a) clarifying the decision and the person's needs, (b) providing facts and probabilities, (c) clarifying values, (d) guiding/coaching/supporting in deliberation and communication, and (e) monitoring/facilitating progress.

Health professionals tend to overuse factual information about options and to underuse other strategies. Specific strategies tailored to patients' needs are described in Table 1. Generic and condition-specific decision aids have been developed to assess needs and plan decision support. An example of a generic aid is the Ottawa Personal Decision Guide. It is a framework-based tool to help people and their practitioners structure, record, and communicate decisional needs and plans. The guide incorporates a short version of the decisional conflict scale. It can be self-administered or practitioner-administered. A computer-based 1-page PDF version, as well as a 2-page paper version, is available from the Ottawa Health Research Institute's Web site.

Condition-specific patient decision aids are interventions designed to prepare people for decision making; they do not replace counseling. They help people (a) understand the probable benefits and risks of options, (b) consider the value they place on the benefits versus the risks, and (c) participate actively with their practitioners in deciding about options. According to the International Patient Decision Aids Standards (IPDAS) Collaboration, patient decision aids provide the following: (a) information on the disease/condition, options, benefits, harms, and scientific uncertainties; (b) the probabilities of outcomes tailored to a person's health risk factors; (c) values clarification such as describing outcomes in functional terms, asking patients to consider which benefits and risks matter most to them; and (d) guidance in the steps of decision making and communicating with others. Decision aids may be administered using various media before, during, or after counseling. Most developers are moving toward Web-based materials that can be printed or used online.

Patient decision aids have been developed for a variety of screening, diagnostic, medical, therapeutic, and end-of-life decisions. A list of currently available decision aids is found in the A to Z Inventory of Decision Aids at the Ottawa Health Decision Centre Web site. Reviews of randomized controlled trials of decision aids conclude that they are better than standard care in terms of the following: (a) increasing participation in decision making without increasing anxiety, (b) improving decision quality (improved knowledge of options, benefits, harms), (c) more realistic expectations of the probabilities of benefits and harms, (d) better match between personal values and choices, (e) lowering decisional conflict, and (f) helping undecided people to decide. Patient decision aids may also have a role in addressing underuse and overuse of

Table 1 Decision support strategies tailored to decisional needs

Knowledge deficits. Exposure to information about the health condition, options, and outcomes improves knowledge.

- Help the person access information. Balanced presentations of available options and both potential benefits and harms should be presented in sufficient detail for decision making.
- Adapt medium and pace of information delivery to the person's needs (literacy, numeracy, impairments in sight, hearing, cognition).
- Assess the person's comprehension of the information after it is provided; the focus should be on information that is "essential" for decision making.

Unrealistic expectations. Exposure to probabilities of benefits and harms creates realistic expectations.

- Present probabilities in ways that are understandable to patients, for example, event rates using common denominators and time periods and with mixed frames.
- In labs, the chances of outcomes are perceived to be more likely when they are easier to imagine and when you can identify with the people experiencing them. Therefore, in cases where a person overestimates the chances of an outcome occurring, the practitioner may acknowledge the possibility but then describe anecdotes (vivid stories) in which the outcome did not happen. In cases where a person underestimates the chances of an outcome occurring, the practitioner may acknowledge the possibility but then describe anecdotes in which the outcomes did happen. The use of narratives to change patients' expectations has not been evaluated in clinical trials.

Unclear values (personal importance). Values clarification and communication is under active study and debate.

- A person cannot judge the value of unfamiliar outcomes. Therefore, outcomes need to be described in familiar, simple, and experiential terms to help the person judge their personal importance. This means that, rather than providing a label for an outcome (e.g. pain from osteoarthritis), a person is helped to understand how the outcome will affect him or her physically (characteristics of pain, effects on ability to walk, work, and carry on daily activities), emotionally (discouraged), and socially (withdrawn, avoid social activities). Other examples of meaningful outcomes are (a) for depression: You are more likely to answer the phone or go out with your family; (b) for attention deficit disorder: Your child is more likely to read at grade level or to have friends.
- Ask the person to implicitly consider the personal importance of the positive and negative outcomes. Sometimes decision support includes explicit values clarification exercises using numerical approaches (e.g., rating scales (0 = *not at all important* to 10 = *very important*). The relative value of explicit approaches is under investigation.
- A person needs a strategy for communicating his or her values when discussing the options with others. People, including clinicians and family members, are not very good at judging the values of others. It may be helpful to use rating scales or balance scales showing what is important that can be viewed "at a glance."

Unclear or biased perceptions of others' opinions. The optimal method for presenting the experiences of others in the form of narratives is under active investigation.

- Explain available options to broaden personal awareness of alternatives.
- Present examples of others' choices, in a balanced manner, so that a person is aware that people choose different options and there is no "one size fits all" answer.
- Provide statistics on variation in choice (e.g., the percentage of people who choose the different options that are available; the differences in practitioners' opinions; or the differences in practice guidelines). It is also helpful to present the rationales behind the differing opinions. Often, differences in choices reflect scientific uncertainty, or differences in people's circumstances, tolerance for risk or uncertainty, or values.

Social pressures. Conflict resolution approaches may be useful but have not been tested.

- Explore the nature of the pressure, including its source, the areas of agreement and disagreement, and the reasons behind differences in points of view.
- Guide the person to (a) verify his or her perceptions of others' opinions in case there are misconceptions; (b) focus on those whose opinions matter most; and (c) handle relevant sources of pressure.
- Strategies for dealing with people who are exerting pressure include (a) planning how to communicate information and values; (b) inviting others to discuss their perceptions of options, benefits, harms, and values to find areas of agreement and disagreement; (c) mobilizing social support; and (d) identifying a mediator, if needed. Role play and rehearsal of strategies may help.

Lack of support or resources. Help a person access support or resources needed to make the decision. Resources may include health professionals who are personal advocates, family and friends, support groups, or services from voluntary or government sectors. In some cases the practitioner's support is all that is needed to make the decision.

Lack of skills or confidence in decision making. Provide structured guidance, or coaching in the steps of decision making (i.e., deliberating about a decision) and communicating preferences. There is limited evidence that coaching in addition to information improves decision making.

Preferred role in decision making. The type of guidance will depend on the role people prefer to take in decision making. According to Rothert and Talarczyk, the clinicians' expertise lies in providing information about the options available, their outcomes, the associated risk/probability, and the healthcare resources required and available. The patients' expertise includes their preferences or values and personal, social, and available economic resources.

- Degner and colleagues identified three profiles of preference for decisional control: those who want to *keep*, *share,* or *give away* control of decision making. "Keepers" might guide the deliberation and ask their practitioner for input on the scientific facts. Practitioners might start by providing guidance to "sharers," who would then become actively involved in the decision. A more advisory role might be used by practitioners with those who want to give away control, who would then be asked to provide informed consent. It is important, however, for practitioners not to take preferred roles in decision making completely at face value; providing people with decision support often increases their desire for active participation in decision making. Therefore, people need adequate information about the issues and time to consider which decision-making role they prefer to take.

Decision type, timing, stage, and leaning. Practitioners need to tailor decision support to the type of decision. For example, the approach may differ if the focus is on screening for prostate cancer, treatment of early-stage disease, treatment of recurrence, or end-of-life care. Tailoring support also depends on timing. Short timelines to make big decisions often increases stress, but very long timelines may increase decision delay. In the very early and very late stages of decision making it is important to gauge a person's receptivity to new information and further deliberation. Otherwise, decision support may be irritating or unproductive. The aim of decision support is to help the person progress in his or her stage of decision making, not necessarily "change." Sometimes "maintaining the status quo" is a reasonable option (e.g., forgoing PSA testing, amniocentesis, or hormone therapy).

Personal and clinical characteristics. Decision support should be gender-sensitive and appropriate for an individual's age, developmental stage, education, socioeconomic status, and ethnicity. Adjustments should be made to accommodate a person's physical, emotional, and cognitive capacities. Involving the family or a personal advocate is important when the person's capacities are limited. The characteristics of the practitioner will also influence decision support, based on a person's training, experience, and counseling style.

Monitoring and facilitating progress. Once needs have been addressed, monitor progress in resolving needs, moving through the stages of decision making, and achieving the goal of decision quality (informed, choice matches features that matter most to the informed patient). Decision tools help a patient consider and become committed to taking the next steps.

options. They reduce the uptake of discretionary surgical options that informed people don't value when baseline rates of these procedures are high. They also increase the uptake of colon cancer screening options, which are underused, and lower the rates of prostate cancer screening tests, which are overused.

Gaps in Research

Although decisional conflict is common and decision support interventions can address its modifiable contributing factors, there are three major knowledge gaps. First, most large studies describing people's decisional conflict are from North America. Therefore, more descriptive research is needed on the prevalence of decisional conflict and related factors for the many decisions people face in more diverse populations. Second, the Decisional Conflict Scale elicits people's "overall comfort level" with their knowledge, values, and support. These comfort levels are only modestly correlated with a person's knowledge test scores and their match between their values and the chosen option. Researchers still don't know the relative contribution of each of these variables to downstream behavior. Third, practitioners should be trained to recognize and screen for decisional conflict in their patients so they can refer those who require assistance in resolving their decisional needs. A 4-item clinical version of the Decisional Conflict Scale may hold promise in this regard.

Annette O'Connor and France Légaré

See also Decision Making in Advanced Disease; Patient Decision Aids; Shared Decision Making

Further Readings

Bunn, H., Lange, I., Urrutia, M., Campos, M. S., Campos, S., Jaimovich, S., et al. (2006). Health preferences and decision-making needs of disadvantaged women. *Journal of Advanced Nursing, 56*(3), 247–260.

Carroll-Johnson, R. M., & Paquette, M. (1994). *Classification of nursing diagnoses: Proceedings of the tenth conference*. Philadelphia: Lippincott.

Gattellari, M., & Ward, J. E. (2005). Men's reactions to disclosed and undisclosed opportunistic PSA screening for prostate cancer. *Medical Journal of Australia, 182*(8), 386–389.

Janis, I. L., & Mann, L. (1977). *Decision making*. New York: Free Press.

O'Connor, A. M. (1995). Validation of a decisional conflict scale. *Medical Decision Making, 15*(1), 25–30.

O'Connor, A. M., Drake, E. R., Wells, G. A., Tugwell, P., Laupacis, A., & Elmslie, T. (2003). A survey of the decision-making needs of Canadians faced with complex health decisions. *Health Expectations, 6,* 97–109.

O'Connor, A. M., Stacey, D., Entwistle, V., Llewellyn-Thomas, H., Rovner, D., Holmes-Rovner, M., et al. (2003). Decision aids for people facing health treatment or screening decisions. *Cochrane database of systematic reviews,* 1. Art. No.: CD001431. DOI: 10.1002/14651858.CD001431.

Ottawa Health Decision Centre. (n.d.). *A to Z inventory of decision aids*. Retrieved February 4, 2009, from http://decisionaid.ohri.ca/AZinvent.php

Ottawa Health Research Institute. *Ottawa personal decision guide*. Retrieved February 4, 2009, from http://decisionaid.ohri.ca/decguide.html

Sepucha, K. R., Fowler, F. J., Jr., & Mulley, A. G., Jr. (2004, October 7). Policy support for patient-centered care: The need for measurable improvements in decision quality (Web exclusive). *Health Affairs,* DOI: 10.1377/hlthaff.var.54

Decision Analyses, Common Errors Made in Conducting

Decision analytic modeling (DAM) has been increasingly used within the past 30 years to synthesize clinical and economic evidence and support both clinical and policy-level decision making. Decision models often represent complex decision and synthesize data from a variety of sources, and they may be difficult to validate and interpret. Thus, while DAM can be extremely useful, it is also difficult to do well. Errors are common among neophytes and not uncommon even in published decision analyses. This entry reviews the steps associated with constructing a decision model and describes several of the most common errors in model construction, analysis, and interpretation. It considers both conceptual errors in

model construction and errors of computation or calculation. Although DAM is commonly used in economic evaluation, the purview of this entry extends only to model-related aspects of economic evaluation.

Comparators

Every decision analysis compares at least two options. If the decision is a clinical one (e.g., how should localized prostate cancer be treated?) all *feasible* and *practical* options should be considered. These might include doing nothing (or active surveillance), surgery, radiation, brachytherapy, or cryotherapy, and more. If the decision is a policy decision (say, whether a national human papillomavirus vaccination program should be funded), the same criteria apply: Feasible and practical options might include no vaccination, universal vaccination, vaccination targeted at high-risk groups, vaccination targeted at specific age groups, and more. Feasible and practical are clearly subject to interpretation, but the key ideas are that all options that stand a realistic chance of being implemented (feasibility) should be examined, given the resources available to address the problem (practicality).

The decision analysis neophyte often is reluctant to include many options because of concerns that the model will become unmanageably complex. As a result, many models consider only the two or three most intuitively attractive options. Options such as "do nothing" or "supportive care only," or alternate frequencies or intensities of an intervention may be avoided. This is acceptable if the goal is to gain experience in modeling, but it is not acceptable if the goal is to choose the best therapeutic or policy option.

More advanced analysts may also inappropriately constrain the potential options considered. This may be because of a desire to adhere closely to the best quality evidence published in high-impact journals. Or it may be a strategic decision to put a new drug or device in the best possible light by choosing a plausible but weak comparator or by avoiding comparisons across types of interventions (e.g., comparing drugs only to drugs but not to surgery). Regardless of the reason, inappropriately constraining the set of comparators is a common and serious error in modeling.

Model Structure

Decision models represent potential outcomes of alternate strategies using models, which may be simple decision trees, discrete-time state-transition (i.e., Markov) models, discrete-event simulation models, or dynamic infectious disease models. Models may be simple or complex, but should correspond to an underlying theory or biological model of disease.

Underrepresentation

In particular, models must capture important differences across strategies. For example, if two strategies differ mainly in adverse effect profile, the structure of the model must represent adverse effects. An important and common example of underrepresentation is the use of cohort simulation models to represent decision problems in which events within the cohort affect members outside the cohort. For example, vaccination will protect individuals within a cohort, but the herd immunity associated with high rates of coverage will confer benefits beyond the cohort. Failure to represent these additional benefits of vaccination will inaccurately represent the true effect of vaccination on the entire population.

Unclear or Inappropriate Target Population

Neophytes in particular are often unclear about which population is being represented in the model. Models should represent a specific group or population. This includes a value or distribution for age, sex, disease severity, and prevalence and type of comorbid illness.

Perspective

When the perspective of a decision problem extends beyond the individual patient, modeling outcomes only for the patient represents an error. For example, the question of optimal approaches to testing for fetal abnormalities potentially affects the parents, the fetus, and other family members. While appropriate valuation of these outcomes is difficult, constraining the decision problem to one perspective is incorrect, unless the perspective taken is explicitly that of only one individual. Similar errors are often present in models that represent health outcomes of

children but ignore families, and models that consider the elderly or disabled, but ignore caregivers.

Time Horizon Bias

Every model represents a limited period of time. The time horizon of the model should extend to or beyond the point at which there are no differences between strategies in life expectancy and quality of life. Neophytes often prefer short time horizons because this reduces model complexity. For example, a model designed to compare computed tomography with abdominal ultrasound imaging for suspected appendicitis in children might focus on short-term events around the abdominal pain, imaging, surgery, and immediate perioperative outcomes. However, one of the main concerns of parents and clinicians is avoiding unnecessary radiation. Thus, representing the long-term cancer risk associated with radiation is an essential aspect to correctly representing this decision problem.

This error is also common among more experienced modelers. For example, decision models often closely follow randomized trials. Because the time horizon of trials is often short, important differences in quality of life and mortality that extend beyond the horizon of the trials may not be represented in models, and bias is therefore introduced. While experienced modelers are often aware of this problem, they may adopt inappropriately short time horizons, in the interest of enhancing the apparent scientific credibility of the model to clinical or policy audiences, by reducing the complexity and the number of assumptions in the model. This is a common and often serious error.

Mortality From Other Causes

While focusing on a particular disease, modelers may neglect to represent competing causes of mortality in a decision model. This means that subjects in the model remain at risk of disease-related adverse events for longer. Differences across strategies may be exaggerated, and error is therefore introduced.

Half-Cycle Correction Problems

As the name suggests, discrete-time cohort simulation models represent risk in discrete time periods. Models represent events as occurring at the beginning or end of discrete time intervals, whereas events can actually occur throughout the interval. Half-cycle correction adjusts for this property of discrete time models by adding (or subtracting) the value (life expectancy, quality-adjusted life expectancy, or cost) associated with half of one cycle length. Neophytes often neglect to introduce a half-cycle correction or assign the incorrect sign to the correction (subtraction instead of addition of a half-cycle or vice versa).

Symmetry

Symmetry refers to consistent representation of model events and outcomes across strategies. Errors of symmetry often occur when modelers use different structural elements (e.g., tree fragments, Markov states) or variable names and expressions across different strategies that represent the same components of the decision problem. Events and outcomes may not be represented or be represented in a different manner when alternate structures are used. Experienced modelers frequently use a common model structure for all strategies to avoid this error.

Model Data

Obtaining, analyzing, and adjusting data for use in decision analytic models represent perhaps the greatest challenge in developing valid models.

Lamppost Bias

The availability of data may constrain and shape the structure of decision models. While some degree of adaptation may be necessary, it is an error to allow the available evidence to play a fundamental role in shaping the structure of the model, just as it is an error to confine a search to the location of the available light. The structure of the model must be shaped primarily by the decision problem, not the availability of evidence. This refers to inclusion of comparators and other aspects of structure, as described above. For example, representing only treatments or adverse effects for which there is strong evidence represents an error.

Rate to Probability Conversion

Transitions between states in Markov models for a discrete time period are commonly expressed

using probabilities, referred to as transition probabilities. However, these data are commonly abstracted from the literature in the form of rates. While rates are close to probabilities for very small values, this is less true for larger values. A common neophyte error is to neglect the conversion of rates, as obtained from the literature, into probabilities.

Errors in Value Structure

A complete set of utilities for important health outcomes is rarely available from a single source. Many variables may affect utility values reported in the literature, including source of preferences (patients vs. experts or members of the general public), scaling method, use of direct or indirect utility elicitation, instrument used for indirect utility elicitation (e.g., Health Utilities Index vs. EQ-5D), and computer-assisted versus interview-assisted elicitation, among others. Judgment must therefore be applied when utilities from widely disparate sources are used. In particular, the ordinal relationships among important health outcomes should, in general, be reflected in ordinal scores among utility values used in the model. Uncritical use of published data may result in a model value structure that is not internally consistent or does not correspond with an a priori model of disease. A common error is to overweight the size, quality, or place of publication of a utility study, and underweight consistency and appropriate ordinal relationships among model values.

Adjustment for Age and Comorbidity-Related Utility

As patients age, utility scores for current health status decline. This may be due to acquired comorbidity, age-related decline in functional status, change in preference structure as patients age, or a combination of these. A frequent error in representing the value structure of a model is to assume that the utility of individuals without the disease in question can be assigned a value of 1.0. This assumption results in overestimation of the difference in utility scores between those with and without disease, and correspondingly may overestimate the benefit of treatment or prevention.

A companion error is to adjust for age-related comorbidity among patients without disease but fail to adjust among patients with disease. For example, if the mean utility for 70-year-old individuals is .90, and the disutility (1 – utility) associated with renal failure is .40, the utility for a 70-year-old with renal failure should reflect the contribution of both factors. A common method of adjustment is to assume that utility is multiplicative, and simply multiply (e.g., .90 × (1 – .40) = .54).

Internal Consistency

Internal consistency refers to the internal mathematical structure of a decision model. Internal consistency is most frequently evaluated using univariate sensitivity analysis. Changes in values of a single variable should have predictable effects on model outputs. A common error is to evaluate inconsistency in a haphazard or unsystematic way. Every variable should be tested across a broad range. Any deviation from predicted behavior represents either a failure of internal consistency (a "bug") or an insight, but more commonly the former. An equally common error is to identify internal consistency problems but fail to correct them because of time or resource constraints.

Murray Krahn and Ava John-Baptiste

See also Cost-Utility Analysis; Decision Tree: Introduction; Decision Trees, Construction; Decision Trees: Sensitivity Analysis, Deterministic; Markov Models; Markov Models, Applications to Medical Decision Making; Markov Models, Cycles

Further Readings

Krahn, M. D., Naglie, G., Naimark, D., Redelmeier, D. A., & Detsky, A. S. (1997). Primer on medical decision analysis: Part 4: Analyzing the model and interpreting the results. *Medical Decision Making, 17,* 142–151.

Philips, Z., Bojke, L., Sculpher, M., Claxton, K., & Golder, S. (2006). Good practice guidelines for decision-analytic modelling in health technology assessment: A review and consolidation of quality assessment. *PharmacoEconomics, 24,* 355–371.

Smith, K. J., Barnato, A. E., & Roberts, M. S. (2006). Teaching medical decision modeling: A qualitative description of student errors and curriculum responses. *Medical Decision Making, 26,* 583–588.

Decision Board

The decision board is a visual aid to help clinicians present information about different courses of action in an efficient and standardized manner. The sole goal of the decision board is to improve communication (i.e., improve information transfer). The decision board can (and has been) successfully modified for other uses: to describe the options to choose between in willingness-to-pay (WTP) surveys and to elicit treatment options of potential patients and patients for policy decision making.

Context

What is the best treatment for an individual patient? It is important to realize that there is often no right or wrong choice. For example, the case of adjuvant chemotherapy for patients with breast cancer presents a situation of choice between potential morbidity and disability now (due to therapy, if chosen) and potential morbidity and inconvenience later (due to recurrence of the disease). The uncertainty of the outcome at the individual level (i.e., there is no way to know in advance what will happen to an individual patient) further complicates the problem and makes the choice a very difficult one. Thus, the question of which course of action to take becomes a preference judgment. However, to make an informed preference judgment one needs to know the relevant courses of action and their potential risks and benefits.

Treatment decision making typically takes place within the context of a doctor-patient encounter. This process is both complex and dynamic and can be done using different approaches (i.e., paternalistic, shared, and informed approaches, with myriad in-between approaches that combine components of different approaches). Besides the paternalistic approach, the vast majority of approaches require that the physician inform the patient about the relevant courses of action and their potential benefits and risks. This is due to the fact that, typically, a doctor is required to determine the diagnosis about the type and severity of the patient's illness, on the basis of which the determination of the available courses of action will be made.

Communication difficulties between doctors and their patients are a well-known problem. It has been argued that doctors and patients talk to each other with different voices. The voice of medicine is characterized by medical terminology, descriptions of medical symptoms, and the classification of these within a reductionist biomedical model. The voice of patients, on the other hand, is characterized by nontechnical discourse about the subjective experience of illness within the context of social relationships and the patient's everyday world. Many studies document that communication misunderstandings experienced by doctors and their patients are common.

Goals and Benefits

The decision board should provide all the relevant clinical information that a patient needs to make a decision or participate in the decision-making process, if he or she wishes to do so. The potential morbidity and mortality effects are described in a probabilistic manner, acknowledging the fact that the final outcome and course of any intervention are uncertain. In other words, there is no way to predict what will happen to an individual patient. Scenarios are constructed to describe the treatment options (e.g., in the case of early-stage breast cancer after surgery, adjuvant chemotherapy vs. no further treatment) and the potential side effects (e.g., in the case of chemotherapy, hair loss, stomach upset, vomiting). Scenarios are also constructed to describe the potential outcome of each treatment option (e.g., in the case of chemotherapy, a cancer-free scenario and a cancer returns scenario).

The decision board can be seen as a specific form of decision aid. Various types of decision aids have been developed over the years designed to help participants in the medical encounter to make treatment decisions. In terms of the goals to be achieved by using a decision aid, different authors have different ideas about what the primary goal should be. Two goals are most commonly cited: (1) to provide patients with information on the potential benefits and risks of different options and (2) to help patients clarify their values so that they will make treatment choices that are consistent with their values. Some other goals mentioned are lowering the cost of care, reduction of decisional conflict, improving patient satisfaction with

the decision-making process, encouraging patients to be more involved in the decision-making process, and improving clinical outcomes. While there is agreement on the role of decision aids in information transfer (or knowledge acquisition), there is still debate on whether the other goals are appropriate and feasible.

The sole goal of the decision board is to improve communication (i.e., improve information transfer) about potential courses of action by presenting information simply, using spoken and written language supported by the use of visual aids and relying on repetition. It has been found that when the decision board is administered by the clinician (e.g., doctor), it helps build the relationship between the clinician and the patient. It helps facilitate two-way communication and encourages questions from patients and responses from the clinician. This should be seen as an additional benefit and cannot be assumed to happen every time a decision board is used. In many cases, a relationship between the doctor and the patient already exists (e.g., in the case of a family physician or a specialist treating a chronic condition). In other cases (e.g., a cancer patient meeting his or her oncologist the first time), building the relationship between the patient and his or her doctor might be important.

Format

It is important to emphasize that the decision board does not have a fixed format. It can be seen as a "concept" that leaves "artistic freedom" for its creators to modify according to the special features of the medical problem dealt with. However, within that artistic freedom, a few "rules" should be kept to. For example, after describing all the information about the different courses of action (i.e., the different "pieces of the puzzle"), a visual aid where all these elements are integrated (i.e., a full picture) should be available. This is because it is known that most individuals cannot judge a situation only by valuing the different parts separately. They need to see the full picture to be able to compare the different options. Also, a take-home version should be available for patients, because few decisions are so urgent as to need immediate answers. Where it is feasible, agreeing to defer the decision to allow time for further understanding of the options and for deliberation would be helpful. Finally, the decision board should be easy to administer, inexpensive to produce, and easily modified to incorporate local variations in practice or new clinical information that becomes available.

The first decision board was developed in 1990 for use in the situation of adjuvant chemotherapy for node-negative breast cancer. The board was made of foamcore, which was found to be both lightweight and more durable than cardboard. With the advent of computer capability, the decision board was computerized, too. The move to a computer-based version has opened new opportunities (e.g., ease of providing more tailored information, ease of supplementing core information on an individual basis, and the ability to present technical information in alternative ways to suit patients' needs) but created other challenges (e.g., difficulties in presenting the full picture due to constraints regarding screen size). Examples of schematic presentations of decision boards can be found in articles mentioned in the Further Readings section.

Research Findings

The decision board was tested in several well-conducted studies (including several randomized controlled trials, where it was compared with current practice). It was found to be clear and understandable, valid, and reliable, and improves information transfer (e.g., knowledge about potential treatment options, their potential benefits and risks). It was also found to be easy to administer and use. It was well accepted by clinicians and patients and is currently being used as part of regular practice in different places. Even though it is not the goal of the decision board, it is interesting to note that it was also found that patient satisfaction with decision making was improved. When tested, it was found that the average time of consultation with the decision board was not increased as compared with the average time of consultation without the board. While it is not the goal of the decision board to maintain (or even reduce) the time of consultation, this is still an interesting finding.

Nonclinical Uses

The decision board can be easily modified to serve as an instrument describing the options to choose between in WTP surveys. In WTP studies, individuals

are asked to (a) choose a preferred course of action (or program) and then (b) indicate the maximum amount they are willing to pay to ensure that their preferred option will be available if needed. A weakness that was identified in many WTP questionnaires is the lack of clarity in describing the options compared in terms of their potential benefits and risks. This cast doubts on the validity of the WTP values provided by respondents. The modified decision board was offered as a way to explain the choices to participants in surveys. Because the concept was found to be useful in explaining treatment options to real patients (who are often anxious and confused), it seemed that it would work with healthy people. Indeed, using a modified decision board to explain the different courses of action was shown to be helpful. It was also felt that the use of the decision board can also enhance the credibility of the results among users of information, as it makes explicit the exact question faced by the respondent in the study. However, this point has not been tested yet.

A modification of the decision board is required, because typically subjects in a WTP survey are not patients who suffer from the disease. They should be members of the general population who are typically healthy people. The modification of the decision board depends on whether the WTP question is being asked ex post (i.e., WTP at the point of consumption) or ex ante (i.e., insurance-based approach). For an ex post–type WTP instrument, a preamble is required to describe the medical conditions for which the different courses of action described are required. This helps healthy respondents imagine that they are at the point of consumption of the services described. For an ex ante–type WTP instrument, the preamble should have additional information about the risk of the condition/disease to the individual (or loved ones or other people in the population, depending on the nature of the disease and the question asked). In other words, for an insurance-based question, the respondents need to know the likelihood of their being at the point of consumption.

The decision board can also be used to elicit the preferences about treatment options of potential patients and patients for policy decision making (rather than clinical decision making). An example of such use is a study which attempted to assess if potential patients prefer tissue plasminogen activator (tPA) over streptokinase (SK). In patients with acute myocardial infarction, tPA (compared with SK) has been shown to reduce the 30-day mortality rate at the expense of an increased rate of stroke. The assumption in the literature was that, were it not for cost issues (tPA is much more expensive), all patients presenting with myocardial infarction would choose tPA. A decision board describing the treatment options (without mention of the drug names) was used in face-to-face interviews with individuals at risk for having the event in two hospitals (as it is not possible to ask patients who are experiencing the event). It was found that a substantial proportion of individuals who could potentially require thrombolytic therapy chose SK over tPA. This finding, if found to be consistent, has significant implications for clinical decision making as well as economic and policy implications.

Amiram Gafni

See also Patient Decision Aids; Shared Decision Making; Willingness to Pay

Further Readings

Charles, C., Gafni, A., & Whelan, T. (1999). Decision making in the physician-patient encounter: Revisiting the shared treatment decision making model. *Social Science and Medicine, 49,* 651–661.

Charles, C., Gafni, A., Whelan, T., & O'Brien, M. A. (2005). Treatment decision aids: Conceptual issues and future directions. *Health Expectations, 8,* 114–125.

Gafni, A. (1997). Willingness-to-pay in the context of an economic evaluation of healthcare programs: Theory and practice. *American Journal of Managed Care, 3,* S21–S32.

Heyland, D., Gafni, A., & Levine, M. (2000). Do potential patients prefer tissue plasminogen activator (tPA) over streptokinase (SK)? An evaluation of the risks and benefits from the patient perspective. *Journal of Clinical Epidemiology, 53,* 888–894.

Levine, M. N., Gafni, A., Markham, B., & MacFarlane, D. (1992). A bedside decision instrument to elicit a patient's preference concerning adjuvant chemotherapy for breast cancer. *Annals of Internal Medicine, 117,* 53–58.

Matthews, D., Rocchi, A., & Gafni, A. (2002). Putting your money where your mouth is: Willingness-to-pay for dental gel. *PharmacoEconomics, 20,* 245–255.

Nelson, W. L., Han, P. K. J., Fagerlin, A., Stefanek, M., & Ubel, P. A. (2007). Rethinking the objectives of decision aids: A call for conceptual clarity. *Medical Decision Making, 27,* 609–618.

O'Brien, B., & Gafni, A. (1996). When do the "dollars" make sense? Toward a conceptual framework for contingent valuation studies in health care. *Medical Decision Making, 16,* 288–302.

Whelan, T., Levine, M., Gafni, A., Sanders, K., Willan, A., Mirsky, D., et al. (1999). Mastectomy versus lumpectomy? Helping women make informed choices. *Journal of Clinical Oncology, 17,* 1727–1735.

Whelan, T., Levine, M., Willan, A., Gafni, A., Sanders, K., Mirsky, D., et al. (2004). Effect of a decision aid on knowledge and treatment decision making for breast cancer surgery: A randomized trial. *Journal of the American Medical Association, 292,* 435–441.

Whelan, T. J., Sawka, C., Levine, M., Gafni, A., Reyno, L., Willan, A. R., et al. (2003). Helping patients making informed choices: A randomized trial of a decision aid for adjuvant chemotherapy in node negative breast cancer. *Journal of the National Cancer Institute, 95,* 581–587.

Decision Curve Analysis

Decision curve analysis is a straightforward technique for evaluating diagnostic tests, prediction models, and molecular markers. Unlike traditional biostatistical techniques, it can provide information as to a test's clinical value, but unlike traditional decision analytic techniques, it does not require patient preferences or formal estimation of the health value of various health outcomes: Only a general clinical estimate is required. Differences between biostatistical techniques, decision-analytic techniques, and decision curve analysis are shown in Table 1.

A common clinical problem is when a physician can easily obtain information about *T*—the result of a diagnostic test, the level of a molecular marker, or a probability from a statistical prediction model—but wants to know *D,* whether or not a patient has, or will develop, a certain disease state. From a research perspective, the analyst's task is to determine whether doctors should obtain *T* in order to make decision about *D*.

In this entry's motivating example, *D* is whether the patient has prostate cancer and is used in decisions about whether or not to conduct a prostate biopsy; *T* may be the result of a digital rectal examination (normal vs. abnormal) or the level of prostate-specific antigen (PSA), or it may be a prediction model based on multiple factors (such as age, race, and family history). This example is used to discuss drawbacks of the traditional biostatistical and decision analytic approaches to evaluating the value of *T,* whether a binary diagnostic test, a statistical prediction model, or a molecular marker. Then this entry discusses the novel method of decision curve analysis.

Biostatistical Approaches and Their Drawbacks

Biostatistical analysis of prediction models, diagnostic tests, and molecular markers is largely concerned with accuracy. Such metrics have been criticized by decision analysts as having little clinical value. An accurate test, prediction model, or marker is, in general, more likely to be useful than one less accurate, but it is difficult to know for any specific situation whether the accuracy of a test, prediction model, or marker is high enough to warrant implementation in the clinic. For example, if a new blood marker for prostate cancer increased the area under the curve (AUC) of an established prediction model from .77 to .79, would this be sufficient to justify its clinical use?

Decision Analytic Approaches and Their Drawbacks

Decision analysis formally incorporates the consequences of test results and can therefore be used to determine whether use of a prediction model, diagnostic test, or molecular marker to aid decision making would improve clinical outcome. A typical approach is to construct a decision tree as shown in Figure 1. We denote probabilities and values of each health outcome, respectively, as p_{xy} and as b_{xy}, where x is an indicator for the test result and y is the indicator for disease. To determine the optimal decision, the values of each outcome are multiplied by their probability and summed for each decision; the decision with the highest expected value is chosen.

To obtain p_{xy}s for a statistical model or molecular marker, the analyst has to choose a cut point in order to dichotomize results into positive and negative. Different analysts can disagree about the appropriate cut point, entailing that the analysis may need to be run several times for a range of

Table 1 Comparison of decision curve analysis with traditional statistical and decision analysis

	Traditional Statistical Analysis	*Traditional Decision Analysis*	*Decision Curve Analysis*
Mathematics	Simple	Can be complex	Simple
Additional data	Not required	Patient preferences, costs or effectiveness	Informal, general estimates
Endpoints	Binary or continuous	Continuous endpoints problematic	Binary or continuous
Assess clinical value?	No	Yes	Yes

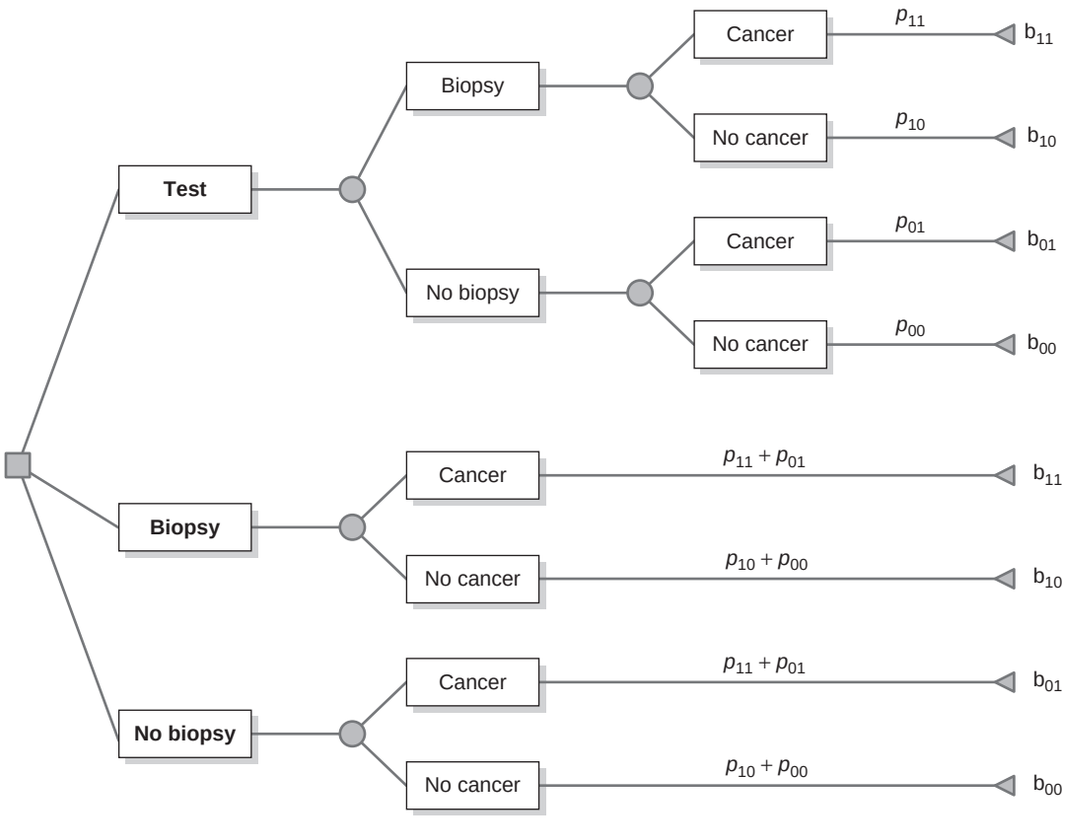


Figure 1 Traditional decision tree to evaluate a test for prostate cancer in men with elevated prostate-specific antigen (PSA)

reasonable alternatives. Choice of b_{xy}s can be even more difficult. A b_{xy} may require data from the literature that can be hard to come by or controversial; moreover, a b_{xy} may require judgments that may reasonably vary from patient to patient. The need for additional data may be one of the reasons why the number of biostatistical evaluations of tests, prediction models, and markers dwarfs the number of decision analyses: In one systematic review of more than 100 papers on cancer markers, researchers failed to find a single decision analysis.

Theoretical Background to Decision Curve Analysis

Traditional decision analysis, unlike biostatistical analyses, can determine the clinical value of a test, prediction model, or marker; however, it requires additional parameters, p_{xy} and b_{xy}, that can be difficult to specify. Decision curve analysis starts by showing that p_{xy}s and b_{xy}s can be related through a simple, clinically interpretable quantity: the threshold probability of disease at which a patient or clinician would opt for further action. The threshold probability of disease is then used to calculate the "net benefit" of different treatment strategies, such as biopsying all men, or biopsying on the basis of a marker. The strategy with the highest net benefit should be used in the clinic.

Threshold Probability of Disease

It is highly unlikely that a man would consent to a prostate biopsy if he was told that his probability of prostate cancer was 1%. Conversely, if the man was told that he had a 99% probability of cancer, there is little doubt as to his course of action. If we were to increase the probability of cancer gradually from 1% to 99%, there would come a point where a man would be unsure of whether or not to be biopsied. We define p_t, the threshold probability of disease for taking some action, such as biopsying a man for prostate cancer: If a patient's estimated probability of disease is greater than p_t he will opt for biopsy; if it is less than p_t, he will not opt for biopsy. When the probability of disease is equal to the threshold probability p_t, the benefits of opting for biopsy or no biopsy are equal:

$$b_{11} \times p_t + b_{10} \times (1 - p_t) = b_{01} \times p_t + b_{00} \times (1 - p_t),$$

and, therefore,

$$\frac{b_{00} - b_{10}}{b_{11} - b_{01}} = \frac{p_t}{1 - p_t}. \quad (1)$$

Now $b_{00} - b_{10}$ is the benefit of true negative result compared with a false positive result; in clinical terms, the benefit of avoiding unnecessary treatment such as a negative biopsy. Comparably, $b_{11} - b_{01}$ is the benefit of a true positive result compared with a false negative result; in other words, the benefit of treatment where it is indicated, such as a biopsy in a man with cancer. Equation 1 therefore tells us that the threshold probability at which a patient will opt for treatment is informative of how a patient weighs the relative benefit of appropriate treatment as compared with the benefit of avoiding unnecessary treatment. As an example, if a man stated that he would opt for biopsy if his risk of prostate cancer were 20% or higher, but not if his risk were less than 20%, we can say that this man thinks that finding a prostate cancer early is worth four times more (i.e., .20 ÷ (1 – .20)) than avoiding the risks, pain, and inconvenience of an unnecessary biopsy.

We can rearrange Equation 1 to obtain

$$-(b_{10} - b_{00}) = (b_{11} - b_{01})\left(\frac{p_t}{1 - p_t}\right). \quad (2)$$

Net Benefit

The idea of net benefit is similar to that of profit. A business owner choosing between several possible investment opportunities will estimate the expected income and expenditure for each and then choose the option that maximizes the difference between the two.

In medicine, the corollary to income and expenditure is benefit and harm; more specifically, in the case of a diagnostic test, prediction model, or molecular marker, benefit is true cases identified and appropriately treated (T^+, D^+, or true positives); harm is unnecessary treatment (T^+, D^-, or false positives). In our prostate cancer example, we want to biopsy men with prostate cancer (true positives) and avoid unnecessary biopsies of men without cancer (false positives). However, "finding cancer" and "avoiding unnecessary biopsy" are not equivalent in value. Equation 2 gives the number of false positives we would exchange for a true positive in terms of the threshold probability. This becomes our way to convert between "finding cancer" and "avoiding unnecessary biopsy." Where n is the total number of men in the cohort, net benefit is given as

$$\frac{\text{True positives} - \text{False positives} \times \left(\frac{p_t}{1 - p_t}\right)}{n}. \quad (3)$$

As an illustration, in a cohort of 728 men undergoing biopsy, 202 had cancer; 479 of the men had a risk of cancer of 20% or higher using a prediction model, of whom 163 had cancer. The

net benefit at a threshold probability of 20% for biopsying all men is (202 (true positives) − 526 (false positives) × .25) ÷ 728 = .0968; the net benefit of using the prediction model is (163 (true positives) − 316 (false positives) ×.25) ÷ 728 = .1154. Hence, use of the prediction model would lead to a higher net benefit and better clinical outcome.

The unit of net benefit is the number of true positives per patient: It therefore has a maximum at the prevalence, but no minimum. A net benefit has a simple clinical interpretation. For example, a difference in net benefit between two prediction models of .02 could be interpreted as "Using Prediction Model A instead of Prediction Model B is equivalent to a strategy that increased the number of cancers found by 2 per 100 patients, without changing the number of unnecessary biopsies conducted."

Decision Curve Analysis

The threshold probability p_t can be used both to define positive and negative test results and to provide a decision analytic weight. The first stage of decision curve analysis is therefore to use logistic regression to convert the results of the test, marker, or prediction model into a predicted probability of disease $\hat{p}$. Decision curve analysis then consists of the following steps.

1. Choose a threshold probability (p_t) for treatment. Here, "treatment" is defined generally as any further action, such as drug therapy, surgery, further diagnostic work-up, or a change in monitoring, depending on the particular clinical situation.
2. Define patients as test positive if $\hat{p} \geq p_t$ and negative otherwise. For a binary diagnostic test, $\hat{p}$ is 1 for positive and 0 for negative.
3. Calculate net benefit of the test, marker, or prediction model using the formula for net benefit in Equation 3.
4. Calculate clinical net benefit for the strategy of treating all patients. Where π is the prevalence, this simplifies to

$$\pi - (1 - \pi) \times \left(\frac{p_t}{1 - p_t} \right). \quad (4)$$

5. The net benefit for the strategy of treating no patients is defined as zero.
6. The optimal strategy is that with the highest clinical net benefit.
7. Repeat Steps 1 to 6 for a range of threshold probabilities.
8. Plot the net benefit of each strategy against threshold probabilities.

Interpretation of Decision Curves

To illustrate decision curve analysis, data from men undergoing prostate biopsy in Göteborg, Sweden, as part of a randomized trial of PSA screening for prostate cancer (ERSPC) are used. One of the drawbacks of the PSA test is that it has a positive predictive value in the 20% to 30% range, such that most men with PSA levels above the cut point for biopsy do not have prostate cancer.

Figure 2 shows decision curves for various biopsy strategies in men with elevated PSA in the first round of the ERPSC. These strategies are as follows: biopsy all men (thick grey line); biopsy no man (thick black line); biopsy only those men with an abnormal clinical examination (the digital rectal examination [DRE]; thin grey line); biopsy on the basis of a statistical prediction model incorporating PSA level and DRE (dashed line); biopsy on the basis of a statistical prediction model of PSA, DRE, and an additional molecular marker, the ratio of free-to-total PSA (thin black line). Note that the decision curves are shown only for probability thresholds of 10% to 40%. Only these thresholds are shown because we have asked clinicians about what would constitute a reasonable range: A typical response is that few men would opt for biopsy if they were told they had a risk of prostate cancer less than 10%; on the other hand, it is hard to imagine that a man taking a PSA test would want at least a 50:50 chance of cancer before agreeing to biopsy. The decision curve shows that the statistical prediction model including PSA, DRE, and free-to-total PSA ratio has the highest net benefit across the whole 10% to 40% range. We can therefore conclude that using this prediction model, and the new marker, will improve clinical outcome.

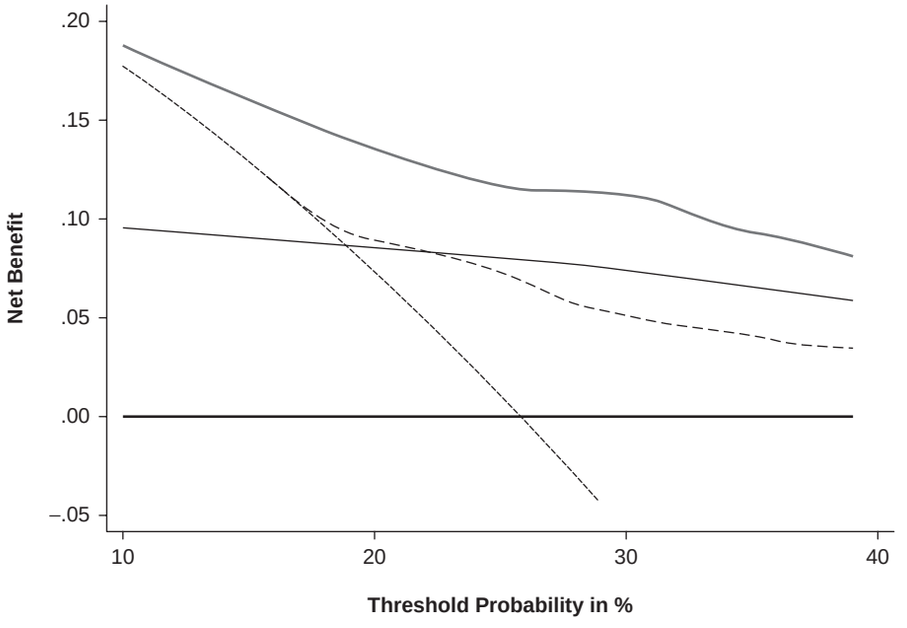


Figure 2 Decision curve analysis for previously unscreened men with elevated PSA

Notes: Biopsy all men (short dashes); biopsy no man (thick black line); biopsy only those men with an abnormal clinical examination (thin grey line); biopsy on the basis of a statistical prediction model incorporating PSA level and DRE (long dashes); biopsy on the basis of a statistical prediction model of PSA, DRE, and free-to-total PSA ratio (thick grey line, top).

It is informative to imagine Figure 2 if it had only shown the decision curve for the PSA and DRE prediction model (the dashed line). The net benefit for this prediction model is only superior to the alternative strategy of biopsying all men for probability thresholds of 20% or higher. We would interpret this as showing that use of the prediction model would help some, but not all men. However, it might also be pointed out that the prediction model only requires data that is routinely collected and is never worse than biopsying all men, so there is little harm in using it, perhaps advising risk-averse men to biopsy irrespective of their risk from the prediction model. If, on the other hand, the prediction model required an additional invasive test or measurement of a novel marker, the decision curve analysis can be described as "equivocal," and it is recommended that a more formal and complex decision analysis be conducted.

Comparison of Decision Curves With Conventional Decision Theory

Figure 2 shows several characteristics of decision curves that are congruent with conventional decision theory. First, the decision curve for "biopsy all men" crosses both the x and the y axis at the prevalence (26%). If the threshold probability is 0 (i.e., $x = 0$), then false positives have 0 weight, and so net benefit becomes the proportion of true positives, which, in the case of biopsying everyone, is the prevalence. For $y = 0$, imagine that a man had a risk threshold of 26% and asked his risk under the "biopsy all" strategy. He would be told that his risk was the prevalence (26%). When a man's risk threshold is the same as his predicted risk, the net benefit of biopsying and not biopsying are the same. Second, the decision curve for the binary test (DRE) crosses that for "biopsy all men" at 1 – negative predictive value, and again, this is easily explained: The negative predictive value is 81%, so a man with a negative test has a probability of disease of 19%; a man with a threshold probability less than this—for example, a man who would opt for biopsy even if his risk was 15%—should therefore be biopsied even if he was DRE negative. Furthermore, although this cannot be seen in Figure 2, the decision curve for DRE is equivalent to "biopsy no one" at the

positive predictive value. This is because for a binary test, a man with a positive test is given a risk at the positive predictive value.

Comparison of Decision Curves With Accuracy Metrics

To illustrate how decision curves and accuracy metrics may diverge, consider the case of a man with elevated PSA after repeat screening. It is reasonable to suppose that different statistical models will be needed for prostate cancer detection, depending on whether a patient has a recent history of screening. Men without recent PSA testing may have an advanced cancer with a high PSA or a localized cancer with a moderately elevated PSA; only the latter is likely for a man undergoing regular screening. Accordingly, both the mean probability of cancer and the relationship between PSA and cancer will differ for previously screened men.

A statistical model for prostate cancer in recently screened men was created using data from rounds 2 to 6 of the ERSPC Göteborg. Differences between prediction models are shown in Table 2. We would expect these different prediction models to have different properties when applied to a data set. Yet when the prediction models are applied to the recently screened men, the predictive accuracies are virtually identical, with AUCs of .6725 and .6732 for the "Round 1" and "Rounds 2 to 6" prediction models, respectively. Figure 3 shows the decision curves for the two prediction models. Although net benefits are close at low threshold probabilities, the "Rounds 2 to 6" prediction model is always superior. An even more extreme case is where we compare a prediction model with just PSA and DRE. The "Round 1" prediction model built on unscreened men has an AUC of .6038 when applied to men with a recent PSA test, again very similar to a prediction model built on this data set (AUC of .6056). However, "Round 1" prediction model has absolutely 0 clinical value with net benefit never higher than those of both "biopsy all" and "biopsy none" (data not shown).

Extensions to Decision Curve Analysis

The formula for net benefit is given in units of true positives but is easily rearranged to give units of false positives.

$$\text{Reduction in False Positives} = \text{Net Benefit} \times \left(\frac{1 - p_t}{p_t}\right)$$

This net benefit can be interpreted as, for example, "Using Prediction Model A instead of Prediction Model B is equivalent to a strategy that reduced the number of biopsies by 10 per 100 patients, without changing the number of cancers found."

Decision curve analysis can also easily incorporate harm, for example, if a test was costly or invasive. The analyst needs to obtain a clinical judgment as follows: "If the test were perfect, how many patients would you submit to the test to find one case?" The reciprocal of this number is the harm and is simply subtracted from the net benefit. For example, if there was an additional test for prostate cancer that was very costly, and clinicians informed us that they would not subject more than 20 patients to the test to find one cancer, the harm of the test would be .05, and the net benefit of any prediction model incorporating the test would be reduced by .05 for all threshold probabilities.

Several other traditional aspects of prediction model evaluation can also be applied to decision curve analysis, including correction for overfit; confidence intervals for net benefit; application to time-to-event data, such as cancer survival; and including competing risks. Simple-to-use R and Stata software for decision curve analysis is available from www.decisioncurveanalysis.org.

Andrew J. Vickers

See also Decision Trees, Construction; Decision Trees, Evaluation; Receiver Operating Characteristic (ROC) Curve; Test-Treatment Threshold

Further Readings

Elkin, E. B., Vickers, A. J., & Kattan, M. W. (2006). Primer: Using decision analysis to improve clinical decision making in urology. *Nature Clinical Practice Urology, 3*(8), 439–448.

Steyerberg, E. W., & Vickers, A. J. (2008). Decision curve analysis: A discussion. *Medical Decision Making, 28*(1), 146–149.

Vickers, A. J., & Elkin, E. B. (2006). Decision curve analysis: A novel method for evaluating prediction models. *Medical Decision Making, 26*(6), 565–574.

Table 2 Differences in prediction comparing men with and without prior screening

	"Round 1 Prediction Model" Created Using Results From Men Without Prior Screening	*"Rounds 2 to 6 Prediction Model" Created Using Results From Men With Prior Screening*
Prevalence of cancer	25.90%	18.90%
Standardized odds ratio from multivariable prediction model		
PSA	1.56	1.19
DRE	4.67	3.34
Free-to-total PSA ratio	0.37	0.58

Note: Change in odds for a 1-standard-deviation increase in the marker.

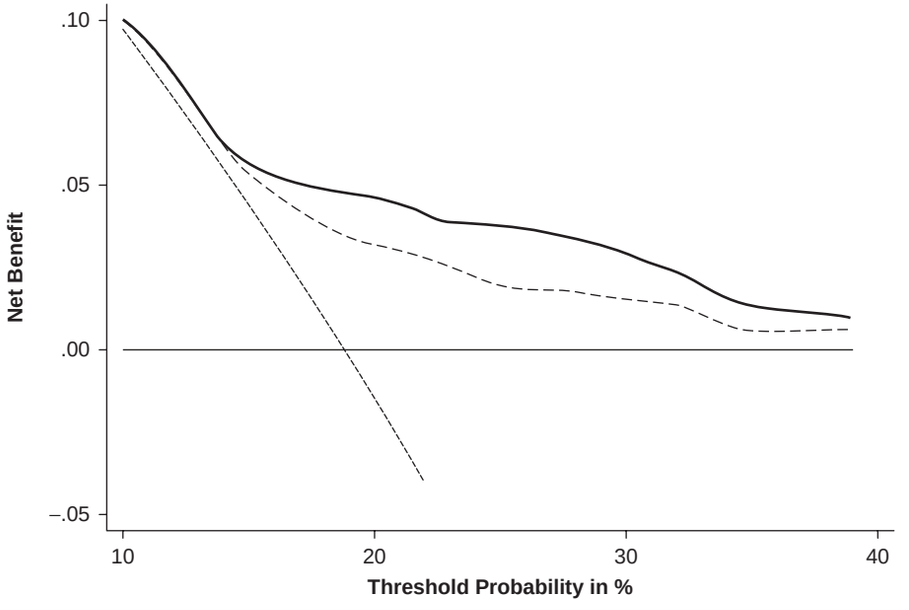


Figure 3 Decision curve analysis for men with elevated PSA on repeat screening

Notes: Biopsy all men (short dashes); biopsy no man (thin black line); biopsy on the basis of a statistical prediction model of PSA, DRE, and free-to-total PSA ratio; prediction model created using data from unscreened men ("Round 1": long dashes); prediction model created using previously screened men ("Rounds 2 to 6": thick line).

Decision Making and Affect

Intuition allows us to make quick decisions in an uncertain environment, not wasting too much time on analyzing possible consequences. Evaluative judgments and decisions are quite often influenced by intuitive feelings rather than analytical conclusions. A doctor in an emergency room, for instance, won't have the time to evaluate the benefits and risks of two similar treatments analytically. The emotion which helps us boost our decision process is called *affect*.

Affect is used as a cue when people define the positive or negative quality of a stimulus; it is experienced as a state and is used whenever quick assignments or attributions are needed to make decisions or judgments. Hence, *affect* is used as an umbrella term referring to states of valence and arousal; it sometimes even includes states of mood, although these are of a more diffuse, low-intensity and long-lasting character. To give an example of experiencing affect, just imagine how fast we associate feelings with words like *cancer* or *emergency*. Thus, some researchers call the reliance on such feelings and their utilization in decision making the *affect heuristic*.

In this entry, a short theoretical background of affective influence in cognition is given, followed by a brief description of psychological models on this topic. Then, various examples according to the affect heuristic and its possible effects in the medical context are examined.

Theoretical Background

Two main attempts can provide a theoretical background for findings on the affect heuristic: First, Epstein's dual-process theory separates "two modes of thinking" into analytical and intuitive, emotional ways of information processing. Secondly, Damasio's theory of "somatic markers" accounts for the importance of affect in decision making.

Epstein's development of the cognitive-experimental self theory introduces a dual process of thinking, assuming two major systems by which people adapt to the world: rational and experimental. Constructs about the self and the world in the rational system refer to beliefs, whereas those in the experimental system refer to implicit beliefs. Neither of the two thinking styles is predominant; they rather function simultaneously. The experimental system is developed through a very long historical evolution and therefore operates more intuitively and automatically. In contrast, the rational system needs more effort to operate; it is mostly used within the medium of language due to its shorter evolutionary history. A wide range of research supports the theory, emphasizing the use of the experimental system in heuristic processing.

Damasio's concept explaining the importance of intuition or affect in decision making was developed by asking the question, "What in the brain allows humans to behave rationally?" His observations led him to the conclusion that human behavior is influenced by "somatic markers" learned in a lifetime. The theory assumes that people mark images with positive or negative feelings, which are directly connected to bodily states. As a result, images can be associated with negative markers that imply an alarming state, or they can be linked to positive markers, meaning a beacon of incentive feeling linked to a bodily state. These assumptions were tested in experiments with patients who had damage to the ventromedial frontal cortices of the brain. Patients with this damage are unable to experience "feelings" and are impaired in their ability to associate affective feelings and anticipated consequences. A gambling game was provided to the participants, where they had to choose cards from any of four card decks. Each chosen card resulted in a gain or loss of a certain amount of money. Patients with the damage to the ventromedial frontal cortices showed their impairment in anticipating future outcomes by their inability to avoid card decks with great outcomes but also great losses. In contrast, "normal" subjects and patients with brain damage outside the prefrontal sections "learned" how to choose the card decks with the lower but continuous payoff. These findings proved that somatic markers increase the accuracy and efficacy of the decision process.

Models of Affective Influence

Psychological models explaining the affective influence on decision making and judgments are often divided by two general categories. One category subsumes associative attempts, when affect is activated in the *semantic memory network* or the *motor network*. Research on semantic memory models analyzes the influence of affective states on the encoding, retrieval, and interpretation of new information. Experiments on affective congruency are derived from this attempt, stating that individuals in a happy mood are more likely to interpret ambiguous information in a positive and more generalized way. For instance, a patient in a good mood might be too positive in describing his or her symptoms, which could complicate the assessment of the right diagnosis. Findings concerning the motor network focus on approaching and avoiding movements

depending on the positive or negative affect, respectively. Therefore, positive valence could be especially useful in stimulating motor action.

A second category refers to inferential models. These are based on the influence inferred by current or anticipated absence or presence of an affective experience. On the one hand, affect can serve *as information*, using a shortcut to decisions and judgment when no alternative explanation is available—as explained in the following examples about the affect heuristic. On the other hand, the influence of affect can occur due to the *intended regulation* or maintenance of an emotional state and therefore lead to accordant decisions. Following this attempt, individuals are not only intending a mood-congruency due to their affective state, but also seek to modulate their mood depending on the contextual needs. Hence, the good mood of the patient could be "adjusted" when telling the risks of a possible disease—and might lead to an adequate description of experienced symptoms.

The Affect Heuristic

A wide range of research is done referring to the affect heuristic, mostly associated with the affect-as-information model—using affective states as useful tools when no other information is available. Findings due to this affect heuristic have various aspects and may interfere with decisions in the medical context. Each of these aspects is described, followed by examples of the effects affect might have on medical decision making.

Preference

Early research already proved that the repetitive presentation of objects leads to positive attitudes and affect toward these objects—independently of any cognitive evaluation. Even more, adding positive or negative meaning to objects guides evaluative judgments, respectively. People therefore are much more susceptible to the affective meaning, albeit any cognitive scrutiny. For instance, the preference of certain drugs and other medical treatments might stem from familiarity or iterated application without taking into account other possibilities. Therefore, medical staff has to be cautious, not wearing blinders or ignoring alternative treatments.

Proportion

Another source of affective influence could be observed by experiments dealing with people's willingness to save a stated number or proportion of lives. Although not rationally comprehensible, the preference of a life-saving intervention is rather evaluated by the proportion than by the numbers of lives that could be rescued. This tendency only changed when two or more interventions could be compared—then the number of lives became more important. Similar findings revealed a study on the support for airport safety. To evaluate benefits of treatments, health professionals are often provided with numbers and statistics—it might be advantageous for them to be aware of the fallacies followed by presented proportions and to always compare different sources of information.

The Evaluation of Risk

A further example in using affect rather than analytical thought concerns the correlation of risks and benefits. Although there is a positive relation in the world, people perceive a negative relation when it comes to everyday decisions: If the benefit is perceived as high, risk is perceived as low and vice versa. Examples can be found in the use of drugs (which are perceived to have a low benefit and a high risk potential) and also medical treatments (e.g., X-rays or antibiotics) that are perceived to have a high benefit and a low risk.

Moreover, despite rational knowledge or evaluation, people often respond rather emotionally in considering dangerous stimuli. For instance, fear can much easier be experienced when people are confronted with dangerous stimuli that evolution has prepared us for (e.g., spiders, snakes, or heights), even when they are cognitively harmless. In contrast, stimuli without an evolutionary history tend to evoke little fear (e.g., guns, smoking)—although they can actually harm us. In the same vein, addictive behaviors tend to be underestimated. Thereby, the strength of a positive or negative affect guides the perception of risks and benefits of an activity.

Numeracy Formats

Quite often, people make a nonoptional choice by "feeling" that this would be the better option. Hence, numeracy is found to have a positive

influence in comprehending probability numbers. In a study analyzing the accuracy in decision making of forensic psychologists and psychiatrists, they were asked to determine whether a patient would commit an act of violence in the following 6 months. As an orientation, clinicians were provided with an assessment of another expert that was either given in terms of relative frequency (e.g., "of every 100 patients similar to Mr. Jones, 10 are estimated to commit violence to others") or statistical probability (e.g., "10% of patients similar to Mr. Jones are estimated to commit violence to others"). Although both probabilities were similar, Mr. Jones was evaluated to be more dangerous when clinicians were informed in terms of relative frequency. Consequently, experts are not resistant against their affective influence on decision making. However, also patients run the risk of misinterpreting information when seeking healthcare decisions (e.g., cancer screening).

Communication of Medical Risk

Risks and benefits of medical treatments are of high relevance for the care seeker. However, as decision options in the medical context are mostly unfamiliar to the patient, "affective cues" could assess meaning to the provided information. In a study analyzing people's ability to perceive the quality of healthcare information, positive and negative affective attributes were included to a presented health plan. Findings showed that participants preferred the health plan more often when positive affective categories were added. Furthermore, the risk of a certain disease is influenced by people's experienced worry rather than actual numbers of deaths from this disease. Therefore, it is important to communicate risks and benefits of illnesses and treatment options to give patients an adequate opportunity to make their right choice.

Stephanie Müller and Rocio Garcia-Retamero

See also Emotion and Choice; Errors in Clinical Reasoning; Mood Effects; Numeracy; Risk Perception

Further Readings

Alhakami, A. S., & Slovic, P. (1994). A psychological study of the inverse relationship between perceived risk and perceived benefit. *Risk Analysis, 14*(6), 1085–1096.

Damasio, A. R. (1994). *Descartes' error: Emotion, reason, and the human brain.* New York: Avon.

Fetherstonhaugh, D., Slovic, P., Johnson, S. M., & Friedrich, J. (1997). Insensitivity to the value of human life: A study of psychophysical numbing. *Journal of Risk and Uncertainty, 14*(3), 282–300.

Loewenstein, G. F., Weber, E. U., Hsee, C. K., & Welch, E. S. (2001). Risk as feelings. *Psychological Bulletin, 127,* 267–286.

Peters, E., Hibbard, J., Slovic, P., & Dieckmann, N. (2007). Numeracy skill and the communication, comprehension, and use of risk-benefit information. *Health Affairs, 26*(3), 741–748.

Peters, E., Lipkus, I., & Diefenbach, M. A. (2006). The functions of affect in health communications and in the construction of health preferences. *Journal of Communication, 56,* S140–S162.

Slovic, P., Finucane, M. L., Peters, E., & MacGregor, D. G. (2004). Risk analysis and risk as feelings: Some thoughts about affect, reason, risk and rationality. *Risk Analysis, 24*(2), 311–322.

Slovic, P., Peters, E., Finucane, M. L., & MacGregor, D. G. (2005). Affect, risk, and decision making. *Health Psychology, 24,* S35–S40.

Winkielmann, P., Knutson, B., Paulus, M., & Trujillo, J. L. (2007). Affective influence on judgments and decisions: Moving towards core mechanisms. *Review of General Psychology, 11*(2), 179–192.

Decision-Making Competence, Aging and Mental Status

The term *competence* in decision making is often linked to the phrase *legal competence*. Here, for example, a judge may go by a patient's bedside to determine if the patient is legally competent to make a medical decision on his or her own behalf. The term *decisional capacity* and notions related to the assessment of decisional capacity belong to the realm of physicians in two areas: (1) assessing the capabilities of patients to make medical decisions in medical care and (2) assessing the capabilities of individuals to make decisions on whether to participate in human research studies as study volunteers. This entry provides an overview of decisional capacity; addresses the role of physicians

in assessing decisional capacity of patients; discusses various assessment procedures for patients, including aging patients and patients with compromised mental states; and closes with a brief look at the implications for future research.

Overview

Decisional capacity is often phrased as being a question of whether a particular individual has the ability to make choices on his or her own behalf. It has been argued that decisional capacity itself has many components, including but not limited to cognition, memory, mood, emotion, and valuation, among others, which can be affected by age and/or mental status. But the above definition overstates the concept of decisional capacity because—even if all components are intact in an individual—in reality very few individuals make decisions solely on their own behalf.

In reality, individuals come to and make choices after considering opinions from others and make decisions of accepting or rejecting alternatives offered to them from a set of options on the basis of the opinions of others in any number of areas: on what to base a decision, on how to choose among a set of alternatives, and on how much to value a benefit in the context of related harms. And all this information is received and processed into a decision where it is virtually impossible to ensure that all intentional or nonintentional attempts to manipulate the information are able to be identified and extracted from the information and the decision. This extraction of manipulated information is essential in any decision that is worth making.

Physician Assessment

Medical Decision Making

In medical decision making—in absence of an emergency where the individual needs to be acted upon medically to save his or her life, with that emergency further characterized as lacking an advance directive developed and signed by this patient at some time before the emergent event—an individual has three options open to him or her when offered a medical opinion. First, the individual can accept the proffered opinion. Second, the individual can reject the proffered opinion. Third, the individual can elect to delay choice until a later time in the hope that something will be developed scientifically (or that more understanding will be gained scientifically) before the medical condition or disease under consideration takes the upper hand in the individual and before that medical condition reaches a state where it can no longer be reasonably eliminated, slowed, or otherwise managed medically no matter what attempts are made to do so.

The phrase *decisional capacity* is used in a much more basic sense in medical decision making than in other arenas of competency of judgment. Issues of rejecting physician-recommended medical interventions may mean that the patient will die from a medical condition or disease process that is otherwise medically considered to be curable, eliminable, eradicable, treatable, or at least manageable by the physician.

Human Subjects Research

Physicians also have to assess the decisional capacity of individuals volunteering their services for research. As noted in the U.S. Code of Federal Regulations, because the goal of human subjects research is to possibly develop scientific knowledge for use in future generations, participation in a research study may not guarantee benefit to the individual study volunteer. Given this, the individual study volunteer may be asked to bear considerable risk of morbidity and mortality in the name of the advancement of scientific knowledge for future generations. To what extent an individual with mental health conditions, such as severe schizophrenia, understands that research is not aimed at helping the individual patient is an active research question.

Assessment Procedures

The question of how decisional capacity is best assessed is also an active research question. Assessment procedures may be unstructured or structured. Structured approaches to decisional capacity may be by a computer-generated tool, by a handheld device, or by a paper-and-pencil assessment. The issues regarding assessment of decisional capacity are not related to not having the instruments to record responses to questions. There is a

wide range of tools to record the answers to assessment questions. Rather, what is missing in the assessment of decisional capacity in medical care and medical research is an answer to the following question: What are the questions that should be asked in the assessment of decisional capacity both in medical care and research on human subjects?

Decisional capacity can be assessed as a changing (evolving) state over time across different choices; on a choice-by-choice basis; or on a hierarchy of choice, where an individual is able to make a choice at a very basic level but yet incapable of choosing at a more complex level where choices have to be made.

Hierarchical Choice

The most basic question in decisional capacity in medical care of patients is the following: Do you want to live or die? For example, if a patient with symptomatic valvular heart disease is asked, "Do you agree to having the doctor perform surgery on your valve?" The patient may respond, "No, I do not want surgery on my heart." The question here is, what is the missing piece in the discussion? If the patient does not understand that the doctor not performing the valvular surgery means that the patient will die, then the patient has not understood the surgery question being posed. Two questions remain: First, has the surgeon given the patient enough information to understand that the present choice, to accept or reject the valve operation, has the consequence the patient will die, or die sooner without the surgery than if he or she had elected to have the surgery and made it through the surgery without dying? Second, does the patient have the decisional capacity to accept or reject the surgical intervention, a surgical operation on his or her heart valve, on his or her own behalf?

Surrogate Decision Making

There is a third underlying question: How would the surgeon respond, regarding his or her operating on a patient, if the patient did not have decisional capacity but the patient's designated surrogate decision maker wanted the surgery to be done on the patient's behalf? This question about surrogate decision making and the response of the surrogate decision maker does not stand alone but again begets a set of questions: What if the patient grimaced each time the surgeon asked the patient directly if he or she wanted to have the operation, would the surgeon still be willing to operate? What if the patient screamed at the surgeon every time the surgeon asked the patient directly whether he or she wanted to have the operation and only screamed at the surgeon when the surgeon broached the issue of the operation, would the surgeon still be willing to operate? What if the patient grimaced, screamed, and attempted to grab and hold onto anything the patient could grab onto each time an attempt was made to place the patient on a gurney to take the patient anywhere outside of the patient's room, would the surgeon still be willing to operate?

This illustrates that there may be definite circumstances in which the surgeon may object to performing a medical procedure on a decisionally impaired patient even if the procedure was necessary to save the patient's life (as in severe valvular heart disease) and even if the patient's designated surrogate agreed to the operation (such as valvular heart repair) on the patient's behalf.

Advance Directives

A final question comes up when a patient begins to lose decisional capacity, recognizes such, and then develops advance directives in clinical care and in research specifying what he or she would be willing to have done in the clinical and research arenas in a variety of circumstances. Here again, simply the placement of a preference in a written and signed advance directive does not necessarily mean that the preference will be carried out or acted on in any way. The carrying out of a decision in an advance directive assumes that those physicians responsible for caring for the patient (or principal investigators and researchers involved in recruitment of patients into studies and their institutional review boards) also agree with what is to be done as specified in the advance directive.

Advance directives that specify nonaction (e.g., do not resuscitate, do not intubate, do not place a feeding tube, do not treat an infection with antibiotics) are more likely to be respected than are certain types of advance directive that may specify action (e.g., do take me to surgery for valvular heart repair should I need it in the future, or do involve me in all invasive research studies in schizophrenia, which

is a disease that I possess, even if I do not have decisional capacity to make the statement of my willingness to participate). One of the problematic characteristics of neurodegenerative disease is the change in personality that can accompany the neurodegenerative process. Here the patient who led a very mild life of careful decision making may become an irascible person quick to anger, and the question can be legitimately asked, is the irascible patient now present in the room with his or her physician or surgeon or with a research principal investigator the "same person" as the patient who signed the advance directive at an earlier time in either a clinical or a research context?

While the phrase *decisional capacity* often connotes the cognitive realm, one of the human mind's key features related to decision making involves not only cognition but memory. Without memory of past and present events, philosophers have argued that there isn't a thread of holding the "same person" together as one unified whole who is to be counted as the person who is the decision maker choosing among sets of options on his or her own behalf.

There is much that is not known about decisional capacity in medicine. For example, depression has in many mental health circles been considered a disorder of mood, yet severe depression is also a disorder affecting cognition and memory, where the severely depressed individual may pay little attention to consideration of any option in his or her care while in the severely depressed state.

In addition to considering issues related to what constitutes the "same person" in the area of advancing neurodegenerative disease, consideration should also be given to patients who face similar issues with other neurologic conditions (e.g., memory problems due to traumatic brain injury) and mental health conditions (e.g., alternating states of severe mania and severe depression). Is the patient in a state of severe mania the same person as the patient in the state of severe depression? Here, the body may be the same but the mental states may be dramatically different.

In addition, memory is no longer viewed in terms of the presence or absence of short- versus long-term memories. Contemporary research on memory includes descriptions of gist versus verbatim memory in normal persons. Individuals volunteering their participation in research studies have helped the acquisition of further scientific delineations of memory including episodic memory; semantic memory; the distinction between implicit and explicit memory; recollection in anterograde and retrograde amnesia; autobiographical memory and autonoetic consciousness; long-term memory following transient global amnesia; the prospect of new learning in amnesia; and the fate of recent and remote memory for autobiographical and public events, people, and spatial locations.

Implications for Future Research

The development of the notion of substitute consent (advance directives and surrogate decision makers) is essential for future scientific research in all medical conditions that break down the person beyond what he or she was in terms of memory and thinking. Yet there is much research to be done in identifying what are the key questions that humans need to be approached with to determine their capacity for decision making at a given time and over time to ensure that they are protected from intrusions that they not only prefer not to have, but that they outright object to as humans.

Dennis J. Mazur

See also Decisions Faced by Institutional Review Boards; Informed Consent

Further Readings

Bravo, G., Duguet, A. M., Dubois, M. F., Delpierre, C., & Vellas, B. (2008). Substitute consent for research involving the elderly: A comparison between Quebec and France. *Journal of Cross-Cultural Gerontology, 23*(3), 239–253.

Dunn, L. B., Nowrangi, M. A., Palmer, B. W., Jeste D. V., & Saks, E. R. (2006). Assessing decisional capacity for clinical research or treatment: A review of instruments [review of the current status of decisional capacity assessment tools]. *American Journal of Psychiatry, 163,* 1323–1334.

Dunn, L. B., Palmer, B. W., Appelbaum, P. S., Saks, E. R., Aarons, G. A., & Jeste, D. V. (2007). Prevalence and correlates of adequate performance on a measure of abilities related to decisional capacity: Differences among three standards for the MacCAT-CR in patients with schizophrenia. *Schizophrenia Research, 89,* 110–118.

Guillery-Girard, B., Quinette, P., Desgranges, B., Piolino, P., Viader, F., de la Sayette, V., et al. (2006). Long-term memory following transient global amnesia: An investigation of episodic and semantic memory. *Acta Neurologica Scandinavica, 114,* 329–333.

Jefferson, A. L., Lambe, S., Moser, D. J., Byerly, L. K., Ozonoff, A., & Karlawish, J. H. (2008). Decisional capacity for research participation in individuals with mild cognitive impairment. *Journal of American Geriatric Society, 56*(7), 1236–1243.

Piolino, P., Desgranges, B., Belliard, S., Matuszewski, V., Lalevée, C., de la Sayette, V., et al. (2003). Autobiographical memory and autonoetic consciousness: Triple dissociation in neurodegenerative diseases. *Brain, 126*(Pt. 10), 2203–2219.

Reyna, V. F., & Hamilton, A. J. (2001). The importance of memory in informed consent for surgical risk [gist vs. verbatim memory]. *Medical Decision Making, 21,* 152–155.

Rosenbaum, R. S., Köhler, S., Schacter, D. L., Moscovitch, M., Westmacott, R., Black, S. E., et al. (2005). The case of K.C.: Contributions of a memory-impaired person to memory theory. *Neuropsychologia, 43,* 989–1021.

Saks, E. R., Dunn, L. B., Wimer, J., Gonzales, M., & Kim, S. (2008). Proxy consent to research: The legal landscape. *Yale Journal of Health Policy, Law, and Ethics, 8,* 37–92.

U.S. Code of Federal Regulations. (2007). 45.46.102.d.

Decision Making in Advanced Disease

Decision making in advanced disease is complex and challenging. Decisions are emotional and often have irreversible outcomes (e.g., death). For many, the desire to live longer is strong, but unrealistic, when faced with advanced illness, and goals must be realigned toward comfort and quality of life. Some medical interventions are invasive and detract from quality of life without lengthening life. Decision makers include professionals, patients, their families, and external parties (institutions, insurers, governments). These decision makers may have diverse goals, priorities, values, and cultural backgrounds, affecting their beliefs about care near the end of life. Prognostication and communication are critical to good decision making. In advanced disease, preferences for decision-making style are individual, often unstated, and change over the illness course. Advance directives involve decisions about theoretical future events. Near the end of life, many people lose capacity to make decisions and this responsibility falls to their families. This is often at a time of great stress and influenced by emotions, grieving, and caregiving burdens. Sometimes a time-limited trial of therapy is used to facilitate decision making in these difficult situations.

Why Is Decision Making Needed in Advanced Disease?

Patients with advanced disease are faced with complex treatment options (disease-focused or supportive therapy, hospice, clinical trials) and choices about commencement, continuation, or withdrawal of interventions such as artificial hydration and nutrition, blood transfusion, cardiopulmonary resuscitation, circulatory support, dialysis, and invasive ventilation.

Studies of quality of death in America have found that death frequently occurs in hospitals and is accompanied by the use of highly technical interventions (e.g., invasive ventilation, cardiopulmonary resuscitation) and significant pain and distress. Invasive medical interventions close to death are not associated with better outcomes and are sometimes against the expressed wishes of patients. Trials of interventions to improve quality of care at the end of life (the SUPPORT study) have so far been unsuccessful.

When Are Decisions Needed?

Decision making in advanced disease requires recognition (usually by the clinician) that a decision needs to be made. Even not making a decision may be a decision itself. Timing of the decision requires recognition and communication of the following: incurable disease, limited prognosis, potential future-course and alternative-management options. Decision making may be impaired by the assumption that only one option is available (e.g., active treatment is pursued due to failure to recognize supportive care as a valid treatment option).

In many advanced illnesses, especially neurological illnesses, ability to communicate is lost as

disease progresses. Decision-making capacity may also be lost due to an acute crisis requiring intubation or sedation or to delirium, which commonly occurs close to death. Ideally, patients with advanced illness are able to express their treatment preferences, write advance directives, and appoint a surrogate decision maker (medical power of attorney) before they lose capacity.

Prognostication

Decision making in advanced disease relies on prediction of prognosis: the expected duration and quality of life, and likely future course of the disease. Advanced cancer is often characterized by a short decline in function toward death. Advanced nonmalignant diseases (chronic organ failure) have a more gradual decline worsened by recurrent exacerbations. Death is the result of an acute exacerbation that fails to respond to treatment, thus timing is less predictable. Chronic frailty or dementia follows a slow, drawn-out decline. Instruments are available to predict prognosis based on type and stage of disease, symptoms, physical function (performance status), and test results. These instruments predict chances of being alive at a certain point or give a median survival for a similar group of patients; they cannot predict how long an individual will live. Physician predictions of individual prognosis are often inaccurate and tend to be overly optimistic. Physicians are reluctant both to make prognostic estimates and to communicate them to patients. Patients' estimates of their own prognosis are also often inaccurate. These failures of prognostication and communication hinder decision making.

Communication

Physicians may be reluctant to initiate discussions about end-of-life care for fear of removing hope. However, patients often do prefer to receive prognostic information, and denying them this knowledge may impair preparation for death. Patients are willing to discuss preferences but rarely initiate these conversations; thus clinicians need to be proactive. Communication goals include eliciting preferences (for information, decision making, and treatment), understanding values and beliefs, and establishing goals of care—priorities (for quality or quantity of life), hopes, and legacies. Patients may have specific wishes to fulfill, events to live for, and preparations to make (financial, practical, or legal). Clinicians also must provide information about diagnosis, prognosis, and treatment options. A majority of patients in English-speaking countries want detailed information, while patients in other countries may prefer less information. When presenting treatment options, clinicians have an obligation to be realistic. Rather than present a laundry list of all possible treatments, only options that are feasible given the circumstances should be discussed.

Decision-Making Styles

The prevailing attitude in Western medicine is respect for individual autonomy, and thus shared or autonomous decision making is preferred. However, patients express a range of preferences, with between 30% and 60% preferring shared decision making. Age, gender, and ethnicity may influence preferences, but inconsistently; thus, individual preferences need to be elicited. Preferences may alter with each decision and with disease course; patients closer to death are more likely to delegate responsibility to their physician. Decision style also varies with the magnitude of the decision and the certainty of the outcome. For example, a decision about which antibiotic to prescribe for pneumonia is usually made by a physician based on established medical knowledge. These unilateral decisions are usually communicated to the patient, who then may choose to accept or reject the recommendation.

While Western culture highly values autonomy, other cultures value family decision-making styles. Such families may request that information regarding diagnosis, prognosis, and treatment be withheld from the patient. This can cause conflict with a clinical team focused on individual autonomy; however, autonomy includes the right to defer decision making to one's family. These issues can be addressed by eliciting cultural beliefs about truth telling and decision making of patients and their families.

Surrogate decision making is required if patients lose decision-making capacity. It is most often performed by a close family member. Ideally surrogate decisions are based on substituted judgment (what the patient would want in this circumstance) and

best interest (what is thought to be best for the patient at this time). Substituted judgment is best derived from previous conversations or statements; however, patients infrequently express their wishes to family members. Families are often inaccurate in predicting patients' treatment preferences. They also consider factors such as quality of life, emotions, and their own values when making decisions. Caregiver anxiety or depression may also influence surrogate decisions. Surrogate decision making can be burdensome for families if they are asked to make decisions without information, assistance, and recommendations from clinicians. They may feel guilt or responsibility for their relative's death. Surrogate decision making can also be a source of conflict if clinicians consider further aggressive treatment futile and families insist that it continue.

Perspectives of Decision Makers in Advanced Disease

Clinicians

Models of care for advanced disease promote multidisciplinary teams; thus, multiple clinicians may be involved in decision making. These clinicians may have diverse backgrounds and training (e.g., physicians, nurses, social workers, psychologists, and chaplains) and thus diverse views on care in advanced disease.

Treatment recommendations may be influenced by specialist training and practice location. Clinicians untrained in principles of palliative care may not feel confident in offering this option. Conventional medical teaching (e.g., antibiotics for pneumonia) may not always be the best option for a person close to death. Clinicians have a professional responsibility to provide recommendations rather than abdicating all decisional responsibility to the patient. Recommendations should be based on both medical knowledge and the priorities and values of the patient and family.

Clinicians are also influenced by real and perceived ethical dilemmas. Consensus supports the ethical nature of treatment withdrawal and withholding artificial nutrition and hydration in terminal illness, surrogate decision making, and the principle of double effect (unintentional hastening of death with treatment aimed at comfort). Physician-assisted suicide and euthanasia are illegal in most countries and American states, exceptions being the Netherlands, Switzerland, Washington, and Oregon.

Physicians sometimes use futility to facilitate treatment decisions in advanced disease. This principle holds that physicians are not obliged to provide treatment considered futile. The definition of futility is controversial and lacks consensus. Futility definitions may be quantitative—the treatment won't work (e.g., cardiopulmonary resuscitation in advanced disease), or they may be qualitative—treatment will only prolong a state of poor quality of life (e.g., persistent vegetative state). Definitions of futility are subject to value judgments; thus, a process of communication and negotiation is recommended when futility issues arise.

Patients

Patients faced with life-threatening illness are more willing to accept aggressive and toxic treatment, with minimal chance of benefit, than their clinicians and healthy people. Decision making is influenced by values and priorities (for quantity or quality of life), past experiences, family, friends, and presence of children. Patients' perception of their prognosis (which is often inaccurate) influences their treatment choices. Access to and availability of services may influence treatment decisions (e.g., geographic access to radiotherapy is often limited). Patient priorities near the end of life may include pain and symptom management, sense of control, avoiding prolonged dying, relieving families' burden, strengthening relationships, and preparation, including financial and funeral arrangements. Concerns may include treatment toxicity, burden (appointments, tests, side effects), and financial costs.

Family

Families of people with advanced illnesses may be hoping for cure or prolongation of life while also experiencing anticipatory grief. Caregiving is often characterized by loss of employment and financial security and stresses of maintaining family function and their own health. Information needs of families may be different from those of patients, especially as disease progresses. Families may not be in close proximity and thus may be faced with difficult decisions of timing travel to be with their

family member. Surrogate decision making places additional burden on families, and preexisting conflicts are likely to be escalated by decision-making responsibilities. Family conferences are often used in advanced disease, especially in the intensive and palliative care units. These meetings usually involve at least two clinicians (often physician and social worker) and all family members (including friends or other caregivers) relevant to the patient.

External

External organizational, cultural, and political factors influence decision making in advanced disease. These factors may not be evident on an individual level but influence the experiences of groups. For example, the number of regional hospital beds is a stronger predictor of place of death than patient preference. Availability of hospice services reduces hospital deaths.

Health insurers and reimbursement options also influence decisions. In the United States, the Medicare Hospice Benefit is available to patients with an estimated prognosis of less than 6 months. Because of difficulties in prognosticating for nonmalignant diseases, these patients are underserved by hospice. Patients often need to forgo disease-modifying treatments to be eligible for hospice. This condition causes some people to delay hospice until terminal stages of illness.

Decision-Making Processes

Six Thinking Hats

This model was developed by Edward de Bono to promote parallel thinking in group decisions. Decision makers consider issues from one perspective simultaneously and then move on to the next. Principles of this strategy can also be applied to advanced disease.

Information (White Hat)

Information needed includes prognosis, options available, and likely outcomes of each option. Often, information gathering and provision is the role of the clinician. Different styles of presenting and framing information influence patients' decisions. Patients and their families are increasingly accessing Internet sources, which may lead to misinformation. Cancer Web sites frequently discuss treatment options and side effects but rarely prognosis. In advanced disease there may be limited evidence, thus uncertainty and probabilities play a large role. A treatment-response rate may be small, but who responds or experiences side effects is largely unpredictable.

Emotion and Intuition (Red Hat)

Patients' and families' emotions may include denial, hope, anger, or a sense of abandonment. Patients may be concerned about being a burden to others, loss of control, and dignity. Both patients and clinicians are influenced by spiritual, religious, and cultural beliefs about death. Patients and families may also have emotional reactions to the decision-making process itself, for example, feelings of anger and resentment toward the process or clinical team.

Caution (Black Hat) and Optimism (Yellow Hat)

Consequences of each option (positive and negative) need to be considered. While active treatments may extend life, supportive therapy also has positives of symptom control and quality of life. Costs may include treatment burden, side effects, caregiving burden, and financial costs. Patients and their families may hold false hopes for prolongation of life or cure. Hope may need to be redirected toward comfort and quality of life.

Creativity (Green Hat)

Creative solutions in advanced disease may include flexibility of decisions (e.g., pursue Plan A with Plan B if unsuccessful), or two options simultaneously (supportive care and active treatment). Second opinions and advice from colleagues may also suggest creative options.

Process Control (Blue Hat)

The clinician's role is to summarize, conclude, and make plans for follow-up. Retention of information in times of stress is poor, and questions are often thought of after conversations. Time may be needed to consider options and make a decision.

Decision Aids

Question prompt lists are available for advanced cancer or those seeing a palliative care team to help

patients gather information. Decision aids are available for early-stage cancer, but few are available for advanced disease. Guidelines, care plans, and hospital policies can be used to facilitate decision making, but few have been published.

Practical Decision Making: Issues to Consider

In clinical practice, certain practical considerations may facilitate treatment decisions. These include the following:

- What is the performance status and extent of disease? These may make aggressive treatment unrealistic.
- Is the condition reversible or treatable?
- What are the possible complications or worst outcome, and are these acceptable to the patient?
- Does treatment contribute to patient comfort and/or safety?
- Is it logical, appropriate, and humane?
- Does it make good medical and common sense?
- What are the costs?
- What do the patient and family want?

Conflict

Conflict may arise between patient and family and clinician (e.g., wanting to continue treatment that clinician considers futile), between members of the treating team, or between various family members. Conflict may be avoided by clear communication about prognosis, expected outcomes, and goals of care. Communication can be facilitated by family or team meetings. Approaches to conflict resolution include ethics or palliative care consultation or independent mediation. If conflict cannot be resolved, strategies involve a time-limited treatment trial (with explicit outcome measures), or transfer of care to another clinician. Unfortunately, some conflicts have ended in legal and public disputes.

Katherine Hauser and Declan Walsh

See also Advance Directives and End-of-Life Decision Making; Decision-Making Competence, Aging and Mental Status; Decisions Faced by Surrogates or Proxies for the Patient; Physician Estimates of Prognosis

Further Readings

Back, A. L., & Arnold, R. M. (2005). Dealing with conflict in caring for the seriously ill. "It was just out of the question." *Journal of the American Medical Association, 293,* 1374–1381.

Council on Ethical and Judicial Affairs, American Medical Association. (1992). Decisions near the end of life. *Journal of the American Medical Association, 267*(16), 2229–2233.

Council on Ethical and Judicial Affairs, American Medical Association. (1999). Medical futility in end-of-life care: Report of the Council on Ethical and Judicial Affairs. *Journal of the American Medical Association, 281*(10), 937–941.

Covinsky, K. E., Fuller, J. D., Yaffe, K., Johnston, C. B., Hamel, M. B., Lynn, J., et al. (2000). Communication and decision-making in seriously ill patients: Findings of the SUPPORT project (The Study to Understand Prognoses and Preferences for Outcomes and Risks of Treatments). *Journal of the American Geriatrics Society, 48*(Suppl. 5), S187–S193.

de Bono, E. (1999). *Six thinking hats*. Boston: Back Bay Books.

Matsuyama, R., Reddy, S., & Smith, T. J. (2006). Why do patients choose chemotherapy near the end of life? A review of the perspective of those facing death from cancer. *Journal of Clinical Oncology, 24*(21), 3490–3496.

Meisel, A., Snyder, L., Quill, T., & American College of Physicians-American Society of Internal Medicine End-of-Life Care Consensus Panel. (2000). Seven legal barriers to end-of-life care: Myths, realities, and grains of truth. *Journal of the American Medical Association, 284*(19), 2495–2501.

Parker, S. M., Clayton, J. M., Hancock, K., Walder, S., Butow, P. N., Carrick, S., et al. (2007). A systematic review of prognostic/end-of-life communication with adults in the advanced stages of a life-limiting illness: Patient/caregiver preferences for the content, style, and timing of information. *Journal of Pain and Symptom Management, 34*(1), 81–93.

Quill, T. E., & Brody, H. (1996). Physician recommendations and patient autonomy: Finding a balance between physician power and patient choice. *Annals of Internal Medicine, 125,* 763–769.

Stagno, S. J., Zhukovsky, D. S., & Walsh, D. (2000). Bioethics: Communication and decision making in advanced disease. *Seminars in Oncology, 27,* 94–100.

Weissman, D. E. (2004). Decision making at a time of crisis near the end of life. *Journal of the American Medical Association, 292*(14), 1738–1743.

Decision Modes

A decision is a commitment to a course of action that is intended to serve the interests and values of particular people, which often differ sharply from one person to the next. A good example is a patient's choice of radical mastectomy over lumpectomy as a treatment for breast cancer, where the patient seeks to do what is best for both herself and her family, especially her young children. There is considerable variability in not only *what* different people (and even the same person on different occasions) decide when facing the same dilemmas, but also in *how* they decide. The term *decision modes* is used to characterize such qualitatively distinct means by which people reach their decisions. This entry describes and reviews several of the major decision modes that have been acknowledged. It also discusses their conceptual and practical significance, particularly in medicine.

A Big Picture

There are myriad decision modes. But almost all of them can be classified into a small number of categories defined according to several metadecisions that are made, consciously or otherwise, in virtually every decision situation. Here the expression *metadecision* refers to a decision about how to decide. The decision mode tree in Figure 1 provides a big-picture view of the decision modes that result from these metadecisions. The discussion proceeds from the "Responsibility" node near the top of the tree down to the bottom.

Responsibility

In every decision situation, someone—either an individual person or a collective—must assume responsibility for making the decision in question. Thus, for example, in the contemporary United States, it is understood that the patient herself has the responsibility—or "right," "privilege," "authority," "obligation," "burden," even "duty"—for deciding how her breast cancer will be treated. Usually, on a local basis, at least, assumptions about decision-making responsibility are so broadly accepted, so "natural," that the issue never crosses people's minds. Discussions of responsibility do not occur except under extraordinary circumstances, such as when the assumptions are contested. Only then do people realize that responsibility typically has been established via earlier metadecisions made by others, including society, as suggested by the "Prior metadecisions" node in the decision mode tree. For instance, many Americans are first spurred to think about responsibility for cancer treatment decisions when they learn that Japanese responsibility customs are different from their own. They are surprised to learn that in some long-standing Japanese traditions, a cancer patient might not even be told by her physician and her family that she has the disease. Or take the case of end-of-life decisions. When the patient is incapacitated, as in the Terri Schiavo case in Florida, which ended on Schiavo's death in 2005, who has the right to decide—the patient's spouse, the patient's parents, the state legislature, Congress, or the courts? Many people had never pondered such knotty questions until media coverage of the Schiavo case forced them to do so.

Digression: Adequacy of Mode Metadecisions

Part of the full scientific story of human decision behavior is an understanding of how and why people make the mode metadecisions that they do. But there is a practical side, too. Suppose that, at some metadecision choice point in the mode tree, the decider goes down one path rather than some other. Furthermore, suppose that this increases the odds that the eventual decision will be effective. Then it is legitimate to say that that metadecision is better than it would have been otherwise. The following discussion briefly addresses adequacy concerns as well as questions about how particular metadecisions are reached.

Choice Point ❶: Reauthorization

The first metadecision facing the responsible or recognized decider—one person or several—is about whether and how to shift at least some of that responsibility to others, authorizing them to take part in the decision process. At one extreme, the recognized decider might do nothing, *retaining* full responsibility. For instance, a heart disease patient might declare, "Whether I receive

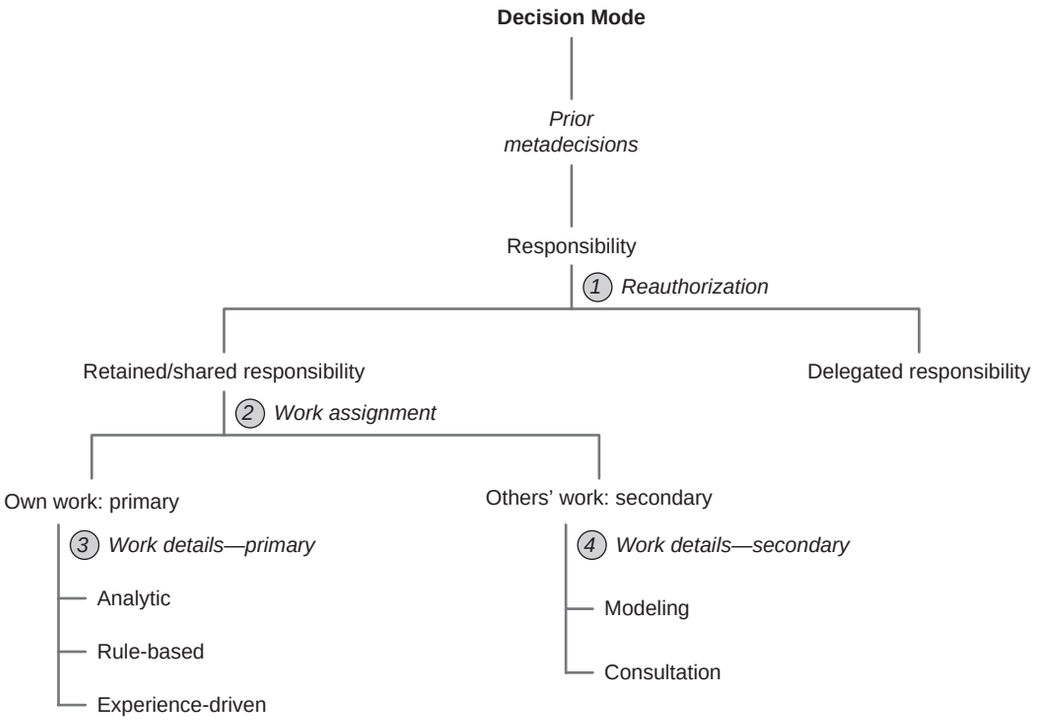


Figure 1 Decision mode tree

angioplasty is my decision and mine alone." At the opposite extreme, responsibility is *delegated* or relinquished entirely, a circumstance sometimes described as "agency," as when a patient says to his doctor, "I'm no expert, but you are, so would you please decide for me, as if you were choosing for your own father?" Between those extremes, the recognized decider might elect to bring others into the picture to *share* decision-making responsibility. That is, the decision process becomes a (more) collective one, as in the shared medical decision-making paradigm, where the patient and the physician assume joint responsibility for choosing medical treatments.

There are several reasons that recognized deciders sometimes favor either retaining or relinquishing some or all of the responsibility they inherit. Especially significant among those reasons are prevailing customs or even laws. A good illustration is provided by socially (and sometimes legally) sanctioned personal autonomy principles that encourage patients to lean toward making their own treatment decisions. Other motivations as well as consequences of authorization choices are sketched presently.

Retaining Responsibility

Two aims are common for those choosing to retain full decision-making responsibility. The first is preserving the perquisites of decision-making authority. People who have such authority cannot help using it to serve their own, personal interests, it seems. Thus, relinquishing that authority, or even sharing it, poses a significant risk. This is part of the rationale for patient autonomy principles such as those embodied in informed consent requirements: Empowered patients will not knowingly choose treatments that harm them. A second common goal in holding onto decision-making responsibility consists of short-term time constraints. If a decision needs to be made in a hurry (e.g., triage in an emergency room), all else being the same, the fewer people there are who must reach agreement, the better things are.

Cast against the sought-after aims of retaining extant decision-making authority are several threats to decision effectiveness. There is evidence that people often overestimate their own skills. Such overconfidence would induce deciders to believe that they can perform essential decision-making

tasks better than others whose true expertise is actually superior. Such overconfidence is not universal, however. It appears weakly if at all when there is unambiguous evidence of one's incompetence. Thus, although a trained physician might overestimate his ability to make decisions better than those of his peers, similar overestimation should be less common among naïve patients. A second major drawback of retaining decision-making authority is related to the first—excessive workload. Suppose that a physician refuses to delegate decision tasks because she incorrectly believes that no one else can make those decisions as well as she can. Then, in short order, she is likely to be so overwhelmed that all her decisions suffer and so does her personal well-being.

Sharing Responsibility

The most compelling motivation for a decider choosing to share decision-making responsibility is embodied in the adage, "Two (or more) heads are better than one." That is, sharing decision-making responsibilities seems to promise better decisions because of the energy and also the specialized knowledge that other people bring to the table. "Political" benefits beckon, too. The people brought into the decision process are more likely to accept instead of resist the resulting decisions, since people rarely protest against themselves. Similarly, sharing partly shields the originally recognized decider from the blame that often spews forth when decisions turn out badly (e.g., the wrath of a family whose loved one dies in a surgical procedure chosen solely by a physician).

The anticipated rewards of sharing decision-making responsibility sometimes go unrealized or are overshadowed by the costs of sharing. One threat to the effectiveness of sharing is free-riding, the tendency for members of a group to do less than their fair share of the work, partly because they expect that others will pick up the slack for them. Another is the documented phenomenon whereby information that is possessed by every participant in a meeting tends to be overly represented among the topics actually discussed. This means that knowledge possessed uniquely by individual discussants is neglected. This defeats a primary aim for broadening participation in the decision-making process in the first place, the exploitation of specialized expertise (e.g., the unique insights that an endocrinologist, an oncologist, and a gynecologist can bring to a case conference). And then, of course, there are the increased coordination costs demanded by the sharing of decision-making responsibility (e.g., the hassles of finding mutually suitable meeting times for all participants in the decision process, to say nothing of the time spent in the meetings themselves).

Delegating Responsibility

The advantages envisioned for delegating decision-making responsibility to others are partly the same as those for sharing (e.g., taking advantage of specialized knowledge). But the prospects of lower costs are especially alluring. After all, the originally recognized decider is freed entirely (although typically for a fee, broadly defined) from having to work through the decision problem in question; that problem would belong to someone else altogether.

Yet delegation carries with it burdens and risks that are easy to overlook. First off, to delegate properly, a recognized decider must understand decision processes as well as the current decision problem in sufficient detail to know what kinds of expertise are required to solve that problem effectively. Consider, for example, the challenge of determining whether the training of a physician's assistant is sufficient to allow her to decide whether to send home patients who do not need further attention. Furthermore, the decider must know how to appraise others' expertise (e.g., "Is this *particular* assistant up to the task?"). Ample research indicates that our ability to evaluate expertise is less than ideal, being vulnerable to numerous potentially misleading indicators, such as candidates' skills at mimicking the speaking style of recognized authorities. An especially important challenge is assuring *incentive alignment*. This means that those to whom decision-making authority is delegated would gain no benefit from making decisions that are contrary to the interests and values of the people the decisions are supposed to serve. That is, they have no conflicts of interest. Incentive alignment is at the heart of controversies about physicians' dual responsibilities to patients and insurers.

Choice Point ❷: Work Assignment

Once responsibility for making a decision is settled, the actual work of reaching that decision must be carried out. There are two alternatives for who performs particular aspects of that work—the recognized deciders or someone else. When those deciders execute the required tasks themselves, the modes are referred to as *primary*; otherwise they are *secondary*. Decision making typically encompasses a wide variety of chores. Therefore, work assignment for a given decision problem can easily involve both primary and secondary modes for different elements of the overall effort. The problem of determining whether a decision task should be assigned to a secondary mode is largely the same as that of determining whether to delegate an entire decision problem to someone else. Thus, the same principles apply.

Choice Point ❸: Work Details—Primary

As indicated in the decision mode tree, there are three major classes of primary modes: analytic, rule-based, and experience-driven.

Analytic Decision Making

The essential, distinguishing feature of *analytic* decision making is that the decider reasons through what makes sense as a solution to the decision problem at hand, with no constraints on the inference process. When most people hear and use the term *decision making,* this is what they have in mind. There are two principal variants of analytic decision making, substantive and formal.

In *substantive* analytic decision making, the decider reasons according to a conception of how nature (broadly conceived) works, that is, how one event or action leads to another, which in turn yields other occurrences, and so on. Effectively, the decider relies heavily on mental simulations of the chains of events that plausibly might ensue if various alternative actions were chosen. Then the decider pursues the option whose simulation turns out best in the decider's eyes. Consider, for example, how a physician might reason through the sequences of potential biological consequences if she were to recommend alternative drug therapies for a patient experiencing both hypertension and diabetes. If the sequence for one particular therapy includes a highly probable severe drug interaction, the physician backs away from that course of action.

The defining characteristic of *formal* analytic decision making is that significant elements of the decider's reasoning entail operations on symbolic representations of key elements of the decision situation. These operations might be carried out in the decider's head or perhaps via a computer. As an example, consider a decision analysis in which a kidney patient's utilities for various health states are, via expected utility formulas, aggregated with probability assessments for potential outcomes, to yield treatment recommendations.

Rule-Based Decision Making

Rule-based decision making relies on decision rules of this form: *If Conditions C1, C2, C3, . . . hold, then pursue Action A.* Sometimes deciders develop such rules on their own, summarizing personal observations and arguments (e.g., when a physician says, "Over the years, I have noticed that . . ."). But some rules are provided by experts, as in the case of the National Comprehensive Cancer Network's practice guidelines for treating osteosarcoma. Rule-based decision making is not as simple as it might seem. For instance, it requires a prior decision about whether to accept a particular decision rule, say, on the basis of its developers' reputations. And applying that rule often demands a tough judgment as to whether the current situation matches the rule's preconditions sufficiently closely.

Experience-Driven Decision Making

The word *experience* is used in two distinct but related senses in the expression *experience-driven decision making*. The first sense implicates decision making that is nearly the antithesis of analytic decision making in that it does not entail breaking decision problems down into their components, such as utilities versus probabilities. Instead, the decider has an undifferentiated psychological experience that somehow pushes the decider toward one potential action rather than its competitors. Furthermore, the decider typically cannot explain the decision process and it may well be nonconscious. Instead, those asserting a reliance on such nondeliberative, "intuitive," "recognition-primed,"

or "System 1" decision making often say things such as "For some reason, it just felt like the right thing to do." One prominent line of scholarship on such decision making is commonly identified with the somatic marker hypothesis for risk taking. According to this theory, over repeated experiences, people gradually develop biologically mediated associations with high-risk alternatives, associations that compel them to shy away from those alternatives even before they can offer reasons for why they feel the way they do. Significantly, individuals with damage to the medial prefrontal cortex, who are notorious for poor decision making, do not develop these risk-repelling associations.

The second sense of *experience* in some kinds of experience-driven decision making refers more directly to the decider's past cognitive activities. The core idea is that, as the decider repeatedly encounters—experiences—a particular situation, the decider learns, in the broad sense of the term. Imagine a teenager who chooses to light up a cigarette for the first time in a certain type of situation, as in the presence of particular friends. If this scenario is repeated over and over, eventually the teenager no longer deliberately and reflectively "chooses" to smoke. Nevertheless, he routinely finds himself smoking in that scenario. What was once an analytic decision has evolved into an automatic, experience-based one. More generally, automatic decision making is such that *If Conditions C1, C2, C3, . . . present themselves, then the decider* will *pursue Action A*. Furthermore, the process has the usual characteristics of automaticity: The decider has no control over the process, the process is virtually effortless, and the decider has minimal awareness of it.

Considerations

In a given situation, the primary modes are likely to be attempted in this order: experience-driven, rule-based, analytic. By its nature, experience-driven decision making, particularly the automatic, habitual variety, just pops out when the given routine has been established and when the triggering conditions are encountered. Otherwise, the decider has no choice but to seek an applicable decision rule or, if that fails, make the decision analytically. The latter is a last resort since it is so labor-intensive. But there may very well be no choice since, although decision rules are indeed common, they do not fit or exist for every situation (e.g., not every patient situation matches an available practice guideline).

It is important to recognize that in a given decision episode, more than one primary mode might be invoked. However, because they "run" so rapidly and effortlessly, if they are available, experience-driven modes are likely to exert inordinate influence as compared with more deliberative analytic and rule-based modes. This can be worrisome since the arbitrariness of the events giving rise to experience-driven modes (e.g., chance peer encounters that nurture smoking habits) provides no assurance that these modes yield effective decisions. Similar considerations apply to the problem of improving decision-making practices. Clearly, approaches that work for reshaping analytic decision-making practices would be useless for experience-driven ones.

Choice Point ❹: Work Details—Secondary

For the most part, secondary decision modes can be viewed as tools for assisting analytic decision making. That is, in the process of reaching a decision analytically, the decider draws on the efforts of other people (or devices) as special resources.

Modeling

The *modeling* mode is deceptively simple. The decider identifies another decider who has faced the same dilemma and just mimics that person's decision, allowing that model to do all the work of thinking through what is reasonable to do. Although seldom discussed, modeling occurs often, as when a patient chooses as his own physician the same one he learns was selected by the boss he admires at work. More generally, modeling is the mode implicit in herding behavior, which is observed among both lower animals in stampedes and humans in financial markets. Modeling is unquestionably easy, but it is beset by significant risks, too. Simply observing the model's decision without also learning whether it was effective for the model is one risk. Another is assuming that the model's interests and values are identical to the decider's own, an assumption that is often highly suspect, as when choosing doctors.

Consultation

In the *consultation* mode, the decider acquires advice about the decision problem from either a real person or a device, such as a computer program—advice that the decider is free to accept or reject. The advice might be a "bottom line" recommendation as to the action the decider should pursue. Such is the case when a patient asks for a second opinion: "Should I have the proposed surgery or shouldn't I?" (Note that the consultant might have arrived at a recommendation via any of the primary decision modes distinguished previously, e.g., analytic, rule-based, or experience-driven.) Alternatively, the advice could pertain to some specific element of the decision problem, as when a patient asks, "What are the available treatments for my condition, and what can go wrong with each of them?"

In principle, consultation seems almost perfect as a complement to analytic decision making. After all, it allows for the application of specialized expertise to every critical aspect of the decision problem. But therein lies perhaps the greatest hazard: assessing such expertise. Ideally, deciders' conclusions about the expertise of their potential consultants should be based on the track records of the candidates; that is, they should be evidence based. Studies have shown, however, that conclusions are strongly affected by factors that easily can have nothing to do with track records (e.g., an authoritative manner). They have also demonstrated that, left to their own devices, people often fail to seek such records and are confused about how to best use them when they are available.

J. Frank Yates

See also Automatic Thinking; Intuition Versus Analysis; Judgment Modes; Shared Decision Making

Further Readings

Bechara, A., Damasio, H., Tranel, D., & Damasio, A. R. (1997). Deciding advantageously before knowing the advantageous strategy. *Science, 275,* 1293–1295.

Christensen, C., Larson, J. R., Jr., Abbott, A., Ardolino, A., Franz, T., & Pfeiffer, C. (2000). Decision making of clinical teams: Communication patterns and diagnostic error. *Medical Decision Making, 20,* 45–50.

Evans, J. St. B. T. (2008). Dual-processing accounts of reasoning, judgment, and social cognition. *Annual Review of Psychology, 59,* 255–278.

Helfand, M. (2007). Shared decision making, decision aids, and risk communication. *Medical Decision Making*, *27*(5), 516–517.

Klein, G. A. (1998). *Sources of power: How people make decisions*. Cambridge: MIT Press.

Yates, J. F. (1990). *Judgment and decision making.* Englewood Cliffs, NJ: Prentice Hall.

Yates, J. F. (2003). *Decision management*. San Francisco: Jossey-Bass.

Yates, J. F., Price, P. C., Lee, J.-W., & Ramirez, J. (1996). Good probabilistic forecasters: The "consumer's" perspective. *International Journal of Forecasting, 12,* 41–56.

Decision Psychology

Decision psychology is a scientific discipline with two main dimensions: choice and values underlying choice. Decision psychology can be undertaken to describe how humans make decisions; how humans should make decisions; what humans can do if they would like to change the way they make decisions; how much an individual understands about a decision; how much risk an individual is willing to take in an uncertain decision; how to influence the decision making of others; how to control or prevent unwanted influences by others on decision making; if and when to implement surrogacy decision making (the individual or someone on behalf of the individual deciding to give over decision making to another); whose beliefs and preferences should be incorporated in a decision; and how that process of incorporation of beliefs and preferences into a decision should be carried out.

The psychology of medical decision making can focus on individuals making decisions on their own; doctors and patients making decisions together; competent patients giving over decision-making authority to others; patients who today are fighting to preserve their decision-making abilities against progressive neurodegenerative diseases or other mental-impairing conditions that if continue unabated will eventually lead to those patients being characterized as being without decisional

capacity; and finally, patients who are now without decisional capacity and for whom decisions need to be made.

Types of Decision Making

Descriptive Versus Normative

In all types of research on the psychology of decision making, including medical decision making, the question arises whether the researcher is interested in describing how decisions are actually made by individuals confronted with a decision-making task (descriptive decision making) or whether the researcher is interested in describing how decisions that are actually made by individuals compare with a model or framework of how decisions should be made (normative decision making).

Population Versus the Individual

The basic distinction with medical decision making is whether the decision making under consideration is decision making regarding the population as a whole (or at least sizeable groups within that population) versus the individual patient. The difference between these types of decision making can be seen in the arena of immunization against disease, where the population may benefit by the immunization program but the individual may bear the brunt of death or severe morbidity from the adverse outcomes associated with the vaccine.

Studies of Decision-Making Psychology

Subjects

The subjects of studies of decision-making psychology may include hypotheses about choices, judgments, or other types of reasoning and include as study participants citizens, patients, healthcare providers (physicians and nurses), and other members of the healthcare team (social workers, chaplains, among others) and associated administrative teams (for example, information technologists), as well as students or trainees in all of these areas and more.

Psychological Models

When normative models are tested in the psychology of decision making, these models may include expected utility theory or game theory as well as psychological models, such as prospect theory.

These normative models do not exhaust the models of decision-making psychology, which also include preference theories, emotive theories, and ethical and moral theories.

Decision Making Under Risk

The classic decision-making situation is one that is common in all models of decision-making psychology: having to choose between alternatives, each of which is characterized by an estimated risk. The classic examples of medical decision-making psychology are characterized by psychologists Amos Tversky and Daniel Kahneman as embodying a form of decision making under risk.

The fact that patients do not necessarily think solely in terms of risk in decision making has also been noted, primarily in ethical perspectives on decision making. However, the focus on risk has perhaps been singled out because of the overattention that is often paid to discussions of benefits. This point was noted by Barbara J. McNeil and colleagues in one of the earliest patient-preference papers in the medical literature, published in the *New England Journal of Medicine*. In this scientific paper, McNeil and colleagues identify the fact that the very data that physicians use in published research papers in the area of oncology is the "5-year survival curve," which sets forth the best treatment as the treatment that offers the best 5-year survival. But in their study, McNeil and colleagues note that some study participants preferred not to take the short-term risks of a treatment that are often necessary to achieve this 5-year survival and would rather go with a treatment that had a better chance of short-term survival and forgo the better chances of long-term survival offered by the rival treatment.

Risky Versus Riskless Choices

The basic decision study in medical decision making is typically between a gamble (trade-off) and a sure thing. Do patients and physicians go for the gamble or do they prefer the sure thing? Risky choice, as noted by Kahneman and Tversky, is undertaken in a circumstance where there is no future knowledge about consequences. In addition, in medical decision making, the risky choice is

made on the basis of data where it may not be clear how the individual fits into the published, peer-reviewed medical literature related to the decision. Indeed, here, it is assumed that peer-reviewed medical literature associated with the medical condition or disease process that the patient has can shed light on the diagnostic or therapeutic decisions. In the real world of decision making, many times there is no published, peer-reviewed medical literature that fits the patient's case, and medical decision making thus depends highly on physician clinical experience and opinion.

Elicitation Versus Construction of Preferences

One of the key areas of research in the psychology of decision making is in the very basic notion that underlies its research: Are preferences about risky versus riskless choices actually involving preferences that are being elicited from study participants who already have formulated their preference (regarding the point about which a preference is being elicited) in the past and—as much research has assumed—are now just retrieving their previously constructed preference in response to a question being asked by the researcher? Or are the study participants actually formulating (constructing) the preferences they are offering to the researchers at the time they are being asked the question?

Future Research Areas

Jerome P. Kassirer, former editor of the *New England Journal of Medicine*, offers the following questions that need to be asked as future research areas in the psychology of decision making: First, have most of the subjects in the published medical literature to date experienced the outcomes they were asked to assess? Second, have most of the subjects undergone a preference elicitation procedure before in their life when they agree to the researcher's request to participate in the researcher's study? Third, what is to be done about the fact that preferences may well change over time? Fourth, what is to be done if different preference procedures lead to different results in the same subject? Fifth, what is to be done when the same subjects reports that he or she places the same value on a state of morbidity associated with a medical condition or disease process as on being in a state of perfect health?

To Kassirer's questions can be added two others: How do aspects such as a patient's emotions, which are present or actually elicited during the preference elicitation procedure, to be accounted for in the psychology of decision making? How are patients with strong belief and value systems to be approached by such procedures and methodologies of assessing preferences when they object to the taking of gambles?

Continued research in the psychology of decision making is needed to better understand how humans make decisions now and in what sense these same humans may want to change the way they make decisions and opt for another framework to achieve in some sense a better decision.

Defective Decision Making

Psychophysiologic correlates of defective decision making are most often discussed in relation to the dementias, yet contemporary researchers study those seemingly healthy older adults who seem to be free of obvious neurologic or psychiatric disease, but have deficits in reasoning and decision making. We will first consider decision making in the dementias and then decision making in apparently normal aging.

Decision Making and Dementia

The impact of dementias on cognitive processes and the psychology of decision making often includes a fluctuating cognition with variations in attention, alertness, and visual-perceptual problems with complex (well-formed and often detailed) visual hallucinations. Contemporary research in aging, neuropsychology, imaging, and neurophysiology are attempting to distinguish early versus later stages of dementia of various types (e.g., Alzheimer's disease, Lewy body dementias, dementia of Parkinson's disease) to aid in research on prevention of dementia. Yet contemporary research is still trying to distinguish dementias from what otherwise seem to be apparent changes of normal aging in various groups of people.

Decision Making and Apparently Normal Aging

Natalie L. Denburg and colleagues are interested in defining the psychophysiologic correlates of defective decision making in normal aging. These

researchers investigated the scientific hypothesis that some seemingly normal older persons have deficits in reasoning and decision making due to dysfunction in a neural system. The authors argued that this hypothesis (a) is relevant to the comprehensive study of aging and (b) addresses the question of why so many older adults fall prey to fraud.

The authors (in a series of three studies) investigated a cross-sectional sample of community-dwelling participants and argue that they demonstrated that a subset of older adults (approximately 35%–40%) do not perform well and appear to be working with a disadvantage on a laboratory measure of decision making that closely mimics everyday life by the manner in which it attempts to factor in reward, punishment, risk, and ambiguity.

The authors found that the same poor decision makers may also display defective autonomic responses such as those previously established in patients with acquired prefrontal lesions.

Finally, the authors present data demonstrating that poor decision makers are more likely to become the victims of deceptive strategies such as deceptive advertising. Examples of such deceit may include fraud but may also include misrepresentations encountered in other daily activities, including television broadcast advertising designed to motivate the sale of prescription medicines within the broadcast. Here, we see that the intricacies of decision making—initially described by Daniel Bernoulli in 1738 in his work on the exposition of a new theory on the measurement of risk and built by Kahneman and Tversky in the 1970s—are in turn affected by the processes of normal aging on the human brain; that is, these normal aging processes affect the very activities described above as the psychology of decision and decision making. Clearly, the psychology of decision making has to be better understood in terms of the contemporary research on normal and abnormal processes of aging and the way these affect risk and benefit consideration by humans.

Decisional Capacity

Far away from the universities where university students served as study participants in the work of Tversky and Kahneman, the concept of decisional capacity began to be developed. The notion of decisional capacity—that is, the capacity to make a decision—is often raised in two arenas: clinical care and research on human subjects.

1. In the *clinical setting*, the question is raised whether that individual has the capacity to consent to or reject the medical intervention being offered.
2. In the *research setting*, the question is raised whether that individual has the capacity to consent to or reject participation as a study volunteer in a research study to which he or she is being asked by a principal investigator.

The issue of decisional capacity is appropriately raised in areas of cognitive decline or cognitive problems, such as dementia; confusion or delirium; and mental-health cognitive biases and disease processes. The issue of decisional capacity can also be raised in the case of adults with symptomatic or asymptomatic medical conditions and disease states in intensive care units or simply in hospital wards.

Decision making in both clinical care and research on humans is complex, whether the individual is declared to have decisional capacity or not. Let us consider each domain separately.

Clinical Care

In the clinical setting, if the patient is in the hospital and declared to have decisional capacity, the question is: How long will that capacity be manifest in that individual?

In the clinical setting, if the patient is in the hospital—without an advance directive and without that individual declaring at some prior time a family member or significant other to serve as his or her surrogate decision maker—and is declared not to have decisional capacity, the question becomes: What form of substituted judgment will be used on the patient's behalf?

Research on Humans

In the research setting, if the individual has come to a hospital emergency room for care, or is admitted to the hospital for care, or is transferred from the medical ward to the intensive care unit for care, how is decisional capacity to be assessed in each of these areas? The areas of an "advance

research directive" to be prescribed by individuals before they lose decisional capacity are also open areas for research on human study volunteers.

The domains of substituted judgment in clinical care or in research on humans, as yet, have only been minimally explored by medical decision makers, with plenty of opportunities for research as our population ages and these issues multiply.

Future Research

All research underlying the psychology of decision making, from Tversky and Kahneman to Denburg and colleagues, depends heavily on questionnaire studies. The accuracy of future research depends heavily on the development of the best questionnaires to diagnose and follow individuals with cognitive and decision-making decline to make certain accurate diagnoses are made at each point in development and aging. The evaluation of patient response to therapy will also depend on treatment versus treatment comparison, which in turn will depend on optimal questionnaire studies to demonstrate the efficacy between therapies used to prevent, manage, and ideally treat, slow, and cure these conditions.

Dennis J. Mazur

See also Human Cognitive Systems; Risk Attitude; Unreliability of Memory

Further Readings

Bechara, A., Damasio, H., & Damasio, A. R. (2000). Emotion, decision making and the orbitofrontal cortex. *Cerebral Cortex, 10,* 295–307.

Bechara, A., Damasio, H., & Damasio, A. R. (2003). Role of the amygdala in decision-making. *Annals of the New York Academy of Sciences, 985,* 356–369.

Bechara, A., Damasio, H., Tranel, D., & Damasio, A. R. (2005). The Iowa Gambling Task and the somatic marker hypothesis: Some questions and answers. *Trends in Cognitive Sciences 9,* 159–162, Discussion 162–164.

Collerton, D., Burn, D., McKeith, I., & O'Brien, J. (2003). Systematic review and meta-analysis show that dementia with Lewy bodies is a visual-perceptual and attentional-executive dementia. *Dementia and Geriatric Cognitive Disorders, 16,* 229–237.

Denburg, N. L., Cole, C. A., Hernandez, M., Yamada, T. H., Tranel, D., Bechara, A., et al. (2007). The orbitofrontal cortex, real-world decision making, and normal aging. *Annals of the New York Academy of Sciences, 1121,* 480–498.

Ernst, M., Bolla, K., Mouratidis, M., Contoreggi, C., Matochik, J. A., Kurian, V., et al. (2002). Decision-making in a risk-taking task: A PET study. *Neuropsychopharmacology, 26,* 682–691.

Eslinger, P. J., & Damasio, A. R. (1985). Severe disturbance of higher cognition after bilateral frontal lobe ablation: Patient EVR. *Neurology, 35,* 1731–1741.

Grober, E., Hall, C. B., Lipton, R. B., Zonderman, A. B., Resnick, S. M., & Kawas, C. (2008). Memory impairment, executive dysfunction, and intellectual decline in preclinical Alzheimer's disease. *Journal of the International Neuropsychological Society, 14,* 266–278.

Mazur, D. J. (2007). *Evaluating the science and ethics of research on humans: A guide for IRB members.* Baltimore: Johns Hopkins University Press.

Terry, A. V., Jr., Buccafusco, J. J., & Wilson, C. (2008). Cognitive dysfunction in neuropsychiatric disorders: Selected serotonin receptor subtypes as therapeutic targets. *Behavioural Brain Research, 195*(1), 30–38.

Decision Quality

In its landmark report, *Crossing the Quality Chasm*, the Institute of Medicine set out six aims for high-quality medical care: that care should be effective, efficient, equitable, safe, timely, and patient-centered. A significant part of the quality of medical care is determined by the large and small decisions that doctors and patients make every day about seeking care, having tests, starting treatments, and stopping treatments. It is important to know to what extent decisions contribute to or detract from quality of care. To a great extent, the quality of a decision depends on the decision situation, on the perspective of the person who is judging the quality, and on what is being judged (e.g., whether it is the decision or the decision maker that is the unit of analysis). Careful attention to these issues is important to create valid and reliable assessments of decision quality.

Decision Situation

The decision situation plays a big role in determining the quality of a medical decision. There are

situations in medicine where a treatment or approach has considerable evidence of a significant benefit with considerable evidence of minimal harm. Most clinical guideline committees, such as the U.S. Preventive Services Task Force (USPSTF), set out explicit criteria for grading the clinical evidence for the benefits and harms of different tests and treatments. When the benefits are determined to outweigh the harms, there is a "right" answer that can be termed *effective care*. For example, the use of beta blockers following a heart attack fit the criteria for effective care. Most patients have a strong desire to reduce the risk of repeat heart attacks and death, and most feel the modest side effects of the medicines are worth the benefit. As a result, high-quality decisions in these situations are about efficiently delivering proven, effective care to all those who may benefit. Decision quality can be inferred by the percentage of eligible patients who receive the proved treatment, or the percentage of care that is "consistent" with the guidelines.

Not all situations in medicine are examples of effective care. In fact, a surprising number of decisions do not have sufficient evidence of benefit for one option over another, or have evidence of equivalence of two or more options, or have evidence of substantial harm that accompanies the benefit. In these situations, often called *preference-sensitive decisions,* there is not a clearly superior approach, and the preferences of individual patients are critical to selecting the "best" choice. Simply examining treatment rates will not provide enough information to determine the quality of decisions—a more sophisticated approach is needed.

Many different stakeholders have recognized this complexity and have called for attention to this challenge. The Institute of Medicine has defined patient-centered care as "healthcare that establishes a partnership among practitioners, patients and their families (when appropriate) to ensure that decisions reflect patients' wants, needs and preferences and that patients have the education and support they need to make decisions and participate in their own care" (p. 7). Researchers in the field of medical decision making have also focused on two themes that are in this definition—that patients are informed and that choices for tests and treatments reflect patients' goals and preferences. In an international consensus process, researchers, providers, policy makers, and patients overwhelmingly supported a definition of decision quality as the extent to which a decision reflects the considered preferences of a well-informed patient, and is implemented.

To assess decision quality in preference-sensitive decision situations requires assessing the extent to which patients are informed, for example, through a set of multiple-choice knowledge items. It also requires assessing patients' *considered* preferences for the potential health outcomes, their risk attitudes, and their willingness to make trade-offs over time. And finally, it requires assessing the treatment implemented. The patient's preferences would then be used to calculate value concordance, or the amount of association between their preferences and the treatments received. One approach is to aggregate over a group of patients, controlling for other factors that may influence treatments, to determine the extent to which the variation in treatments is explained by variation in patients' preferences. This could be used to compare different hospitals or providers. For example, those who are consistently able to inform their patients about the key facts of the situation, and those who are able to document that patients' preferences for key health outcomes are significantly associated with their treatment rates would be able to demonstrate higher decision quality than those who cannot inform their patients and who are not able to show any consistent association between patients' preferences and treatments.

Conceptual Framework

This definition reflects the commitments of normative decision theory, which holds that a decision should be judged by the process by which an alternative was selected rather than by the outcomes that resulted from the decision. Most normative theories make the assumption that individuals are self-interested and goal-directed in their behavior. Actual behavior shows that people make decisions that are not consistent with self-interest, as people often pay attention to irrelevant factors and make suboptimal decisions. Many normative theorists ascribe these gaps to inattention, ignorance, or lack of adequate elicitation of preferences. Others use these as a starting point for research into the heuristics that people use and how they attempt to simplify complex decision situations so as to minimize

cognitive load, conflict, and other issues. These prescriptive approaches still assume a fairly analytic mode of processing, although recognizing boundaries and limits.

An alternate view of decision making emphasizes a very different cognitive mode of information processing. Intuitive modes are fast, unconscious, and rely on heuristics and rules of thumb. The processing is tacit and is colloquially referred to as "going with your gut" or "sleeping on it." Studies in the laboratory and in real life have found that consumers view some choices more favorably when made in the absence of attentive deliberation. In these studies, the "quality" of the decision was evaluated by the consumers' satisfaction or happiness with their choice. Whether consumers' views (e.g., happiness with their selection) are an appropriate proxy for the quality of medical decisions has been debated.

The focus on process does not mean that outcomes are not important in the evaluation of decisions. In fact, over multiple decisions, it is assumed that this type of logical process will achieve better outcomes. In financial or economic decisions where money is used to measure the "value" or outcome, it is fairly easy to evaluate or compare groups of decisions (e.g., return on a stock portfolio using a dartboard to pick stocks as compared with using a more logical process that incorporates information and preferences). For healthcare, however, there is not a fungible outcome measure that is universally valued. Time is not fungible; it cannot be bought or sold, and neither can health. Many might assume that survival could be a clear outcome measure to compare treatments or decision protocols; however, many studies have documented variable tolerance to trading length of life and quality of life. Evaluating medical decisions by whether they produce better outcomes is difficult—in large part because there is no fungible measure.

That does not mean that the appropriate approach is to ignore outcomes. Studies have shown that the laypeople often view the outcomes as more important than the process used to get there. An oft-cited example is that most patients and family members would not be likely to agree that a decision to undergo surgery was good when the patient died during the procedure. Thus, quality of the decision depends greatly on the perspective through which it is being judged and may be evaluated differently on an individual basis than when combined as a group of events. In this same case, the provider, who has a broader context from which to evaluate decisions, may recognize that for the majority of times it has been used, the procedure has helped and that this was a good decision that had a bad outcome.

Policy Makers

Policy makers are important stakeholders in medical decision making, although their influence is often opaque to patients and sometimes to providers as well. Through decisions about benefits, coverage, access, accreditation, accounting, and financing of care, policy makers and administrators enable and constrain the options that providers may offer and that patients may accept, the amount of time they have to discuss choices, and the cost to the patient of various alternatives. For policy makers, medical decisions are statistical groupings that have economic and health implications. The tension between making decisions that benefit an individual and decisions that benefit a group are real and challenging. The quality of decisions from their perspective may not be evaluated on the extent to which individuals get what they want, but on whether, on average, the group gets better outcomes at the same or lower costs. Policies may negatively affect a minority of people for the benefit of the majority. Also, instead of individualization, they may favor stability and eliminating variation.

Karen R. Sepucha

See also Informed Decision Making; Shared Decision Making; Utility Assessment Techniques

Further Readings

Bell, D., Raiffa, H., & Tversky, A. (1988). Descriptive, normative and prescriptive interactions in decision making. In D. Bell, H. Raiffa, & A. Tversky (Eds.), *Decision making: descriptive, normative, and prescriptive interactions* (pp. 9–30). Cambridge, UK: Cambridge University Press.

Elwyn, G., O'Connor, A., Stacey, D., Volk, R., Edwards, A., Coulter, A., et al. (2006). Developing a quality criteria framework for patient decision aids: Online international Delphi consensus process. *British Medical Journal, 333,* 417.

Institute of Medicine. (2001). *Envisioning the National Health Care Quality Report* (M. P. Hurtado, E. K. Swift, & J. M. Corrigan, Eds.). Washington, DC: National Academy Press.

Klein, G. (1998). *Sources of power: How people make decisions*. Cambridge: MIT Press.

Mulley, A. G. (1989). Assessing patients' utilities: Can the ends justify the means? *Medical Care, 27,* S269–S281.

Redelmeier, D. A., Rozin, P., & Kahneman, D. (1993). Understanding patients' decisions. Cognitive and emotional perspectives. *Journal of the American Medical Association, 270,* 72–76.

Sepucha, K. R., Fowler, F. J., Jr., & Mulley, A. G., Jr. (2004). Policy support for patient-centered care: The need for measurable improvements in decision quality. *Health Affairs* (Web Exclusive): DOI: 10.1377/hlthaff.var.54.

Decision Rules

A decision rule is a decision-making tool combining fixed history and physical examination items and/or a simple diagnostic test used for explicit application to a clinical decision. Although many decisions about management of patients are accurately made on the basis of clinical judgment, some decision making can be improved through application of a standardized decision rule that has been developed and tested through a rigorous evidence-based process. Implementation of a rule can bring greater certainty to the clinician about the course of action to follow given a particular patient presentation, or it may lead to an improved ability to predict the probability of disease.

A decision rule is developed in a systematic process, using prospective studies often involving large numbers of patients, to meet an outcome determined to be clinically important and necessary for improved healthcare. The three stages of rule development are those of derivation, validation, and implementation. Derivation involves identifying decision items of the rule and ensuring that items are clearly defined and have demonstrated reliability. Validation requires analysis of whether the rule is accurate and reliable and meets the intended outcome; is acceptable to clinicians; can be used by different health professionals; and is suitable for application to diverse patient populations. The final stage of rule development involves analysis of the impact of implementation of a rule on patient management and healthcare.

Course of Action

A decision regarding referral or not for further testing is frequently required in clinical assessment. Referral may be to low-cost tests, as in the case of plain radiographs for identification of fracture, or to more expensive tests such as dual-energy X-ray absorptiometry to assess bone mineral density for osteoporosis screening. Clinical decision rules have demonstrated advantages over clinical judgment in these decisions. Further useful applications of clinical decision rules include guiding referral for cranial computed tomography for minor head injury and venous ultrasonography for lower-limb deep vein thrombosis.

Ankle and knee decision rules are examples of rules designed to explicitly suggest when to refer for radiography. The ankle and knee rules were developed to inform referral to radiography of patients with acute injury and potential fracture in primary care and emergency department settings. Impetus for development of ankle and knee decision rules arose from recognition that plain radiographs were commonly ordered for patients following ankle and knee blunt trauma from blows and falls, in the absence of fracture. High healthcare costs of unnecessary radiographs and patient time spent having the procedure were identified. Although the plain radiograph is relatively low cost, ankle and knee trauma are common, resulting in high volumes of ankle and knee radiographs and therefore substantial healthcare costs. Implementation of ankle and knee rules was intended to impact on these costs and lead to healthcare savings.

Concern of the clinician or, in some cases, the patient that a fracture may be missed can influence clinical decisions. Justification for these concerns is that if radiography is not ordered for a patient with a fracture, there could be serious consequences. Delayed or overlooked diagnosis of fracture can affect clinical outcome and may result in increased healthcare costs and lost productivity. A clinician who misses an ankle or knee fracture may be subject to claims of malpractice. For these reasons, acceptance of a rule by clinicians requires a

guarantee that a rule will identify clinically important fractures.

Ottawa Knee Rule

Although other ankle and knee decision rules exist, the Ottawa ankle and knee rules are credible, with well-documented evidence of rule development in different countries with some studies conducted independent of the developers of the rule. The example used to illustrate rule development is the Ottawa knee rule.

Five history and physical examination items form the basis for decision making in the Ottawa knee rule. Rule items are age 55 years or older, tenderness at head of fibula, isolated tenderness of patella (no bone tenderness of knee other than patella), inability to flex to 90°, and inability to bear weight both immediately and in the emergency department for four steps (unable to transfer weight twice onto each limb regardless of limping). Radiographic examination is suggested for patients with acute knee injuries with any one or more than one of the decision items. Rule items were derived in initial prospective investigation from 23 standardized variables on the basis of interassessor reliability, high correlation with fracture, and mathematical analysis.

Numerous studies have investigated the validity of the Ottawa knee rule in adult patients older than 18 years. Different analyses of sensitivity and specificity of the rule have shown similar results. Sensitivity is the proportion of patients with fracture for whom the results of the rule indicate radiography. Specificity is the proportion of patients without fracture for whom the results of the rule do not indicate radiography. The Ottawa knee rule has high sensitivity with extremely low to zero false negative rates, an important factor in acceptability of the rule to clinicians, whose main concern is not to miss a fracture. Low specificity tends to accompany high sensitivity, and this is true of the Ottawa knee rule. This could mean that healthcare costs would not be reduced as much as anticipated by rule implementation, as some patients would be inaccurately selected for radiographs. Interassessor reliability of physician interpretation of the rule is excellent. Examination of the validity and reliability of the rule in very small patient samples has shown less positive support for use of the rule by triage nurses in emergency departments.

Implementation of the Ottawa knee rule as compared with clinical judgment alone brings significant societal cost savings due to decreased use of knee radiography and, for those patients discharged promptly as a consequence of no radiography, less time spent in the clinic. Economic analysis has also considered estimates of the value of missed fracture in terms of damages awarded for delayed diagnosis of knee fracture in the event of compensation. Although very small change in sensitivity from 1.0 would result in missed fractures, studies consistently report sensitivity of 1.0, and therefore economic analysis finds negligible impact of missed fracture with implementation of the rule. The rule does have exclusion criteria, and any benefits identified of rule implementation may not apply to patients younger than 18 years and all cases where circumstances may make it difficult to obtain reliable information from the patient, such as serious communication problems for whatever reason. The success of implementation of Ottawa knee rules and also Ottawa ankle rules has led to further proposals of decision rules for other regions, with the same intent of assisting decision making regarding referral to further testing in cases of blunt trauma.

Prediction of Probability of Disease

Prediction of potentially life-threatening disease can be problematic as observed in cases of serious illness, including acute myocardial infarction, cancer, and pulmonary embolism. Decision rules are one of the approaches used to improve management of these patients.

Development of decision rules for pulmonary embolism, for example, occurred in response to worrying evidence that this potentially fatal condition frequently goes undiagnosed. Without a decision rule, difficulties in accurate separation of those with and without pulmonary embolism have been demonstrated even when clinical assessment is accompanied by a range of sophisticated and expensive tests. As well as the problem of a potentially fatal missed diagnosis, patients incorrectly diagnosed with pulmonary embolism will receive anticoagulant therapy, which they do not need, with possibility of serious side effects.

Wells Rule

The decision rule developed by Wells and colleagues for pretest probability estimate of pulmonary embolism has been thoroughly investigated and is widely recognized. The seven decision items of the Wells rule and the scoring system were derived from 40 initial items through mathematical analysis. Rule items are clinical signs and symptoms of deep vein thrombosis (leg swelling and pain with palpation of the deep veins); an alternative diagnosis is less likely than pulmonary embolism; heart rate >100 beats per minute; immobilization (bed rest, except to access the bathroom, for at least 3 consecutive days) or surgery in the previous 4 weeks; previous objectively diagnosed deep vein thrombosis or pulmonary embolism; hemoptysis; and malignancy (treatment that is ongoing, within the past 6 months, or palliative). The rule assigns points of 3.00 to the first two items, 1.5 to the next three items and 1.0 to the last two items. Patients are categorized according to their score as low probability if <2; moderate probability if 2 to 6; and high probability if >6. The Wells rule has demonstrated moderate or better interassessor reliability.

In early stages of rule development, it was decided that combining the Wells rule with the D-dimer blood test could bring benefit in identifying those without pulmonary embolism and therefore those with no need for imaging tests. A diagnostic algorithm was created, including Wells rule and D-dimer test, which has validated accuracy for identifying those patients in whom pulmonary embolism can be safely ruled out. A decision regarding probability of pulmonary embolism is made first on the basis of the Wells rule. Low, moderate and high probability groups all then undergo D-dimer test to assess for D-dimer fragments present in pulmonary embolism but also present in many other conditions. On application of the rule and D-dimer test, patients with low probability who also have a negative D-dimer test are separated out as without pulmonary embolism, and anticoagulant therapy is withheld from these patients. On prospective investigation, no low-probability patients in whom pulmonary embolism was excluded on the basis of the diagnostic algorithm subsequently died of pulmonary embolism. Using the algorithm, patients with moderate and high probability on the basis of the Wells rule are D-dimer tested and then are investigated with pulmonary angiography or ventilation perfusion scanning, both of which have demonstrated limitations.

Subsequent refinement has resulted in a diagnostic algorithm with two categories—the simple Wells rule (as compared with the original Wells rule, with three categories). The simple Wells rule categorizes patients as pulmonary embolism–unlikely in a case of a score <4 and pulmonary embolism–likely if the score is >4. Computed tomography is the preferred imaging technique in the algorithm for exclusion or confirmation of pulmonary embolism in patients with a score >4. Those with a score <4 and unlikely to have pulmonary embolism have a D-dimer test as in the original algorithm and, if this is negative, are excluded from diagnosis of pulmonary embolism; if the D-dimer is positive then patients have computed tomography. Prospective investigation has validated the safety of the algorithm. There is low risk for incorrectly diagnosing a patient who subsequently goes on to have pulmonary embolism and enhanced potential for correctly excluding diagnosis of pulmonary embolism with application of this algorithm. Wells scores, original and simple, have acceptable reliability.

Mathematical Techniques in Rule Development

Decision items of a rule are derived from a number of variables selected in a transparent process of review of the literature and consultation with relevant experts. A methodologically sound approach is to evaluate all items with possible relevance to the rule prospectively to assess association with rule outcomes. Investigation of potential rule variables involves univariate and multivariate techniques and estimates of reliability.

An accepted exemplar for rule development is the Ottawa ankle rule. The two rule outcomes are no fracture or insignificant fracture (defined as avulsions 3 mm or less across) or clinically significant fracture. In the preliminary screen of 32 clinical variables, chosen on the basis of evidence and clinical experience of investigators, univariate association and reliability of each variable were assessed. Variables with moderate or better reliability (kappa value > .6) and found to be strongly associated with a significant fracture in univariate logistic regression analysis were then analyzed with multivariate techniques of

multiple logistic regression analysis and recursive partitioning analysis.

Univariate logistic analyses, chi-square test for categorical data, and unpaired *t* test for continuous data compared one variable at a time with the outcome. Although these analyses have the advantage of simplicity, a limitation lies in the inability to demonstrate relationships between the variables. Swelling and tenderness over the medial malleolus were both initially associated with fracture in univariate analyses. However, swelling was excluded on the basis of subsequent multiple regression analyses because of finding of high correlation of swelling with tenderness and superior interassessor reliability of tenderness. Only tenderness was retained in the rule.

Initial multivariate analyses of stepwise logistic regression based on logarithmic equations resulted in a model that missed more than half the ankle fractures. Logistic regression analysis seeks overall accuracy rather than an emphasis on sensitivity and thus provided an unacceptable model. It had been determined that for clinician acceptance the ankle rule had to have 100% sensitivity for detecting clinically significant fracture. Recursive partitioning methods create branches of smaller and smaller subpopulations of patients, and this analysis yielded the accepted ankle rule, with 100% sensitivity though low specificity and the smallest number of variables. Reliability of the combination of rule items was good, kappa = .72.

Accuracy statistics differ depending on the purpose of the decision rule and are not limited to reports of sensitivity and specificity. Confidence intervals (CI) indicate the range of variability associated with rule application and should be reported with results of diagnostic accuracy. The Ottawa knee rule, for example, reported sensitivity of 1.0 (95% CI, .94–1.0) and specificity of .48 (95% CI, .45–.51). The Wells rule for pulmonary embolism, as a further example, reported in terms of probability of the disease for the different categories as follows: low pretest probability (3.4%; 95% CI, 2.2%–5%); moderate pretest probability (28%; 95% CI, 23.4%–32.2%); and high pretest probability (78%; 95% CI, 69.2%–86.0%). Likelihood ratios indicate how much an individual decision item or a rule will raise or lower the pretest probability that a patient has the outcome of interest and can be calculated from sensitivity and specificity data. A nomogram proposed by Fagan presents pretest probability, likelihood ratio, and posttest probability scales in diagrammatic form, allowing simple estimation of posttest probability with the use of a ruler if the other values are known.

Use of Decision Rules

A number of potential barriers to clinical uptake of a rule exist related to clinician knowledge, attitude, and behavior. Acquisition and retention of rule knowledge can be problematic. The volume of new evidence can be overwhelming, and clinicians may have difficulty in selecting out valuable information critical to improving their clinical practice. Application of a decision rule requires precise recall of the rule plus calculations where specified, and this may be difficult without pocket prompt cards or computer assistance aids. Clinicians may have doubts about the quality of a rule, the time it will take to implement it in practice, and uncertainty regarding what the rule may deliver for them and their patient. It may feel better to the clinician to continue to make decisions in the same way as they have always made them.

An important advantage of decision rules is certainty of an accurate decision irrespective of clinician experience. Despite this and other advantages of rule use, widespread clinical implementation does not automatically follow their development, even if the evidence is strongly supportive. Investigation of the clinical uptake of the Ottawa ankle rules has demonstrated this, with unsatisfactory reports that the rule has not been as widely used as anticipated even by informed clinicians. Inadequate use of the rule has now directed interest to barriers to uptake. Healthcare benefits of decision rules will only be fully realized when barriers to their clinical uptake are addressed.

Kate Haswell, John Gilmour, and Barbara Moore

See also Clinical Algorithms and Practice Guidelines; Diagnostic Tests; Logistic Regression; Nomograms

Further Readings

Guyatt, G., Walter, S., Shannon, H., Cook, D., Jaeschke, R., & Heddle, N. (1995). Basic statistics for clinicians: Correlation and regression. *Canadian Medical Association Journal, 152,* 497–504.

Lang, E. S., Wyer, P. C., & Haynes, R. B. (2007). Knowledge translation: Closing the evidence-to-practice gap. *Annals of Emergency Medicine, 49,* 355–363.

Laupacis, A., Sekar, N., & Stiell, I. G. (1997). Clinical prediction rules. A review and suggested modifications of methodological standards. *Journal of the American Medical Association, 277,* 488–494.

McGinn, T. G., Guyatt, G. H., Wyer, P. C., Naylor, C. D., Stiell, I. G., & Richardson, W. S. (2000). Users' guides to the medical literature XX11: How to use articles about clinical decision rules. *Journal of the American Medical Association, 284,* 79–84.

Nichol, G., Stiell, I. G., Wells, G. A., Juergensen, L. S., & Laupacis, A. (1999). An economic analysis of the Ottawa knee rule. *Annals of Emergency Medicine, 34,* 438–447.

Stiell, I. G., Greenberg, G. H., Wells, G. A., McDowell, I., Cwinn, A. A., Smith, N. A., et al. (1996). Prospective validation of a decision rule for use of radiography in acute knee injuries. *Journal of the American Medical Association, 275,* 611–615.

Stiell, I. G., Wells, G. A., Hoag, R. H., Sivilotti, M. L. A., Cacciotti, T. F., Verbeek, R., et al. (1997). Implementation of the Ottawa knee rule for use of radiography in acute knee injuries. *Journal of the American Medical Association, 278,* 2075–2079.

Wells, P. S., Anderson, D. R., Rodger, M., Stiell, I., Dreyer, J. F., Barnes, D., et al. (2001). Excluding pulmonary embolism at the bedside without diagnostic imaging: Management of patients with suspected pulmonary embolism presenting to the emergency department by using a simple clinical model and D-Dimer. *Annals of Internal Medicine, 135,* 98–107.

Wolf, S., McCubbin, T. R., Feldhaus, K. M., Faragher, J. P., & Adcock, D. M. (2004). Prospective validation of Wells criteria in the evaluation of patients with suspected pulmonary embolism. *Annals of Internal Medicine, 44,* 503–510.

Writing Group for the Christopher Study Investigators. (2006). Effectiveness of managing suspected pulmonary embolism using an algorithm combining clinical probability, D-dimer testing, and computed tomography. *Journal of the American Medical Association, 295,* 172–179.

Decisions Faced by Hospital Ethics Committees

Hospital ethics committees (HECs) are relatively new bodies in the field of healthcare. They are multidisciplinary hospital groups that assemble for the purposes of ethics education, case consultation, and policy development. A minority of HECs pursue research activities, typically about effectiveness of case consultations and policy review. Ethics committees may advise healthcare professionals, patients, and family members about dealing with troubling cases, ethical conflicts or dilemmas, and ought to provide a nonthreatening forum that allows for the airing of different opinions, and discussion of the moral justifications for choosing one course of action over another.

Ethics committees may facilitate the decision-making process. Common reasons for consultations are questions about treatment limitations, and/or who ought to be included in the decision-making process. Recommendations from HECs are typically advisory in nature and not binding, but in some jurisdictions ethics committee recommendations may have legal weight.

History

Current HECs are derived and expanded from decision-making groups from the past. In the 1950s, some Catholic hospitals formed "medico-moral" committees, to ensure that Catholic teaching on such matters as contraception, sterilization, and abortion were followed.

In the 1960s, some pioneering hospitals developed committees to choose which patients ought to receive experimental dialysis, treatment with an artificial kidney for kidney failure. Shana Alexander's article in *Life* magazine in 1962 describes a typical meeting of the Seattle Artificial Kidney Committee, comprised of a lawyer, a minister, a banker, a housewife, an official of state government, a labor leader, and a surgeon. The article, titled "They Decide Who Lives, Who Dies," describes a re-creation of the discussion about which of several patients ought to receive life-saving dialysis. The members discussed such topics as the patients' education, employment status, financial status, marital status, how often they went to church, etc. Some have called this type of committee the "God Squad," having the power to choose who would live or die, and many believe the social criteria that were addressed by this committee were unfair and inadequate. Nevertheless, this article highlighted the need for the development of decision-making

bodies that could reflect on the ethical problems posed by new technologies, and that such committees ought to have wide representation.

In 1975, pediatrician Karen Teel suggested that hospital ethics committees be established to help physicians and parents make decisions for impaired newborns, to "provide a regular forum for more input and dialogue in individual situations and to allow the responsibility for these judgments to be shared." Her recommendation was that hospital ethics committees composed of "physicians, social workers, attorneys, and theologians" might help in reviewing difficult cases. Her article was cited by the New Jersey Supreme Court in its decision in the *Quinlan* case.

Karen Ann Quinlan was 21 years old in 1975, when she stopped breathing after taking recreational drugs with alcohol. She was placed on an artificial breathing machine, but had sustained severe brain damage, and remained in a coma. Eventually, her parents asked to have her taken off the ventilator. Prior to that time, the American Medical Association held that withdrawing a ventilator to allow death to occur was unethical. The Court's *Quinlan* decision in 1976 authorized the removal of the respirator, and recognized that HECs, as described by Teel, might be useful in the review of difficult cases, and might possibly keep such cases out of the judicial system. Still, HECs did not become common in the 1970s, although that is the era that Institutional Review Boards (IRBs) were begun, to more closely regulate human experimentation in medicine.

After a series of court cases, "Baby Doe" regulations were devised in the 1980s, in response to parents who chose to refuse medical therapy for infants born with abnormalities. In turn, these cases stimulated some centers to form Infant Care Review Committees, to review which treatments made sense for impaired newborns, and to ensure that treatments were not withheld without careful review.

Despite the sporadic and ad hoc formation of all these ethics committee forebears, by 1983, only 1% of U.S. hospitals had developed HECs. That same year, the President's Commission published the guide *Deciding to Forego Life-Sustaining Treatment* and, as an appendix to this publication, described a recommendation for roles and composition of HECs.

After the Baby Doe cases, and the President's Commission report, HECs became much more common. Finally, in 1992, the Joint Commission on Accreditation of Health Care Organizations (JCAHO) added a requirement that hospitals have procedures for dealing with ethical issues. For a hospital to remain accredited it is required to have an ethics committee or some process that could provide for the functions of an ethics committee; HECs are now found in almost all hospitals.

Hospital Ethics Committee Functions

Ethics committees typically address three main responsibilities: education, policy development, and case consultation. A small minority of ethics committees, usually at larger academic centers, may also be involved in research activities.

Ethics committees typically are composed of physicians, nurses, social workers, and clergy as members. Some committees also have lawyers, hospital administrators, community representatives, and, if available, may have specialists from psychiatry, palliative care, neurology, pediatrics, transplantation, intensive care units, and/or allied health services, such as physical therapy. Larger committees are typically found at academic medical centers. It is critical that committee membership be diverse, to facilitate broad discussion from those of different backgrounds. Committee composition can be widely variable, depending on the size and resources of the hospital.

Education

The first task of an HEC is education of its own members. Members should be willing to attend regular meetings and share in the education of the group. Education may include review of previous ethics consultations, hospital policies, and relevant state laws, as well as reports from various authoritative bodies and ethics commissions. Members may share ethical problems from their discipline, or areas of expertise.

Some committee members may also plan educational activities, such as regular hospital rounds to help identify cases of concern, and conferences to educate hospital staff members and trainees. Ethics presentations, such as *How to Fill out an Advance Directive,* or *How to Select a Health Care Proxy,*

may be offered to educate the community. HEC members may review troubling cases in retrospect, to explore justifications for the actions taken, and may consider whether policy changes are necessary if multiple cases identify a common problem in the hospital.

Policy Development

HECs also help develop or review hospital policies. For example, HECs commonly write or review policies on topics such as informed consent, surrogate decision making, how to enact a do not resuscitate (DNR) order, brain death, blood transfusion and Jehovah's Witnesses, utilization of scarce resources, organ donation after cardiac death, and the hospital's code of ethics, and some may even review business contracts as part of JCAHO's requirement for organizational ethics.

Case Consultation

Not all members of HECs are prepared or trained to perform case consultation, and it can certainly be intimidating for a family member to be asked to meet with a dozen strangers to discuss difficult medical decisions for their loved ones. Instead, most clinical ethics case consultations are performed by small teams or individuals who have undergone more extensive training in bioethics and mediation. The consultant or team of consultants may review the cases later with the larger committee, or a subgroup dedicated to the consultation process. In some centers, particularly smaller hospitals, where the committee may be smaller, the entire committee may review cases. More than 81% of U.S. hospitals now have an ethics consultation service of some kind. In 1998, the American Society for Bioethics and Humanities (ASBH) published *Core Competencies for Health Care Ethics Consultation,* a guide to knowledge areas and skills useful for ethics consultation, which has been followed up by the publication of several books and series of ethics cases by clinical ethics consultants.

Although ethics consultations may be requested for a wide variety of specific problems, there are common underlying themes of conflict. The conflict may be between the patient and the healthcare team, the family and the healthcare team, different specialists within in the healthcare team, or the family and the patient. Generally, conflicts are about two major questions. These general questions are

> *What is the right or best thing to do in this situation?* And, because there may be different choices or opinions about what is the best course of action,
>
> *Who gets to decide?*

So the first category of questions above asks about treatment options or limitations. More specific examples include the following: Should we take this elderly lady off the breathing machine, and allow her to die? Should we put her on multiple machines to keep her alive as long as her heart is beating? Are we allowed to turn off her pacemaker? Should we provide dialysis, now that her kidneys have failed, even though she is unconscious? Should we feed her through a tube into her stomach, since she can no longer eat? Should we attempt resuscitation efforts if her heart stops beating? How long will it take before we can know if she will improve or recover? Can we treat her with natural herbs instead?

The second category of questions asks who ought to make decisions when there are differences of opinions about the best course of action. Specific examples include the following: Does this patient have the capacity to make choices herself? That is, can she express an understanding of risks, benefits, and alternatives to the treatment offered and make an informed choice? If she is too sick to make a choice, did she leave an Advance Directive or Living Will that serves as a written expression of her wishes? Did she name someone to make healthcare decisions for her if she is unable? How do we know what she would want? Did she ever make statements about what types of treatments she would or would not want? Does the whole family have to agree with the treatment option offered? What role does her distant relative have, who hasn't seen her in many years but demands that *everything* be done to keep her alive?

A third category of questions that commonly lead to ethics consults are not truly ethics questions, but are questions about communication, policy clarification, or support. Large hospitals are busy places, and the healthcare system is fragmented. Sometimes, it is hard to know "what to do" or "who gets to decide" because there are communication difficulties. Ethics consultants can help

identify and bring the right people together who need to share in the decision-making process. Consultants often facilitate family meetings, to get family members and key members of the treating healthcare team together at one time to clarify the situation. The following are some examples: What is the prognosis? How ought we decide what to do when one doctor tells us one thing, but another doctor tells us something else? Would a consultation from a specialty service such as neurology clarify the prognosis? Is physician-assisted suicide allowed in this state? What does the hospital policy say about making an order for DNR if the family objects?

Often, a healthcare professional will offer a treatment choice but may want extra support or agreement from other knowledgeable but neutral third parties that the choice offered was reasonable, and the ethics consult service can assist the patient, family, and healthcare professionals about evaluating justifications for making that choice. Ethics consultants must be careful to decide whether the consults they receive are appropriate for themselves to handle, or whether they should be more appropriately directed to a psychiatrist, the hospital lawyer, palliative care team, or chaplain.

Sample Process for Ethics Case Analysis

Ethics consultants must be practical, because solutions need to be found for problems involving real patients, rather than mere theoretical concerns. No one theory of ethics may be sufficient to adequately address all the questions an HEC may encounter.

Many committees use some form of casuistry for case review. In casuistry, there are no absolute moral rules. Casuistry is a method of reasoning that examines known example cases where there is general agreement that certain paradigm cases should be treated in certain ways. The case at hand is then compared with the paradigm cases, to assess the similarities and differences from them to determine an appropriate moral response. Casuistry starts with paradigmatic cases in which principles clearly apply and moves to complex or ambiguous cases.

The more similar a case is to a paradigm, the more clear the recommendation may be, and the better moral justification for a recommendation. A question often asked is, "Are there morally relevant differences why we should treat this case differently than another case?" State and federal laws, hospital policies, and the results of prior well-known cases may set limitations on possible recommendations. It is the specific details of the case in question that determines the final recommendation.

A number of approaches to case review using an underpinning of casuistry have been developed. These approaches are ways of organizing the information of a particular case that may allow for comparison with other cases. One well-known approach has been called the Four Topic method, described by Albert Jonsen, Mark Siegler, and William Winslade. For each case, details must be evaluated in each of four main areas: Medical Indications, Patient Preferences, Quality of Life, and Contextual Features.

Another more recent approach developed by the National Center for Ethics in Health Care of the Veterans Health Administration uses the acronym CASES. The CASES approach recommends the following steps: Clarify the consultation request, Assemble the relevant information, Synthesize the information, Explain the synthesis, Support the consultation process. More details of this method can be found on the National Center for Ethics Web pages.

Authority of Ethics Consultants

In most instances, the role of the ethics consultants is to facilitate the decision-making process, not to make decisions themselves. Decision making in healthcare is properly left between the physician who has knowledge of the treatment options and the patient who has to undergo some treatment or his or her appropriate representative.

Although the ethics consultant may offer a recommendation, the final decisions are often left to those who will be most affected by the decision. Ethics consults are most often advisory in nature, and not binding. However, some jurisdictions allow ethics committees to have more legal weight.

One of the most common reasons physicians request ethics consultation is when they believe the therapy they are providing is futile or nonbeneficial, but the patient or his or her representative asks to continue treatment, even though the chance of meaningful recovery is exceedingly low. Physicians and ethicists tried to better define

"futile" treatment in the 1990s, but were unable to come to an agreed-on definition. In 1999, Texas was the first state to adopt a law regulating end-of-life decisions, providing a due process mechanism for resolving futility disputes. This law, signed by then Governor George W. Bush, has been tested in the courts. The Texas Advance Directives Act directs if an attending physician refuses to honor a patient's or family's request for continued treatment, the refusal shall be reviewed by an ethics committee. If the ethics committee determines that the life-sustaining treatment is medically inappropriate, the family may attempt to transfer the patient to another physician or another facility. If no facility agrees to accept the patient in 10 days, then the life-sustaining treatment can be withdrawn, even over the family's objections. The first case that received national attention was that of Baby Sun Hudson, who was taken off a ventilator in March 2005, after a court reviewed the process followed by the hospital. The Hudson infant had a condition that would not allow his lungs to grow. No other state has enacted such a process yet, but under the support of this law, multidisciplinary ethics consultation has helped families accept treatment limitations in many of the cases brought for review by the ethics committees in larger Texas hospitals.

How ethics committees ought to reach a conclusion is not stated in the Advance Directive Act. There is no regulation that notes whether there needs to be unanimous consensus or simply a majority vote of the ethics committee. Ethical decision making is not typically the result of democratic activities such as voting, it is about determining appropriate justification for individual actions; that is one reason why most ethics committee decisions are advisory in nature.

Common Topics

Common topics addressed by HECs or consultants include decision-making capacity, informed consent, surrogate decision making, advance directives, end-of-life decision making, privacy and confidentiality, reproduction and perinatal issues, failure to cooperate with medical recommendations, decision making for minors, critically ill infants, discharge dilemmas, quality-of-life issues, allocation of scarce resources, and genetic testing and gene therapy. Some HECs may tackle topics and policies on human research, but intensive review of research activities is accomplished by IRBs, which are more closely regulated by federal policy. Finally, decisions about whether a patient meets criteria for listing for organ transplantation is usually addressed by transplantation committees, which may request the presence of an ethics consultant for review of a case or policy development, but is handled much differently than the dialysis committee God Squads of the 1960s.

Richard A. Demme

See also Advance Directives and End-of-Life Decision Making; Bioethics; Decisions Faced by Institutional Review Boards; Decisions Faced by Surrogates or Proxies for the Patient; Law and Court Decision Making; Shared Decision Making

Further Readings

Alexander, S. (1962). They decide who lives, who dies. *Life, 53,* 102–125.

ASBH Task Force on Standards for Bioethics Consultation. (1998). *Core competencies for health care ethics consultation.* Oakbrook, IL: American Society for Bioethics and Humanities.

In re Quinlan (1976) 70 N.J. 10, 355 A. 2d. 647.

Jonsen, A., Siegler, M., & Winslade, W. (2006). *Clinical ethics: A practical approach to ethical decisions in clinical medicine* (6th ed.). New York: McGraw-Hill.

Lo, B. (2005). *Resolving ethical dilemmas: A guide for clinicians* (3rd ed.). Baltimore: Lippincott Williams & Wilkins.

National Center for Ethics in Health Care: http://www.ethics.va.gov

President's Commission for the Study of Ethical Problems in Medicine and Biomedical and Behavioral Research. (1983). *Hospital ethics committees: Proposed statute and national survey. Deciding to forego life-sustaining treatment* (Appendix F, pp. 439–557). Washington, DC: Government Printing Office.

Decisions Faced by Institutional Review Boards

Institutional review boards (IRBs) are part of the main committees within institutions authorized to

provide independent scientific and ethical review and evaluation of research studies on humans. Their tasks are to optimally protect human study participants by the review and evaluation of the risks and benefits of a research study from scientific and ethical perspectives in the context of making as certain as possible that the study is aimed at developing appropriate scientific knowledge for use by future generations of humans. IRBs accomplish these tasks by carrying out their own systematic review and evaluation of the science and the ethics of the study from the primary perspective of protecting human research study participants.

While regional IRBs and private IRBs exist, most IRBs are local to allow for the recognition of and sensitivity to local issues in the consideration of research on humans. Historically, review boards evaluated research by peer review, review by the principal investigator's peers in good standing. Today, U.S. federal regulations require IRBs to be composed of a more representative community membership—including members of vulnerable populations or those knowledgeable about and familiar with the research the IRB is charged with evaluating with respect to vulnerable subjects—in the attempt to bring multiple perspectives into the scientific and ethical review and evaluation of research studies.

Vulnerable subjects include the following:

- children, pregnant women, prisoners, and individuals who are permanently or temporarily challenged or disabled physically, mentally, or emotionally;
- individuals who because of the temporary states when exacerbations of their medical, psychological, or psychiatric conditions occur will have impaired capacity to make medical decisions; and
- individuals who are challenged by their educational level or social status, with regard to their capacity to enter into the discussions entailed in understanding the nature of research on humans and the implications of their participation in research.

While the importance of decisional capacity is essential to the participation of anyone in a research study, IRBs lack precise criteria defining "decisional capacity." This entry reviews key elements related to understanding the nature of voluntary research participation and an IRB's responsibilities relating to research involving human subjects.

Research Participation

Research as defined by the *U.S. Code of Federal Regulations* is a "systematic investigation including . . . development, testing, and evaluation, designed to develop or contribute to generalizable knowledge" (38 *CFR* 16.102 d and 45 *CFR* 46.102 d). Any development of innovative therapies in clinical care needs to be formulated into a research study as soon as feasible. For example, a radiologist may devise a stent for a patient with an abdominal aortic aneurysm where the patient has a variant anatomy that will not allow use of a regular-sized and regular-shaped stent in an emergency. But the concept that a different form of stent for repair of abdominal aortic aneurysms can be designed and developed that better meets the various anatomical requirements of variously sized and shaped human beings is a research hypothesis that needs to be submitted to regulators as a new research medical device and subsequently reviewed by an IRB as a research study with a well-developed scientific protocol and well-developed informed consent form for consideration of approval as a research study within an institution or set of institutions.

Research studies are designed to attempt to develop general knowledge for use by future populations by recruiting study participants to serve as human subject volunteers who will bear risks of study participation even though they may not benefit in any way from their research participation and who may be reversibly or irreversibly mentally, physically, or emotionally harmed by their participation in a research study.

Therapeutic Misconception

Although a survey has found that individuals volunteering their participation in a research study prefer to be referred to as *study participants,* the term *human subject* is often used in an attempt to make certain that study volunteers do not misinterpret that they are involved in research, not clinical care. The term *therapeutic misconception* has been used in the peer-reviewed medical literature to

identify the phenomenon of study participants misunderstanding what they are getting involved with when they volunteer for a research study. Some study participants may mistakenly assume research participation is a form of clinical care.

Clinical Care

Participating in a research study is not clinical care. In clinical care, in absence of emergency, a patient presents himself or herself to a physician for care which typically involves diagnosis (history, physical examination, laboratory testing of bodily fluids, imaging of various parts of the body) to identify the medical cause of a patient's symptoms. Once a clinical diagnosis is made, the physician then develops a management and treatment plan to cure, manage, or alleviate the patient's symptoms. Diagnosis, management, and treatment may be the results (end products) of previous research on humans, but they do not constitute research.

The peer-reviewed medical literature postulates a set of reasons why research can be mistaken for clinical care: (a) Research on humans is conducted in the same clinical environment as patients see their clinicians and within which the patients receive their care, (b) research is conducted by the same physicians who care for the patient for their clinical medical conditions, and (c) research is conducted by the same medical providers who the patient sees assisting in or providing their clinical care.

In 1978, the National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research pointed out in the *Belmont Report* that research is not clinical care because the patient seeing a medical provider is expecting to benefit from the medical opinions and recommendations of that provider. In contrast, research is conducted in those circumstances where there is no answer as to what is the best way to help or benefit a patient or a group of patients with specific medical complaints or condition. Therefore, research is conducted to attempt to get better answers to the medical questions that are not understood in terms of the best way to diagnose, manage, or treat a patient; for example, how a treatment will compare with a placebo (placebo-controlled trial) or how one treatment (Treatment 1) will compare with another treatment (Treatment 2). There is no certainty in the outcome of any research study.

Also, research is not clinical care because the primary goal of the research team is to observe and evaluate potential participants at $time_0$ (the time before any research activities have been started) until $time_n$ (where n = the time of study closure or a time after study closure). The key to the majority of research studies is to determine by observation and measurement whether and to what extent the research intervention causes a change in the study participant. A study intervention may be an exposure to a newly developed medical product (device or prescription medicine), instrument, or intervention (invasive or noninvasive) used to screen (identify disease in asymptomatic individuals), diagnose (identify disease or medical conditions in symptomatic individuals), manage, or treat disease.

The purpose of the research study in the above cases is to determine if the newly developed medical product, instrument, or intervention compares in terms of benefits and risks with the study placebo, product, or intervention to which it is being compared in the research study. These comparisons are done by techniques used to observe, measure, and compare the newly developed research entity with the entity now used in the standard practice of care.

Obligations of an Institutional Review Board

The obligation of an IRB is to best protect human study volunteers. This is accomplished by the thorough systematical review and evaluation of the research study, its scientific objectives and its scientific goals, and the scientific methods selected by the principal investigator and the study sponsor to achieve those objectives and goals. The IRB meets its obligation by reviewing and evaluating each research study in terms of its science and its ethics.

Science and Ethics Evaluation

The IRB has a dual role in review of proposed research studies. First, the IRB must be certain that it fully understands the science that is being undertaken. Second, after fully understanding the science, the IRB must fully explore the ethical issues surrounding that science. While one may argue

that in particular research studies, the evaluation of the science and ethics can go on simultaneously, this is not always true. In the evaluation of some research study proposals, if the IRB does not understand the science that is being proposed, it cannot understand the ethical issues surrounding that science.

The evaluation of the science and ethics of a submitted research study is done to achieve the best possible science to both minimize risks to study volunteers and to generate the best possible scientific knowledge. To begin to achieve the above goals, the principal investigator and study sponsor submit to the IRB for review and evaluation a study protocol and an informed consent form.

Scientific Protocol

The scientific protocol describes the research question (study objective), the scientific methodology that will be used to answer the research question, the composition and qualification of the research team, and the surveillance practices that will be put into place to identify any potential harm to a study participant. The ongoing observation of all study participants to look for any harm occurring related to study participation is an ongoing obligation of the principal investigator, research team, and study sponsor. The aim is to identify a harm and begin the chain of communication that will result in that harm being minimized and the research subject being treated as soon as that harm is identified.

There are three ongoing chains of obligations. The first chain of obligation is to attempt to make sure that no harms befall a study participant. The second chain of obligation is the obligation to recognize the occurrence of a harm to a study participant as soon as possible after that harm occurs. The third chain of obligation is to contact the study participant regarding the harm as soon as possible so that the extent of the harm can be minimized, if possible, through management and treatment. Part of this third chain of obligation is for the research team to contact those who will be responsible for the care of the participant until that harm is optimally treated and managed until resolution. It is the primary obligation of the principal investigator and study sponsor to include explicit descriptions of all three chains (prevention, recognition, and communication and care) in the study protocol and informed consent form and to ensure research team members are trained in prevention, early recognition, early communication, and early establishment of care for the injured study participant.

Informed Consent Form

The informed consent form is a form that is given to the individual considering research study participation that specifies in language that is accessible to nonscientists the study objectives, the risks of the study, alternatives available in clinical care that could be opted for instead of participating in the research study, who is funding the study, who are the members of the study team who are responsible for the conduct of the study, and the chains of obligation of recognition, communication, and care related to any harm that might befall a study participant. The informed consent form also specifies the study participant's rights should an adverse outcome happen to him or her during study participation.

Identification of Conflicts of Interest

Within the overall tasks of review and evaluation of research studies in their ethical and scientific dimensions, the IRB is responsible for identifying any and all conflicts of interest that are present in the research study and in its review. There may be conflicts of interest present in relationships on an IRB with respect to a particular study being evaluated. These conflicts of interest may be financial or nonmonetary. A financial conflict of interest would be illustrated by stock ownership of an IRB member in a company that is the study sponsor of the research study being submitted to the IRB for review. A nonfinancial conflict of interest may be a work relationship between an IRB member and the principal investigator. For example, the principal investigator of a study being submitted to an IRB may be the direct supervisor of the IRB member in question. All conflicts of interest on an IRB with regard to a particular study must be eliminated. Elimination of conflict of interest is the recusal of the IRB member from any participation in the review and evaluation of the particular study in question.

Time-Appropriate Continuing Review

The IRB is also responsible for the identification of time-appropriate continuing review points whereby the IRB rereviews and reevaluates the research study for the development of new risks, excessive risk borne by study volunteers, and prevalence and types of adverse outcomes. When an IRB, after careful systematic review, decides to approve a study it then assigns a date to review that study.

For example, any research study involving new prescription medicines that have new mechanisms of action will require early review of adverse outcome occurrence by the IRB to determine whether new risks are occurring in study participants. Here, the IRB receives, reviews, and evaluates all new risks that are occurring in study participants at all sites where the study is being conducted. The IRB reviews all new risks to make certain that the research team is handling the risks, that the study participants are notified about the adverse outcome that has occurred, and that optimal care is being provided to the injured study participant. Optimal care is provided by making all appropriate contacts with the study participants and the study participants' physicians, ensuring transparent communication about the adverse outcomes and that the study participants receive appropriate care, management, and treatment.

If problems do occur, the IRB makes certain that the study is suspended until the study is modified to minimize any subsequent occurrence of the adverse outcomes in other study participants and to make certain that all study participants are willing to continue in the research study after these new risks are identified. This latter point may require that all study participants reconsent regarding their willingness to continue in the research study with the new identified risks reported. If there is unwillingness on the part of the principal investigator or study sponsor to follow IRB recommendations regarding the safety of the study volunteers, the IRB must terminate the study and notify appropriate regulators and authorities.

Communication of Risks

The IRB ensures that the principal investigator and study sponsor have correctly and clearly identified and communicated all known risks and that reasonably estimated risks are clearly described in the study protocol and in the informed consent form. For example, there must be a true conceptual search for what risks a new prescription medicine with a new mechanism of action might reasonably have. This may demand consultation with experts in the field. And at minimum, the IRB conducts its own searches of the peer-reviewed medical and scientific literature to ensure that all risks are being recognized and stated clearly in the informed consent form.

Patients' Understanding of Rights

The IRB ensures that the informed consent form does not attempt to mislead study volunteers about their rights regarding research participation. Rights here include the right to terminate participation in the research study at any time when they can do so safely. The inclusion of the point of safely terminating research participation is crucial because, for example, in the study of a prescription medicine, it must be recognized by the participant (and made clear in the informed consent form at study entry) that some study prescription medicines (e.g., beta blockers) cannot simply be stopped at any time, but rather must be tapered off safely under a physician's supervision to minimize adverse outcomes.

Another example where a research study cannot simply be stopped is that of a medical device requiring surgical placement. In these cases, an operation must be scheduled to surgically remove the device. Again, it is necessary that all study volunteers understand these points at their entry into the study and all points must be transparently disclosed in the study's informed consent form.

Decision-Making Tasks

The IRB's main decision-making tasks involve protection of human subjects. These tasks at minimum are dependent on the IRB making as certain as possible that (a) it has all known information related to the research study and (b) the information in the informed consent form is translated into nonscientific language and its exposition and presentation are as clear as possible to the study participant.

Relevant Information

Before the IRB can begin to protect study participants, it must have all relevant information from principal investigators and study sponsors regarding what is known about the research entity to be studied. The IRB then on its own systematically rechecks the peer-reviewed medical literature and calls on experts to make certain that the information provided by the principal investigator and study sponsor is consistent with what is medically known and scientifically understood about the research entity being studied, including all risks that are reasonably foreseeable. The concept of what it means for a risk to be "reasonably foreseeable" must be explored by the IRB because a precise operational definition is open to debate. Securing a wide range of expert opinion and the IRB's own thorough exploration of the peer-reviewed medical and scientific literature are good places to start in determining reasonably foreseeable risks.

Scientific and Legal Information

The IRB must be certain that the language used in the scientific protocol and informed consent form is not being used to hide information. For example, it is not sufficient to simply say to the study volunteer considering study participation (or to state in an informed consent form) that there are "unknown risks" when those unknown risks are in fact known risks that can be specified. The principal investigator and study sponsor's obligations are to disclose risks, not to hide risk from disclosure. In addition, medical terms need to be translated into nonscientific language, and legal terms need to be translated into nonlegal language.

From the scientific perspective, while it is possible that any new research entity may possess unknown risks, including risks of severe adverse outcomes or increasing risk factors for other disease or medical conditions, the IRB needs to fully explore all foreseeable risks related to the research entity being studied and independently verify if the estimates provided by principal investigators and study sponsors are declared and fully described to each study participant as he or she considers whether to volunteer participation in a study.

From the legal perspective, there should be no use of language that attempts to minimize the liability of study sponsors, research institutions, and principal investigators. There must also be a full disclosure of participants' rights to seek court opinion in specific areas of liability.

Dennis J. Mazur

See also Informed Consent

Further Readings

Appelbaum, P. S. (1997). Rethinking the conduct of psychiatric research. *Archives of General Psychiatry, 54,* 117–120.

Bonnie, R. J. (1997). Research with cognitively impaired subjects: Unfinished business in the regulation of human research. *Archives of General Psychiatry, 54,* 105–111.

Code of Federal Regulations: http://www.gpoaccess.gov/cfr/index.html

Elliott, C. (1997). Caring about risks: Are severely depressed patients competent to consent to research? *Archives of General Psychiatry, 54,* 113–116.

Lavery, J. V., Grady, C., Wahl, E. R., & Emanuel, E. J. (Eds.). (2007). *Ethical issues in international biomedical research: A casebook*. New York: Oxford University Press.

Manson, N. C., & O'Neill, O. (2007). *Rethinking informed consent in bioethics*. Cambridge, UK: Cambridge University Press.

Mazur, D. J. (2007). *Evaluating the science and ethics of research on humans: A guide for IRB members.* Baltimore: Johns Hopkins University Press.

Mazur, D. J., & Hickam, D. H. (1993). Patient interpretations of terms connoting low probabilities when communicating about surgical risk. *Theoretical Surgery, 8,* 143–145.

National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research. (1979). *The Belmont report* (DHEW Publication No. (OS) 78-0012). Washington, DC: Government Printing Office.

Office of Civil Rights, Department of Health and Human Services. (2002). Standards for privacy of individually identifiable health information: Final rules. *Federal Register, 67,* 53182–53273.

Rosen, C., Grossman, L. S., Sharma, R. P., Bell, C. C., Mullner, R., & Dove, H. W. (2007). Subjective evaluations of research participation by persons with mental illness. *Journal of Nervous and Mental Disease, 195,* 430–435.

Decisions Faced by Nongovernment Payers of Healthcare: Indemnity Products

See Decisions Faced by Nongovernment Payers of Healthcare: Managed Care

Decisions Faced by Nongovernment Payers of Healthcare: Managed Care

The concept of private indemnity insurance in healthcare refers to a fee-for-service plan where beneficiaries are compensated for their out-of-pocket costs, up to the limiting amount of the insurance policy. Unlike managed care organizations (MCOs), which overlay tools to control the utilization and cost of services, private indemnity insurance policies allow beneficiaries unrestricted provider choice and reimburse providers on a fee-for-service basis. Many indemnity plans are offered with deductibles, where the beneficiary will be required to pay copays (generally determined with percentages) for additional services required above the deductible amount.

When a private indemnity plan begins to control costs by restricting the choice of providers, it is typically referred to as an MCO. It is useful to think of MCOs on a continuum of loosely to more heavily managed products and from highest to lowest patient cost-sharing, beginning with private indemnity plans on the loosest end, progressing to preferred provider organizations (PPOs), open panel health maintenance organizations (HMOs), and finally closed panel HMOs with the tightest control over cost combined with the lowest patient cost sharing. In general, physicians are most affected in their medical decision-making ability in dealing with the most tightly managed MCOs.

Introduction to Managed Care

Managed care refers to the systematic method of reducing healthcare costs while attempting to improve the quality of patient care. A managed care organization uses these methods to finance and deliver healthcare to people enrolled in the organization's plan. Prompted by the Health Maintenance Organization Act of 1973, which provided grants and loans to assist in the startup of health maintenance organizations, today's managed care environment consists of a variety of private health benefit programs. Widely credited with restraining the runaway medical cost inflation of the late 1980s, managed care has come under attack in recent years by those who say it focuses on efficiency at the expense of patient care. Despite criticism, managed care has become an entrenched foundation of today's national healthcare system, with roughly 90% of insured Americans enrolled in plans with some form of managed care.

Characteristics of Managed Care Organizations

Managed care organizations typically provide a panel or network of healthcare professionals who deliver a comprehensive assortment of healthcare services to enrollees. MCOs usually have specific standards for selecting the providers in the network and for establishing formal quality improvement and utilization review programs. In addition, they tend to focus on preventive care and to offer economic incentives that encourage enrollees to use care efficiently.

MCOs employ a number of techniques to reduce costs and make the delivery of services more efficient while ensuring high quality standards. These may include the following:

- Financial incentives for physicians and patients to select more efficient forms of care from providers who are in the panel
- Mechanisms for reviewing the medical necessity of services
- Beneficiary cost sharing
- Restrictions on inpatient hospital admissions and length of stay
- Selective contracting with healthcare providers
- Rigorous management of the most costly healthcare cases

Additionally, MCOs often cut expenses by negotiating favorable fees from their panel of healthcare providers, choosing cost-effective providers, and

offering economic incentives for providers to practice more efficiently. MCOs may also rely on disease management, case management, wellness incentives, patient education, utilization management, and utilization reviews as indirect ways of lowering costs.

On the surface, there appears to be a conflict between the goals of an MCO and the goals of a physician, with the MCO focusing on cost and efficiency and the physician focusing on quality of care and the best care for the patient. However, when managed care works as intended, the final outcome is the most appropriate evidence-based care delivered by providers and producing the highest-quality, best value outcomes for the patient.

Contracting With Health Plans

The contracts that physicians develop with health plans can have a major impact on their medical decision making. The following are key factors in selecting contracts.

The scope of the plan's network. This affects a physician's ability to refer patients to the healthcare providers they want.

Carve-out networks. These are specialty-specific and ancillary health networks with which the plan has subcontracted to provide services—such as behavioral health, laboratory work and imaging—that either fall outside of the medical insurance benefit (e.g., vision and dental care) or that traditionally represent a high-cost service.

The plan's medical director. A good rapport with the plan's medical director can be helpful when navigating the health plan rules and appeals processes for noncovered services, such as additional testing and extended lengths of stay for inpatients.

Clinical guidelines. Physician decision making can be affected by the clinical guidelines that health plans follow. Most plans adhere to industry standards, such as those found in *Milliman Care Guidelines*. Such guidelines are typically updated annually with evidence-based authorization criteria that encourage high-quality care through tools such as care pathways, flagged quality measures, and integrated medical evidence.

Premium physician networks and rating systems. Some plans sell their subscribers "premium networks," which are groups of physicians who have demonstrated a high volume of successful outcomes. Certain plans even rate their physicians and publish this information to members. These plans usually provide regular feedback to physicians regarding their performance, with the goal of improving quality and efficiency. A low rating may be cause for removing a physician from the plan's premium network.

Pay-for-performance programs. These programs base provider reimbursement on high-quality results and appropriate decision making. Consequently, pay-for-performance programs tend to encourage physicians to make the most appropriate medical decisions to achieve optimal patient outcomes.

Following Health Plan Rules

Physicians who contract with a health plan are obligated to follow the rules of that plan. These directives will directly affect the physician's medical decisions and may vary from plan to plan. Health plan rules may include the following.

Precertification. Since a lack of precertification may result in the denial of payment, physicians may avoid ordering specific medical treatments or procedures that they know will be denied.

Referrals. Physician referral patterns are affected by the plan's provider network, as patients receive their maximum benefit level when referrals are made within the network.

Disease management programs. This is the process of using integrated care to reduce healthcare costs and improve the quality of life for people with chronic disease, such as coronary heart disease, cancer, hypertension, and diabetes. In the United States, disease management has become a big business, with more than half of all employer-sponsored health plans offering disease management programs. Effective disease management can reduce labor costs by cutting down on absenteeism and insurance expenses. Many disease management vendors even offer a return on investment for their programs. Disease management programs generally have their own sets of evidence-based rules. When a plan

employs disease management programs, physicians will support the care management guidelines of the program for the benefit of the patient.

Medical management. This refers to the activities, such as utilization management and quality assurance, that MCOs employ to control the cost and quality of healthcare services provided to their members. A plan's medical management guidelines may affect the physician's decision making as the plan manages resource utilization to contain costs while promoting high-quality care.

Pharmacy formularies. As established by a health plan, a formulary is a list of approved drugs that physicians may prescribe and that pharmacies may dispense. Health plan formularies are continually evaluated by groups of experts working together in committees that are commonly called "pharmacy and therapeutics" (P&T) committees. Health plans often use formularies as a managed care mechanism for controlling inventories and promoting the use of the most cost-effective products that are safe and beneficial to patients. Formularies can differ from plan to plan, and this will affect the physician's decision making when it comes to prescribing medications for patients.

Denials and appeals. In case payment or a certain treatment is denied, it is helpful if physicians and practice administrators understand the appeals process, so that they can assist patients.

Legal issues. Medical decision making is also affected in general by laws and regulations that govern the healthcare, insurance, and other related industries. These may include federal regulations such as the Health Insurance Portability and Accountability Act (HIPAA), compliance, antitrust, antikickback, and Stark laws, and Medicare fraud and abuse laws. With respect to medical decision making, ethics dictates and laws uphold that physicians should base their medical decisions on what is right for the patient rather than on payments or benefits the physician will receive as a consequence of a medical decision.

Conversations With Patients

When developing treatment plans and discussing medical options with patients, it is helpful if physicians are familiar with the basics of the patient's health insurance coverage. These details will affect the physician's decisions for each patient and may include the following:

- *Health plan and product*, including whether or not the patient is in a high-deductible or consumer-directed health plan that may generate high out-of-pocket costs due to the benefit design
- *Benefits coverage*, including noncovered services and the patient's financial responsibility for treatment
- *Coverage of drug formularies*, which may be limited to a list of preferred drugs or to generics for certain medications
- *Referral network*, keeping in mind that patients appreciate referrals to healthcare providers that are within their network

A physician who is familiar with the main elements of the market's major health plans will be on the same page with many of his or her patients.

Physician Perspectives on Managed Care

Attitudes of healthcare professionals regarding how managed care affects medical decisions may vary, depending on their affiliation and the experiences they have had. While many advocate that measures be taken to reduce unnecessary costs, a large number of physicians are understandably negative about managed care techniques that appear to take medical decision-making capabilities out of the hands of medical professionals.

Managed care has been successful in reining in costs and promoting quality in recent years; however, there is still a way to go in terms of easing the administrative burden that the MCO cost-cutting techniques place on physician practice. Such "hassle factor" issues may include the following.

Requirements regarding drugs. It is a challenge for physicians to keep current on the different plan formularies, recognizing that the patient's financial responsibility will change depending on what drug is prescribed. In addition, some managed care organizations require preauthorizations for certain medications.

Medical testing. Managed care can reduce unnecessary or inappropriate medical tests and procedures. This is a cost-saving technique that can improve patient care, but it also means that physicians who want to order medical tests are required to share clinical information with the patient's health plan. Establishing and managing this process can be resource-intensive.

Medical case management. Most health plans have a medical management function that coordinates the efforts of all healthcare providers and facilitates recommended treatment plans to ensure that appropriate medical protocols are followed and that patients achieve medical rehabilitation. Medical case management can reduce unnecessary costs while streamlining patient care, leading to faster, more successful recoveries. This process can sometimes create an adversarial relationship between physicians and health plans—with physicians wanting patients to stay in hospitals longer, on the one hand, and insurance companies seeking to keep costs low, on the other. As a result, hospitals have had to set up denial databases and invest significant resources to track down and get reimbursed for claims that were initially denied by the health plans.

Advocacy and Stewardship

In general, nonprofit hospitals and other healthcare organizations exist to serve the community by providing healthcare services that the population needs, as well as outreach programs, such as health education and wellness programs. Consequently, most physicians and other individuals who are affiliated with hospitals have a sense of stewardship in regard to the community.

A number of physician leaders urge other doctors to speak up and actively work for what they believe is right for patients and for the healthcare system in general. Through their involvement in legislative reform, the formation of physician and consumer groups, and other activities, physicians can bring about change and protect the interests of themselves, their patients, their communities, and the hospitals they serve.

Michael McMillan and Wendy Kornbluth

See also Government Perspective, General Healthcare; Government Perspective, Public Health Issues

Further Readings

Austrin, M. S. (1999). *Managed health care simplified: A glossary of terms*. Clifton Park, NY: Delmar Thomson Learning.

Blakely, S. (1998, July). The backlash against managed care. *Nation's Business*. Retrieved October 5, 2007, from http://findarticles.com/p/articles/mi_m1154/is_n7_v86/ai_20797610/pg_1?tag=artBody;col1

Cairns, K. D. (2002). *Contemporary managed care issues for physicians* (2nd ed.). Newtown, PA: Handbooks in Health Care.

Harris, D. M. (1999). *Healthcare law and ethics: Issues for the age of managed care*. Washington, DC: AUPHA Press.

The Henry J. Kaiser Family Foundation. (2004). *Kaiser public opinion spotlight: The public, managed care, and consumer protections*. Retrieved October 5, 2007, from http://www.kff.org/spotlight/managedcare/index.cfm

Kaiser Family Foundation Health Care Marketplace Project. (2007). *Health care costs: A primer*. Retrieved October 5, 2007, from http://www.kff.org/insurance/upload/7670.pdf

Kaiser Family Foundation Health Care Marketplace Project. (2007). *Trends in health care costs and spending*. Retrieved October 5, 2007, from http://www.kff.org/insurance/7692.cfm

Kongstvedt, P. R. (2001). *The managed health care handbook* (4th ed.). New York: Aspen.

Price Waterhouse Coopers for America's Health Insurance Plans. (2006). *The factors fueling rising healthcare costs 2006*. Retrieved October 5, 2007, from http://www.ahipbelieves.com/media/The%20Factors%20Fueling%20Rising%20Healthcare%20Costs.pdf

Reschovsky, J. D., Kemper, P., & Tu, H. (2000). Does type of health insurance affect health care use and assessments of care among the privately insured? *Health Services Research, 35*(1, Pt. 2), 219–237.

Tindall, W. N. (2000). *A guide to managed care medicine*. Sudbury, MA: Jones & Bartlett.

Decisions Faced by Patients: Primary Care

Primary care is defined as the level of the healthcare system that provides individuals with (a) the gateway into the system for all their needs and problems; (b) care focused on the individual and

his or her context (not disease-oriented); (c) care for all but very uncommon or unusual conditions; (d) continuity of care; and (e) the coordination or integration of the care provided by other levels of the system or by other professionals. Thus, primary care is defined by a series of functions which, in combination, are unique at this level. Countries with a strong primary care component have better health outcomes and are better at keeping costs under control.

In the United States, the ecology model of medical care reveals that on average each month, out of 1,000 individuals, 800 experience symptoms, of whom 327 will consider seeking medical care. Of those, only 217 will visit a physician in the office (113 visit a primary care physician and 104 visit other specialists). Of those visiting a physician, 21 will visit a hospital-based outpatient clinic, and of these, 8 will be hospitalized. Although it is essential to ensure quality of care at every level of the healthcare system, it is apparent that opportunities are being missed by limiting quality and safety programs to hospitals when the largest proportion of individuals seeking medical advice are doing so in primary care. Consequently, it is important to study communication and decision making in primary care because of the potential beneficial impact on the quality of care for a large number of individuals.

This entry reviews the characteristics and nature of decisions faced by patients in the context of primary care. The first part explores the characteristics and nature of decisions that are most frequently encountered in primary care. The second part outlines some examples of interventions that address the specific challenges that patients face when making decisions in this clinical context. The last section of the entry summarizes the lessons learned from these initiatives.

Characteristics and Nature of the Decisions

The National Ambulatory Medical Care Survey estimated that in 2004, a total of 910.9 million visits were made to physician offices in the United States. Although 58.9% of visits were to physicians in the specialties of general and family practice, internal medicine, pediatrics, and obstetrics and gynecology, 87.2% of all preventive care visits were covered by primary care physicians. The leading illness-related primary diagnoses were essential hypertension, malignant neoplasms, acute upper respiratory infection, and diabetes mellitus. In a large comparative study of 115,692 visits in primary care in Australia, New Zealand, and the United States, in each country, primary care physicians managed an average of 1.4 morbidity-related problems per visit. The relative frequency of health problems managed was similar across the three countries, with the five most frequent health problems covering the following clinical areas: musculoskeletal, cardiovascular, ear/nose/throat, skin, and psychosocial.

Results from cross-sectional studies of decision making also provide valuable insight into the characteristics and nature of decisions that are most frequently faced by patients in primary care. For example, in a study of 1,057 audiotaped encounters of routine office visits to both primary care physicians and surgeons, the authors observed that a total of 3,552 clinical decisions were made. However, only 9.0% of these decisions met the definition of completeness for informed decision making. In another study of family physicians' views on difficult decisions faced by their patients, participants identified the five most frequent decisions as follows: cancer therapy, antidepressant drug therapy, level of care, lifestyle issues, and screening tests. In a third study of 212 video-recorded doctor-patient consultations for routine appointments in 12 general practice surgeries in the United Kingdom, it was observed that in addition to those involving medical treatment, there was a range of decision-making opportunities that were not dealt with satisfactorily. More important, it was also observed that most decisions were made by physicians with little effort on their part to foster active participation of their patients in decisions.

Taken together, results from these studies suggest that decision making in primary care is influenced by the following principal characteristics: (a) Many problems and decisions are experienced in one single clinical encounter; (b) decisions are more likely to be about chronic conditions, preventive care, and lifestyle issues; and (c) primary care providers rarely foster active participation of their patients in decisions, which in turn might partly explain the low prevalence of informed decision making.

Interventions

In population-based surveys, individuals facing health-related decisions indicate that their preferred method for obtaining information remains the counseling offered by their physician. Patients facing decisions in primary care are no exception. Therefore, most patients expect their physician to have the necessary skills to give them adequate support for making informed decisions. Given their systematic approach to evidence, clinical practice guidelines—defined as systematically developed statements to assist practitioners and patients with decisions about appropriate healthcare for specific circumstances—have been very popular with medical organizations. However, most studies that aim at improving adherence of clinicians to recommendations of clinical practice guidelines have met with very little success.

In recent years, growing concerns regarding the absence of evidence about patient preferences in clinical practice guidelines have fostered an international interest in patient decision aids. Patient decision aids are tools designed to help patients participate in clinical decision making. They provide information on the options and help patients clarify and communicate the personal values they associate with different features of an option. When compared with usual care or simple information leaflets, patient decision aids improve decision quality and the measures of feeling informed and clear about values during the decision process.

Single Clinical Encounter

A promising initiative that may help primary care providers and their patients access a wide variety of patient decision aids consists of the implementation of call centers staffed by nurses coupled with a database of patient decision aids made available online. A second strategy that might assist decision making when many problems and many decisions are encountered in one single clinical encounter is to train healthcare providers in a generic manner so that they can improve their own decision-making process, recognize decisional conflict in their patients, and then foster better decisions.

Chronic Conditions, Preventive Care, and Lifestyles Issues

Ongoing intervention initiatives suggest that it is feasible to implement patient decision aids for chronic conditions in primary care. Indeed, many trials of patient decision aids have already focused on chronic conditions such as type 2 diabetes, osteoporosis, benign prostatic hyperplasia, or mental conditions and showed beneficial impact on patients and physicians. Interestingly, in the case of chronic conditions, patient decision aids have the potential to foster quality decision-making processes across time, places, and healthcare providers. The underlying hypothesis is that the decision aid will ensure that all healthcare providers involved in the pharmaceutical care of the patient will use the same evidence-based information to improve the quality of care. Notwithstanding when and where the patient receives care for his or her specific condition and who provides this care, a common procedure to support informed decisions by patients in primary care is being used.

Patient decision aids also reduce overuse of controversial medical procedures such as prostate cancer screening tests and lessen the underuse of beneficial public health measures such as childhood vaccination. Therefore, promoting the use of such aids in the context of primary care has the potential of improving the quality of the decision-making process of patients regarding lifestyle issues and public health recommendations. However, addressing lifestyle issues with patients in primary care contexts will require involving other healthcare professionals and extending the concept of high-quality health-related decision making from the medical office into the mainstream. Thus, in the years to come, it is expected that there will be more initiatives applying an interprofessional approach to decision making in primary care.

Active Participation of Patients

In a review of optimal matches of patient preferences for information, decision making, and interpersonal behavior, findings from 14 studies showed that a substantial portion of patients (26% to 95% with a median of 52%) was dissatisfied with the information given (in all aspects) and reported a desire for more information. Nonetheless, in the context of primary care, although patients and doctors

agree that more information needs to be made available to patients to help them make difficult decisions, they do not agree about patients' acceptance of decision aids or patients' willingness to participate actively in decision making. This is congruent with the existing literature indicating that the current level of participation of patients in decisions in clinical contexts is low. Results from a systematic review of 28 studies on the barriers and facilitators to fostering participation of patients in decisions as perceived by health professionals suggest that health professionals may be screening, a priori, which patients they believe are competent to participate in decisions. This is of some concern because physicians may misjudge patients' desire for active involvement in decision making. Therefore, interventions directed at patients and the system will be needed for patients to have direct access to the needed information.

Lessons Learned

Ensuring quality of care is dependent on ensuring the quality of the decision-making processes in clinical settings at every level of the healthcare system. This entry briefly reviewed the characteristics and challenges of decision making in primary care. It also highlighted how some intervention initiatives have addressed these specific challenges. Although several gaps in knowledge remain, there are signs that the agenda is beginning to focus on improving the quality of primary care patients' decision making by providing them with innovative decision support interventions. In turn, the impact of these interventions should translate into improved patient and population health outcomes, the ultimate goal of improved clinical decision making.

France Légaré

See also Decision Making in Advanced Disease; Patient Decision Aids; Shared Decision Making

Further Readings

Bindman, A. B., Forrest, C. B., Britt, H., Crampton, P., & Majeed, A. (2007). Diagnostic scope of and exposure to primary care physicians in Australia, New Zealand, and the United States: Cross sectional analysis of results from three national surveys. *British Medical Journal, 334*(7606), 1261.

Braddock, C. H., Edwards, K. A., Hasenberg, N. M., Laidley, T. L., & Levinson, W. (1999). Informed decision making in outpatient practice: Time to get back to basics. *Journal of the American Medical Association, 282*(24), 2313–2320.

Ford, S., Schofield, T., & Hope, T. (2006). Observing decision-making in the general practice consultation: Who makes which decisions? *Health Expectation, 9*(2), 130–137.

Gravel, K., Légaré, F., & Graham, I. D. (2006). Barriers and facilitators to implementing shared decision-making in clinical practice: A systematic review of health professionals' perceptions. *Implement Science, 1*(1), 16.

Green, L. A., Fryer, G. E., Jr., Yawn, B. P., Lanier, D., & Dovey, S. M. (2001). The ecology of medical care revisited. *New England Journal of Medicine, 344*(26), 2021–2025.

Hing, E., Cherry, D. K., & Woodwell, D. A. (2006). National ambulatory medical care survey: 2004 summary. *Advanced Data,* (374), 1–33.

Kiesler, D. J., & Auerbach, S. M. (2006). Optimal matches of patient preferences for information, decision-making and interpersonal behavior: Evidence, models and interventions. *Patient Education and Counseling, 61*(3), 319–341.

Légaré, F., O'Connor, A. C., Graham, I., Saucier, D., Cote, L., Cauchon, M., et al. (2006). Supporting patients facing difficult health care decisions: Use of the Ottawa Decision Support Framework. *Canadian Family Physician, 52,* 476–477.

O'Connor, A. M., Bennett, C., Stacey, D., Barry, M. J., Col, N. F., Eden, K. B., et al. (in press). Do patient decision aids meet effectiveness criteria of the international patient decision aid standards collaboration? A systematic review and meta-analysis. *Medical Decision Making.*

Starfield, B. (1998). *Primary care: Balancing health needs, services, and technology.* Oxford, UK: Oxford University Press.

Wennberg, J. E. (2004). Practice variations and health care reform: Connecting the dots. *Health Affairs* (Web Exclusive), DOI: 10.1377/hlthaff.var.140.

Decisions Faced by Surrogates or Proxies for the Patient, Durable Power of Attorney

Ideally, medical decisions are made collaboratively by the patient and the healthcare provider.

However, when patients cannot fully participate in their own decisions, an alternative decision-making model must be implemented. Surrogates may be required for medical decision making regarding issues of morbidity, mortality, hospital discharge, and research participation. Physicians are faced with the challenge of evaluating decisions made by surrogates. Patients who are unable to contribute to their medical decisions include children and adults who lack capacity, because they either lost or never attained capacity. Adults may lose capacity temporarily or permanently. Temporary loss of capacity may be due to a psychiatric or acute illness, while permanent loss may be the result of an acute event such as brain trauma or a degenerative condition such as Alzheimer's disease. When patients are unable to fully participate in medical decisions, healthcare professionals look to a surrogate or proxy to make decisions on behalf of patients. The issues that surrogates and healthcare professionals face, as well as controversies around the role of the surrogate as representative of patient values, are discussed in this entry. Although some details of these issues may be culturally specific, the broad ethical challenges can be found throughout all of Western medicine. For this entry, the context of American healthcare is used to illustrate these challenges.

Patients may appoint a surrogate in advance by using legal forms such as a Durable Power of Attorney for Healthcare. In the absence of such an advance directive, state laws and customs usually dictate who may act as a surrogate. These may be family members, close friends, or legal guardians appointed by the court. To make appropriate medical decisions, the surrogates should have knowledge of the patient's values and be able to adequately represent those values; it is helpful if the surrogate has specific knowledge about the patient's wishes. Most important, the surrogate should understand the role: to decide in the manner in which he or she believes the patient would decide and not based on the surrogate's own wishes in the situation. This standard is usually referred to as *substituted judgment*. When a surrogate does not sufficiently know the patient's wishes or values, the decision should be made based on the patient's best interests. Important cultural aspects of decision making may require increased communication and consideration by the healthcare professional. Decision-making considerations for surrogates vary depending on the nature of the choice, life and death, quality of life, research participation, or a discharge planning decision. A discussion of these areas would be incomplete without highlighting the contemporary controversies involved in the utilization of surrogates in healthcare settings.

Life and Death Decisions

Decisions about life and death include a variety of medical treatment choices, such as the use of dialysis, pressors, antibiotics, chemotherapy, ventilators, and artificial nutrition. Healthcare professionals should be aware of the law in the state in which they practice because the legal scope of the surrogate's role varies greatly among states and regions. For example, the surrogate's decision making may be limited only by the requirement to act consistently with the patient's best interests. However, some states, such as Missouri, New Jersey, and New York, use the *clear and convincing* evidence standard, which requires clear and convincing evidence of the patient's wishes regarding withdrawal of life-sustaining treatment. The U.S. Supreme Court validated this standard in the landmark case of *Cruzan v. Director, Missouri Department of Health,* 497 U.S. 261 (1990). Furthermore, some states place additional limitations on the power of the surrogate. For example, the New York Health Care Proxy Law places decisions regarding artificial nutrition and hydration outside the scope of a surrogate's authority unless a written advance directive specifically grants the surrogate such decision-making power. Healthcare professionals should understand the patient's values and local laws regarding life and death decisions to properly facilitate a surrogate decision maker's role.

Decisions about life and death, whether made in an intensive care unit, on a regular patient ward, or while receiving care at home, may be emotionally burdensome for some surrogates, while it provides a positive opportunity to interact with the patient for others. Although emotions play fundamental roles in good decision making, they also may obscure the decision-making process. The role of deciding for someone else may lead surrogates to second-guess decisions to the extent that they become emotionally paralyzed and incapable of

making good, reasoned medical decisions. Surrogates may mistakenly believe that they alone are responsible for deciding whether and when the patient will die. The residual impact of these emotions may be significant on both individuals and families. Emotions need to be recognized and, when appropriate, either affirmed or redirected. Identifying the emotions overlaying a decision may lead to better decision making by increasing the surrogate's awareness of the impact of his or her emotions on decision making.

Healthcare professionals may lessen the burden on surrogates by providing accurate information as well as clear treatment recommendations. The plan should be goal centered as defined by known patient values. For example, it would be inconsistent for a patient or surrogate who has chosen a pure hospice goal to insist on certain resuscitative measures. A goal-centered plan entails a coherent set of medical choices. It is the responsibility of the healthcare professionals to explain the role of the surrogate and continue to focus the discussion on the patient's global wishes and moral values. This usually includes advising the surrogate to use all sources of information and support, including the patient's friends, family, and spiritual advisors. This collaborative approach has the added benefit of distributing the sense of responsibility for choices that the patient may have wanted but that surrogates find morally troubling.

Research Participation

When medical decisions include the enrollment of a patient into research, issues of surrogate consent become more complex. Research exposes the patient to additional procedures beyond those performed only for the patient's benefit within a preferred treatment regimen established by a clinician. It is unclear in which circumstances a surrogate has the right to enter the patient into a research trial. Controversial instances in which even surrogate consent is waived to conduct research highlight concerns in this area. For example, the study of an artificial blood substitute, PolyHeme, challenges whether patients who lack capacity to consent can be ethically enrolled in any research. In general, enrollment of decisionally incapacitated patients in research where the risks are greater than minimal can only be undertaken when the patient might benefit directly and there are appropriate safeguards. In these cases, the level of justification for enrollment by a surrogate decision maker must meet a higher standard due to the increased degree of uncertainty of harm.

Advance directives for research constitutes one proposal to provide guidance to surrogates and healthcare providers about whether to enroll someone in clinical research. In such directives, a patient agrees to participate in research in general while they have decision-making capacity and before a decision must be made. Despite the attempt by the National Institutes of Health to use these documents, they are very rare and often offer little help in making particular decisions about unique, unanticipatable circumstances. In the end, the decision concerning whether to enroll an incapacitated patient into research falls to the healthcare provider and healthcare surrogate. In rare instances, states have prohibited this type of enrollment in an effort to provide protection from abuse to vulnerable populations. This protection may actually harm patients by not allowing them access to potential therapies in situations without a good treatment standard.

The protections for vulnerable populations were developed in the historical context of significant abuses and a recognition that the clinician researcher may have conflicting motivations. In cases of high-abuse potential or significant harms, a surrogate decision maker may be augmented by a patient advocate or a special independent review committee. A third-party moderator provides a perspective less entangled by the emotional responsibility to the particular person when assessing the level of acceptable risk. These third parties generally have the power to exclude patients from research participation but cannot demand their inclusion.

Controversies

Controversies in surrogate consent include fluctuations in patient capacity, retention of some capacity, evaluation of surrogates for capacity, and variations of standards by country and culture. For some illnesses, a patient's capacity to understand fluctuates over time. In these cases, a patient may have the ability to participate at one point but not at another. For example, a patient may understand the situation in the morning but not later the

same day or the next day. Except in emergency cases, capacity assessments should be performed over a length of time before the utilization of a surrogate for decision making. In cases where capacity may soon be lost, every effort should be made to consent a patient during a lucid time for anticipatable events. It is illegitimate to rely on a surrogate out of simple ease when direct consent can be attained.

Although most literature on surrogate cases describes unconscious patients as the paradigm for discussion, there are many instances where patients retain (or develop) degrees of capacity. In these cases, the surrogate and healthcare providers should allow degrees of patient participation. For instance, a normal teenager who requires a surrogate for legal reasons should still be included in the discussion about healthcare matters. Similarly, a mildly demented patient may still be able to provide broad input on values and pleasures. These cases create increasingly complicated situations for interpreting whether patient expressions are appropriate for consideration in the particular decision.

When a surrogate decision maker is identified, he or she is assumed to have decision-making capacity. However, this assumption may be challenged when inconsistent decision making arises. The healthcare provider must grapple with how to assure that good decision making occurs, while respecting the surrogate. Since the surrogate is not a patient, there may be a limited ability to formally evaluate the surrogate for cognitive capacity. In removing a surrogate from the decision-making role, the healthcare provider must articulate clear reasons for doing so beyond a simple disagreement of choice.

Finally, the way in which surrogates act may vary considerably by region and culture. This becomes most trying when there are mismatched expectations of surrogate decision making between patients and the generally accepted model within the region in which they are being treated. For instance, in the United States, competent patients are fully informed and make their own decisions. However, there are cultures in which healthcare decisions are deferred to a surrogate, often a husband, father, or eldest son. The healthcare provider must adjudicate when the patient has opted out of a cultural background and the degree to which the tradition might be considered unjust. Healthcare providers should carefully account for these various complexities when relying on surrogates.

Conclusion

A surrogate decision maker may be called on to make a variety of difficult decisions. Although only research and life and death decisions have been discussed, a similar set of issues may be applied to choices of quality of life, which may include where to send patients to reside for their best healthcare and social benefit. Because of the general value of patient participation in decision making, there is always a preference to avoid the need for surrogate decision making. However, when there are no better alternatives, the surrogate has an obligation to decide carefully, and the healthcare provider has an obligation to confirm that the surrogate enacts the role properly.

Margot M. Eves and Paul J. Ford

See also Advance Directives and End-of-Life Decision Making; Bioethics

Further Readings

Hyun, I. (2002). Waiver of informed consent, cultural sensitivity, and the problem of unjust families and traditions. *The Hastings Center Report*, *32*(5), 14–22.

Kim, S. Y., Appelbaum, P. S., Jeste, D. V., & Olin, J. T. (2004). Proxy and surrogate consent in geriatric neuropsychiatric research: Update and recommendations. *American Journal of Psychiatry, 161*(5), 797–806.

Muthappan, P., Forster, H., & Wendler, D. (2005). Research advance directives: Protection or obstacle? *American Journal of Psychiatry, 162*(12), 2389–2391.

Rabow, M. W., Hauser, J. M., & Adams, J. (2004). Supporting family caregivers at the end of life: "They don't know what they don't know." *Journal of the American Medical Association, 291*(4), 483–491.

Stocking, C. B., Hougham, G. W., Danner, D. D., Patterson, M. B., Whitehouse, P. J., & Sachs, G. A. (2006). Speaking of research advance directives: Planning for future research participation. *Neurology, 66*(9), 1361–1366.

Tulsky, J. A. (2005). Beyond advance directives, importance of communication skills and the end of life. *Journal of the American Medical Association, 294*(3), 359–366.

Decision Tree: Introduction

A decision tree is a powerful method for classification and prediction and for facilitating decision making in sequential decision problems. This entry considers three types of decision trees in some detail. The first is an algorithm for a recommended course of action based on a sequence of information nodes; the second is classification and regression trees; and the third is survival trees.

Decision Trees

Often the medical decision maker will be faced with a sequential decision problem involving decisions that lead to different outcomes depending on chance. If the decision process involves many sequential decisions, then the decision problem becomes difficult to visualize and to implement. Decision trees are indispensable graphical tools in such settings. They allow for intuitive understanding of the problem and can aid in decision making.

A decision tree is a graphical model describing decisions and their possible outcomes. Decision trees consist of three types of nodes (see Figure 1):

1. *Decision node:* Often represented by squares showing decisions that can be made. Lines emanating from a square show all distinct options available at a node.
2. *Chance node:* Often represented by circles showing chance outcomes. Chance outcomes are events that can occur but are outside the ability of the decision maker to control.
3. *Terminal node:* Often represented by triangles or by lines having no further decision nodes or chance nodes. Terminal nodes depict the final outcomes of the decision making process.

For example, a hospital performing esophagectomies (surgical removal of all or part of the esophagus) for patients with esophageal cancer wishes to define a protocol for what constitutes an adequate lymphadenectomy in terms of total number of regional lymph nodes removed at surgery. The hospital believes that such a protocol should be guided by pathology (available to the surgeon prior to surgery). This information should include histopathologic cell type (squamous cell carcinoma or adenocarcinoma); histopathologic grade (a crude indicator of tumor biology); and depth of tumor invasion (PT classification). It is believed that number of nodes to be removed should increase with more deeply invasive tumors when histopathologic grade is poorly differentiated and that number of nodes differs by cell type.

The decision tree in this case is composed predominantly of chance outcomes, these being the results from pathology (cell type, grade, and tumor depth). The surgeon's only decision is whether to perform the esophagectomy. If the decision is made to operate, then the surgeon follows this decision line on the graph, moving from left to right, using pathology data to eventually determine the terminal node. The terminal node, or final outcome, is number of lymph nodes to be removed.

Decision trees can in some instances be used to make optimal decisions. To do so, the terminal nodes in the decision tree must be assigned terminal values (sometimes called payoff values or endpoint values). For example, one approach is to assign values to each decision branch and chance branch and define a terminal value as the sum of branch values leading to it. Once terminal values are assigned, tree values are calculated by following terminal values from right to left. To calculate the value of chance outcomes, multiply by their probability. The total for a chance node is the total of these values. To determine the value of a decision node, the cost of each option along each decision line is subtracted from the cost already calculated. This value represents the benefit of the decision.

Classification Trees

In many medical settings, the medical decision maker may not know what the decision rule is. Rather, he or she would like to discover the decision rule by using data. In such settings, decision trees are often referred to as classification trees. Classification trees apply to data where the y-value (outcome) is a classification label, such as the disease status of a patient, and the medical decision maker would like to construct a decision rule that predicts the outcome using x-variables (dependent variables) available in the data. Because the data set available is just one sample of the underlying

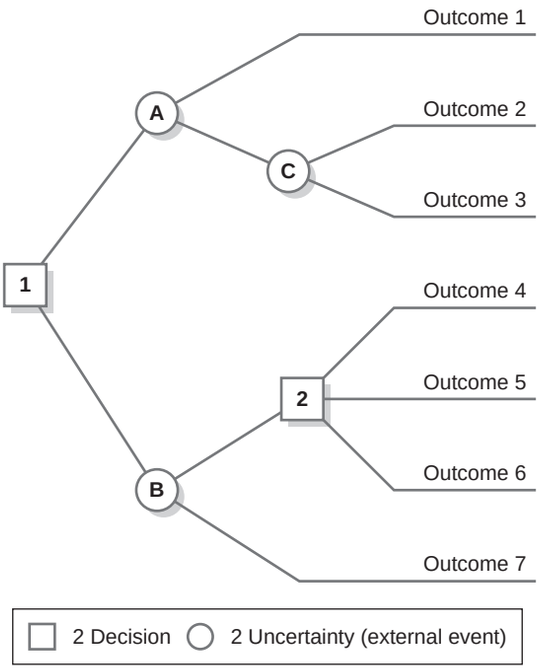


Figure 1 Decision trees are graphical models for describing sequential decision problems.

population, it is desirable to construct a decision rule that is accurate not only for the data at hand but over external data as well (i.e., the decision rule should have good prediction performance). At the same time, it is helpful to have a decision rule that is understandable. That is, it should not be so complex that the decision maker is left with a black box. Decision trees offer a reasonable way to resolve these two conflicting needs.

Background

The use of tree methods for classification has a history that dates back at least 40 years. Much of the early work emanated from the area of social sciences, starting in the late 1960s, and computational algorithms for automatic construction of classification trees began as early as the 1970s. Algorithms such as the THAID program developed at the Institute for Social Research, University of Michigan, laid the groundwork for recursive partitioning algorithms, the predominate algorithm used by modern-day tree classifiers, such as Classification and Regression Tree (CART).

An Example

Classification trees are decision trees derived using recursive partitioning data algorithms that classify each incoming x-data point (case) into one of the class labels for the outcome. A classification tree consists of three types of nodes (see Figure 2):

1. *Root node:* The top node of the tree comprising all the data.
2. *Splitting node:* A node that assigns data to a subgroup.
3. *Terminal node:* Final decision (outcome).

Figure 2 is a CART tree constructed using the breast cancer databases obtained from the University of Wisconsin Hospitals, Madison (available from http://archive.ics.uci.edu/ml). In total, the data comprise 699 patients classified as having either benign or malignant breast cancer. The goal here is to predict true disease status based on nine different variables collected from biopsy.

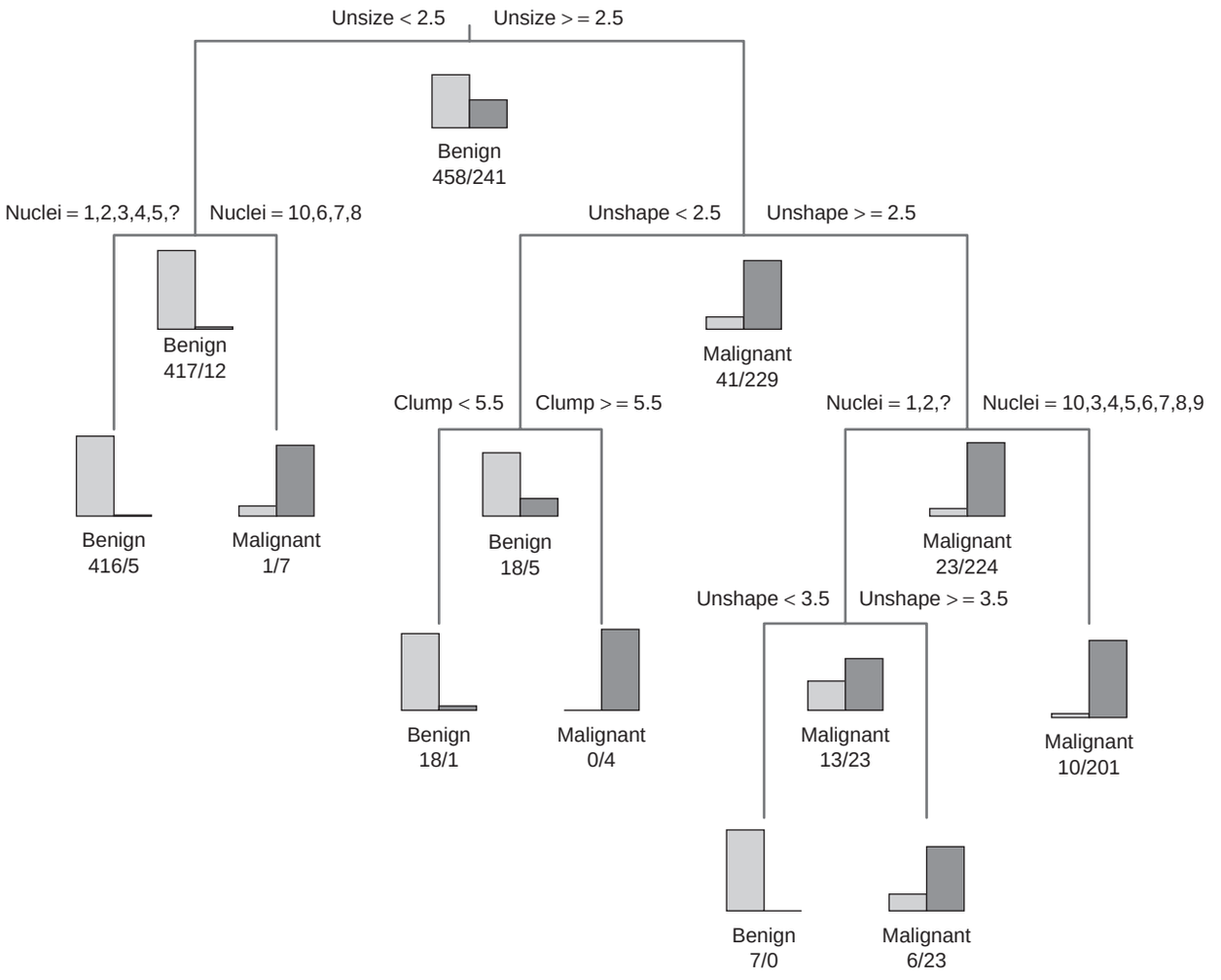


Figure 2 Classification tree for Wisconsin breast cancer data

Note: Light-shaded and dark-shaded barplots show frequency of data at each node for the two classes: benign (light shaded); malignant (dark shaded). Terminal nodes are classified by majority voting (i.e., assignment is made to the class label having the largest frequency). Labels in black given above a splitting node show how data are split depending on a given variable. In some cases, there are missing data, which are indicated by a question mark.

The first split of the tree (at the root node) is on the variable "unsize," measuring uniformity of cell size. All patients having values less than 2.5 for this variable are assigned to the left node (the left daughter node); otherwise they are assigned to the right node (right daughter node). The left and right daughter nodes are then split (in this case, on the variable "unshape" for the right daughter node and on the variable "nuclei" for the left daughter node), and patients are assigned to subgroups defined by these splits. These nodes are then split, and the process is repeated recursively in a procedure called recursive partitioning. When the tree construction is completed, terminal nodes are assigned class labels by majority voting (the class label with the largest frequency). Each patient in a given terminal node is assigned the predicted class label for that terminal node. For example, the leftmost terminal node in Figure 2 is assigned the class label "benign" because 416 of the 421 cases in the node have that label. Looking at Figure 2, one can see that voting heavily favors one class over the other for all terminal nodes, showing that the decision tree is accurately classifying the data. However, it is important to assess accuracy using external data sets or by using cross-validation as well.

Recursive Partitioning

In general, recursive partitioning works as follows. The classification tree is grown starting at the root node, which is the top node of the tree, comprising all the data. The root node is split into two daughter nodes: a left and a right daughter node. In turn, each daughter node is split, with each split giving rise to left and right daughters. The process is repeated in a recursive fashion until the tree cannot be partitioned further due to lack of data or some stopping criterion is reached, resulting in a collection of terminal nodes. The terminal nodes represent a partition of the predictor space into a collection of rectangular regions that do not overlap. It should be noted, though, that this partition may be quite different than what might be found by exhaustively searching over all partitions corresponding to the same number of terminal nodes. However, for many problems, exhaustive searches for globally optimal partitions (in the sense of producing the most homogeneous leaves) are not computationally feasible, and recursive partitioning represents an effective way of undertaking this task by using a one-step procedure instead.

A classification tree as described above is referred to as a *binary recursive partitioned tree*. Another type of recursively partitioned tree is multiway recursive partitioned tree. Rather than splitting the parent node into two daughter nodes, such trees use multiway splits that define multiple daughter nodes. However, there is little evidence that multiway splits produce better classifiers, and for this reason, as well as for their simplicity, binary recursive partitioned trees are often favored.

Splitting Rules

The success of CART as a classifier can be largely attributed to the manner in which splits are formed in the tree construction. To define a good split, CART uses an impurity function to measure the decrease in tree impurity for a split. The purity of a tree is a measure of how similar observations in the leaves are to one another. The best split for a node is found by searching over all possible variables and all possible split values and choosing that variable and split that reduces impurity the most. Reduction of tree impurity is a good principle because it encourages the tree to push dissimilar cases apart. Eventually, as the number of nodes increases, and dissimilar cases become separated into daughter nodes, each node in the tree becomes homogeneous and is populated by cases with similar outcomes (recall Figure 2).

There are several impurity functions used. These include the twoing criterion, the entropy criterion, and the gini index. The gini index is arguably the most popular. When the outcome has two class labels (the so-called two-class problem), the gini index corresponds to the variance of the outcome if the class labels are recoded as being 0 and 1.

Stopping Rules

The size of the tree is crucial to the accuracy of the classifier. If the tree is too shallow, terminal nodes will not be pure (outcomes will be heterogeneous), and the accuracy of the classifier will suffer. If the tree is too deep (too many splits), then the number of cases within a terminal node will be small, and the predicted class label will have high variance—again undermining the accuracy of the classifier.

To strike a proper balance, pruning is employed in methodologies such as CART. To determine the optimal size of a tree, the tree is grown to full size (i.e., until all data are spent) and then pruned back. The optimal size is determined using a complexity measure that balances the accuracy of the tree as measured by cost complexity and by the size of the tree.

Regression Trees

Decision trees can also be used to analyze data when the *y*-outcome is a continuous measurement (such as age, blood pressure, ejection fraction for the heart, etc.). Such trees are called regression trees. Regression trees can be constructed using recursive partitioning similar to classification trees. Impurity is measured using mean-square error. The terminal node values in a regression tree are defined as the mean value (average) of outcomes for patients within the terminal node. This is the predicted value for the outcome.

Survival Trees

Time-to-event data are often encountered in the medical sciences. For such data, the analysis

focuses on understanding how time-to-event varies in terms of different variables that might be collected for a patient. Time-to-event can be time to death from a certain disease, time until recurrence (for cancer), time until first occurrence of a symptom, or simple all-cause mortality.

The analysis of time-to-event data is often complicated by the presence of censoring. Generally speaking, this means that the event times for some individuals in a study are not observed exactly and are only known to fall within certain time intervals. Right censoring is one of the most common types of censoring encountered. This occurs when the event of interest is observed only if it occurs prior to some prespecified time. For example, a patient might be monitored for 2 weeks without occurrence of a symptom and then released from a hospital. Such a patient is said to be right censored because the time-to-event must exceed 2 weeks, but the exact event time is unknown. Another example of right censoring occurs when patients enter a study at different times and the study is predetermined to end by a certain time. Then, all patients who do not experience an event within the study period are right censored.

Decision trees can be used to analyze right-censored survival data. Such trees are referred to as survival trees. Survival trees can be constructed using recursive partitioning. The measure of impurity plays a key role, as in CART, and this can be defined in many ways. One popular approach is to

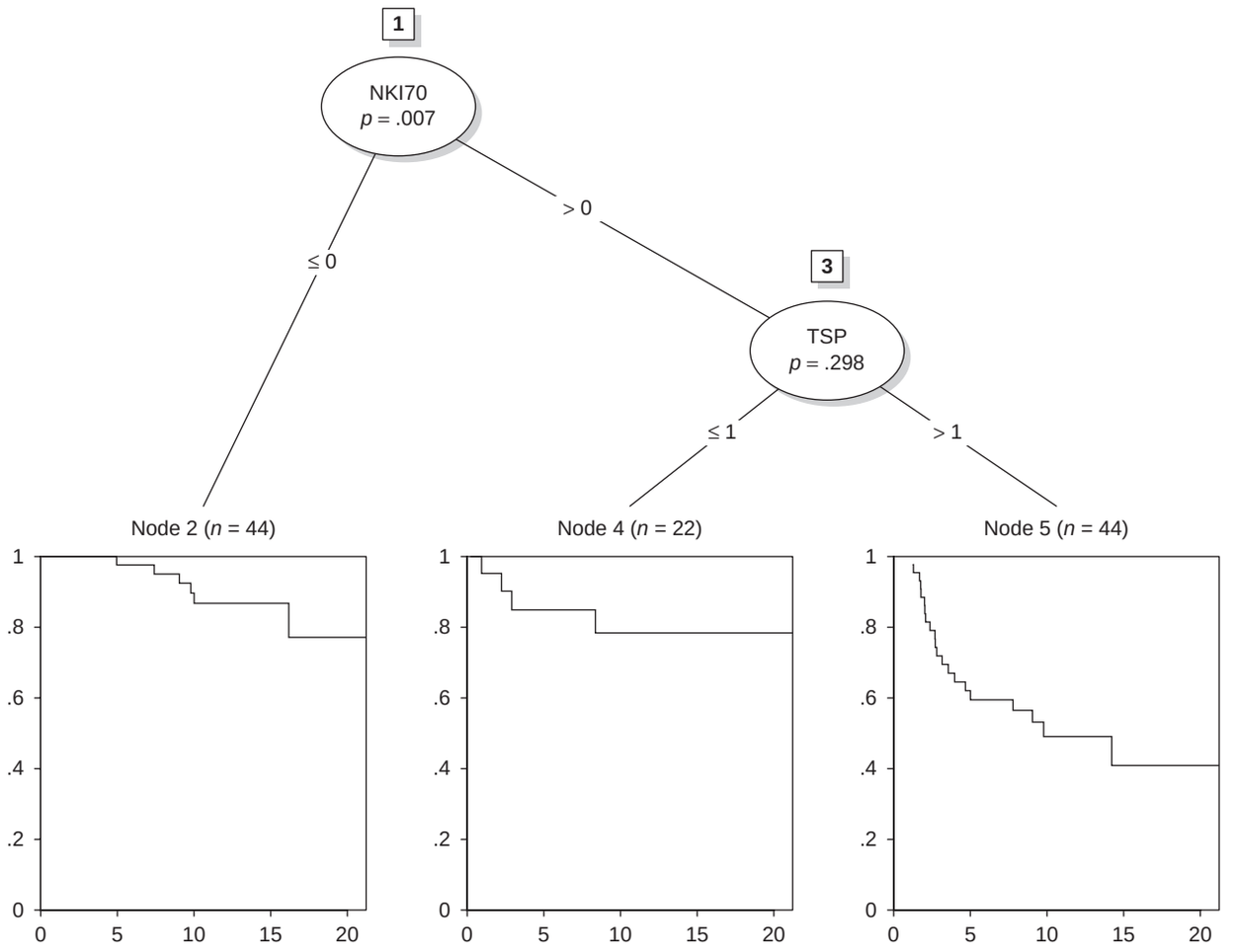


Figure 3 Binary survival tree for breast cancer patients

Note: Dependent variables NKI70 and TSP are gene signatures. For example, extreme right terminal node (Node 5) corresponds to presence of both the NKI70 and TSP gene signatures. Underneath each terminal node are Kaplan-Meier survival curves for patients within that node.

define impurity using the log-rank test. As in CART, growing a tree by reducing impurity ensures that terminal nodes are populated by individuals with similar behavior. In the case of a survival tree, terminal nodes are composed of patients with similar survival. The terminal node value in a survival tree is the survival function and is estimated using those patients within the terminal node. This differs from classification and regression trees, where terminal node values are a single value (the estimated class label or predicted value for the response, respectively). Figure 3 shows an example of a survival tree.

Hemant Ishwaran and J. Sunil Rao

See also Decision Trees, Advanced Techniques in Constructing; Recursive Partitioning

Further Readings

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). *Classification and regression trees*. Belmont, CA: Wadsworth.

LeBlanc, M., & Crowley, J. (1993). Survival trees by goodness of split. *Journal of the American Statistical Association, 88,* 457–467.

Segal, M. R. (1988). Regression trees for censored data. *Biometrics, 44,* 35–47.

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society, Series B, 36,* 111–147.

Decision Trees, Advanced Techniques in Constructing

Decision trees such as classification, regression, and survival trees offer the medical decision maker a comprehensive way to calculate predictors and decision rules in a variety of commonly encountered data settings. However, performance of decision trees on external data sets can sometimes be poor. Aggregating decision trees is a simple way to improve performance—and in some instances, aggregated tree predictors can exhibit state-of-the-art performance.

Decision Boundary

Decision trees, by their very nature, are simple and intuitive to understand. For example, a binary classification tree assigns data by dropping a data point (case) down the tree and moving either left or right through nodes depending on the value of a given variable. The nature of a binary tree ensures that each case is assigned to a unique terminal node. The value for the terminal node (the predicted outcome) defines how the case is classified. By following the path as a case moves down the tree to its terminal node, the *decision rule* for that case can be read directly off the tree. Such a rule is simple to understand, as it is nothing more than a sequence of simple rules strung together.

The *decision boundary*, on the other hand, is a more abstract concept. Decision boundaries are estimated by a collection of decision rules for cases taken together—or, in the case of decision trees, the boundary produced in the predictor space between classes by the decision tree. Unlike decision rules, decision boundaries are difficult to visualize and interpret for data involving more than one or two variables. However, when the data involve only a few variables, the decision boundary is a powerful way to visualize a classifier and to study its performance.

Consider Figure 1. On the left-hand side is the classification tree for a prostate data set. Here, the outcome is presence or absence of prostate cancer and the independent variables are prostate-specific antigen (PSA) and tumor volume, both having been transformed on the log scale. Each case in the data is classified uniquely depending on the value of these two variables. For example, the leftmost terminal node in Figure 1 is composed of those patients with tumor volumes less than 7.851 and PSA levels less than 2.549 (on the log scale). Terminal node values are assigned by majority voting (i.e., the predicted outcome is the class label with the largest frequency). For this node, there are 54 nondiseased patients and 16 diseased patients, and thus, the predicted class label is nondiseased.

The right-hand side of Figure 1 displays the decision boundary for the tree. The dark-shaded region is the space of all values for PSA and tumor volume that would be classified as nondiseased, whereas the light-shaded regions are those values classified as diseased. Superimposed on the figure,

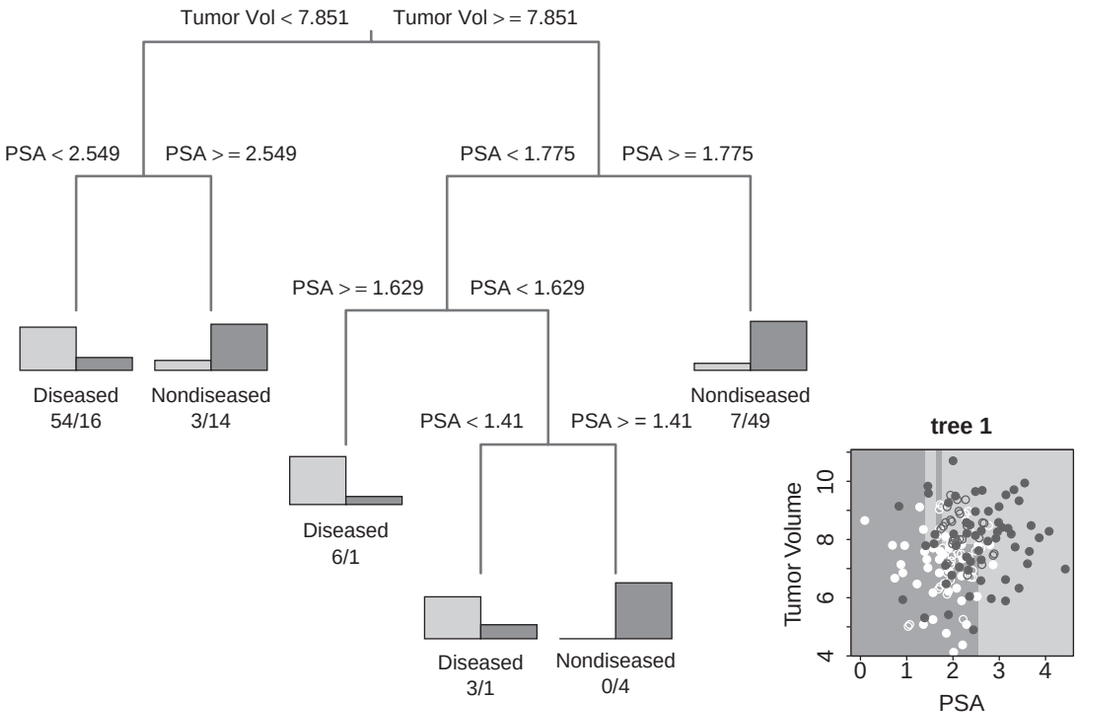


Figure 1 Decision tree (left-hand side) and decision boundary (right-hand side) for prostate cancer data with prostate-specific antigen (PSA) and tumor volume as independent variables (both transformed on the log scale)

Note: Barplots under terminal nodes of the decision tree indicate proportion of cases classified as diseased or nondiseased, with the predicted class label determined by majority voting. Decision boundary shows how the tree classifies a new patient based on PSA and tumor volume. Gray-shaded points identify diseased patients, and white points identify nondiseased patients from the data.

using white and light-gray dots, are the observed data points from the original data. Light-gray points are truly diseased patients, whereas white points are truly nondiseased patients. Most of the light-gray points fall in the light-shaded region of the decision space and, likewise, most of the white points fall in the dark-shaded region of the decision space, thus showing that the classifier is classifying a large fraction of the data correctly. Some data points are misclassified, though. For example, there are several light-gray points in the center of the plot falling in the dark-shaded region. As well, there are four light-gray points with small tumor volumes and PSA values falling in the dark-shaded region. The misclassified data points in the center of the decision space are especially troublesome. These points are being misclassified because the decision space for the tree is rectangular. If the decision boundary were smoother, then these points would not be misclassified. The nonsmooth nature of the decision boundary is a well-known deficiency of classification trees and can seriously degrade performance, especially in complex decision problems involving many variables.

Instability of Decision Trees

Decision trees, such as classification trees, are known to be unstable. That is, if the original data set is changed (perturbed) in some way, then the classifier constructed from the altered data can be surprisingly different from the original classifier. This is an undesirable property, especially if small perturbations to the data lead to substantial differences.

This property can be demonstrated using the prostate data set of Figure 1. However, to show this, it is important to first agree on a method for perturbing the data. One technique that can be used is to employ bootstrap resampling. A bootstrap sample is a special type of resampling

procedure. A data point is randomly selected from the data and then returned. This process is repeated n times, where n is the sample size. The resulting bootstrap sample consists of n data points but will contain replicated data. On average, a bootstrap sample draws only approximately 63% of the original data.

A total of 1,000 different bootstrap samples of the prostate data were drawn. A classification tree was calculated for each of these 1,000 samples. The top panel of plots in Figure 2 shows decision boundaries for four of these trees (bootstrap samples 2, 5, 25, and 1,000; note that Tree 1 is the classification tree from Figure 1 based on the original data). One can see clearly that the decision spaces differ quite substantially—thus providing clear evidence of the instability.

It is also interesting to note how some of the trees have better decision spaces than the original tree (recall Figure 1; also see Tree 1 in Figure 2). For example, Trees 2, 5, 25, and 1,000 identify some or all of the four problematic light-gray points appearing within the lower quadrant of the dark-shaded region of the original decision space. As well, Trees 5, 25, and, 1,000 identify some of the problematic green points appearing within the center of the original decision space.

An important lesson that emerges from this example is not only that decision trees can be unstable but also that trees constructed from different perturbations of the original data can produce decision boundaries that in some instances have better behavior than the original decision space (over certain regions). Thus, it stands to reason that, if one could combine many such trees, the classifier formed by aggregating the trees might have better overall performance. In other words, *the whole may be greater than the sum of the parts* and one may be able to capitalize on the inherent instability using aggregation to produce more accurate classifiers.

Bagging

This idea in fact is the basis for a powerful method referred to as "bootstrap aggregation," or simply "bagging." Bagging can be used for many kinds of predictors, not just decision trees. The basic premise for bagging is that, if the underlying predictor is unstable, then aggregating the predictor over multiple bootstrap samples will produce a more accurate, and more stable, procedure.

To bag a classification tree, the procedure is as follows (bagging can be applied to regression trees and survival trees in a similar fashion):

1. Draw a bootstrap sample of the original data.
2. Construct a classification tree using data from Step 1.
3. Repeat Steps 1 and 2 many times, independently.
4. Calculate an aggregated classifier using the trees formed in Steps 1 to 3. Use majority voting to classify a case. Thus, to determine the predicted outcome for a case, take the majority vote over the predicted outcomes from each tree in Steps 1 to 3.

The bottom panel of plots in Figure 2 shows the decision boundary for the bagged classifier as a function of number of trees (based on the same prostate data as before). The first plot is the original classifier based on all the data (Tree 1). The second plot is the bagged classifier composed of Tree 1 and the bootstrap tree derived using the first bootstrap sample. The third plot is the bagged classifier using Tree 1 and the first four bootstrapped trees, and so forth. As number of trees increases, the bagged classifier becomes more refined. Even the decision boundary for the bagged classifier using only five trees (third plot) is substantially smoother than the original classifier and is able to better classify problematic cases. By 1,000 trees (last plot), the bagged classifier's decision boundary is fully defined. The accuracy of the bagged classifier is substantially better than any single bootstrapped tree. Table 1 records the misclassification (error) rate for the bagged predictor against the averaged error rate for the 1,000 bootstrapped trees. The first column is the overall error rate, the second column is the error rate for diseased patients, and the third column is the error rate for nondiseased patients. Error rates were calculated using out-of-bag data. Recall that each bootstrap sample uses on average 67% of the original data. The remaining 33% of the data is called out-of-bag and serves as test data, as it is not used in constructing the tree. Table 1 shows that

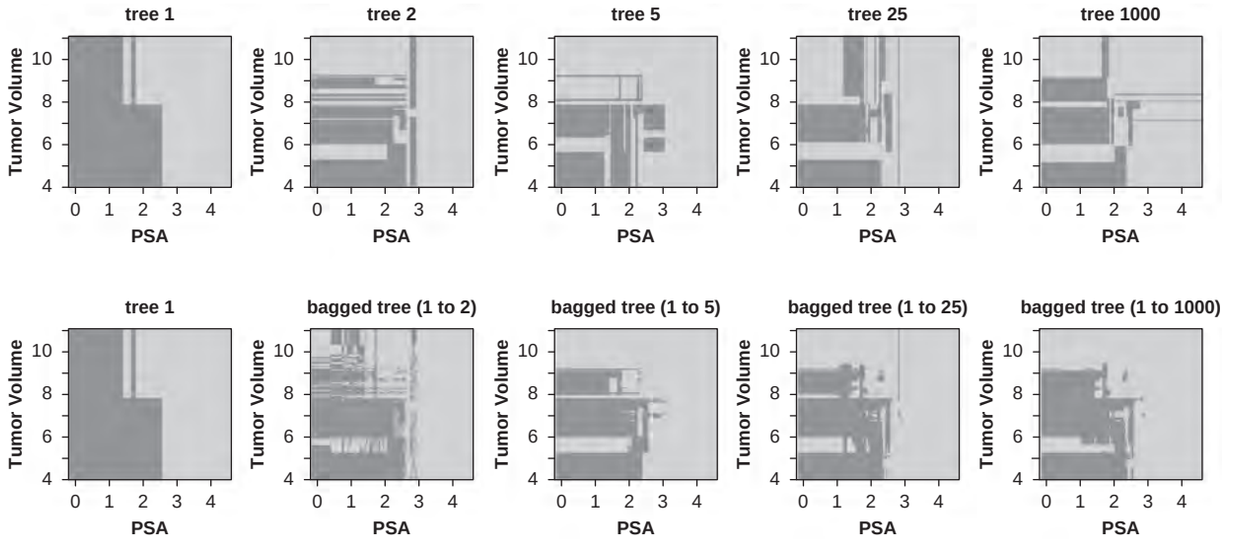


Figure 2 Top row shows decision boundary for a specific bootstrapped tree (1,000 trees used in total), and the bottom plot shows different aggregated (bagged) decision trees

Note: Bagged trees are more robust to noise (stable) because they utilize information from more than one tree. The most stable bagged tree is the one on the extreme right-hand side and shows decision boundary using 1,000 trees.

the bagged classifier is substantially more accurate than any given tree.

Random Forests

"Random forests" is a refinement of bagging that can yield even more accurate predictors. The method works like bagging by using bootstrapping and aggregation but includes an additional step that is designed to encourage independence of trees. This effect is often most pronounced when the data contain many variables.

To create a random forest classifier, the procedure is as follows (regression forests and random survival forests can be constructed using the same principle):

1. Draw a bootstrap sample of the original data.
2. Construct a classification tree using data from Step 1. For each node in the tree, determine the optimal split for the node using *M* randomly selected dependent variables.
3. Repeat Steps 1 and 2 many times, independently.
4. Calculate an aggregated classifier using the trees formed in Steps 1 to 3. Use majority voting to classify a case. Thus, to determine the predicted outcome for a case, take the majority vote over the predicted outcomes from each tree in Steps 1 to 3.

Step 2 is the crucial step distinguishing forests from bagging. Unlike bagging, each bootstrapped tree is constructed using different variables, and not all variables are used (at most *M* are used at each node in the tree growing process). Considerable empirical evidence has shown that forests can be substantially more accurate because of this feature.

Boosting

Boosting is another related technique that has some similarities to bagging although its connection is not as direct. It too can produce accurate

Table 1 Misclassification error rate (in percentage) for bagged classifier (1,000 trees) and single tree classifier

Classifier	*All*	*Diseased*	*Nondiseased*
Bagged tree	27.2	28.8	25.9
Single tree	34.9	36.7	33.0

classifiers through a combination of reweighting and aggregation. To create a boosted tree classifier, the following procedure can be used (although other methods are also available in the literature):

1. Draw a bootstrap sample from the original data giving each observation equal chance (i.e., weight) of appearing in the sample.
2. Build a classification tree using the bootstrap data and classify each of the observations, keeping track of which ones are classified incorrectly or correctly.
3. For those observations that were incorrectly classified, increase their weight and correspondingly decrease the weight assigned to observations that were correctly classified.
4. Draw another bootstrap sample using the newly updated observation weights (i.e., those observations that were previously incorrectly classified will have a greater chance of appearing in the next bootstrap sample).
5. Repeat Steps 2 to 4 many times.
6. Calculate an aggregated classifier using the trees formed in Steps 1 to 5. Use majority voting to classify a case. Thus, to determine the predicted outcome for a case, take the majority vote over the predicted outcomes from each tree in Steps 1 to 5.

The idea of reweighting observations adaptively is a key to boosting's performance gains. In a sense, the algorithm tends to focus more and more on observations that are difficult to classify. There has been much work in the literature on studying the operating characteristics of boosting, primarily motivated by the fact that the approach can produce significant gains in prediction accuracy over a single tree classifier. Again, as with bagging, boosting is a general algorithm that can be applied to more than tree-based classifiers. While these aggregation algorithms were initially thought to destroy the simple interpretable structure (topology) produced by a single tree classifier, recent work has shown that, in fact, treelike structures (with respect to the decision boundary) are often maintained, and interpretable structure about how the predictors interact with one another can still be gleaned.

Hemant Ishwaran and J. Sunil Rao

See also Decision Tree: Introduction; Recursive Partitioning

Further Readings

Breiman, L. (1996). Bagging predictors. *Machine Learning, 26,* 123–140.

Breiman, L. (2001). Random forests. *Machine Learning, 45,* 5–32.

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). *Classification and regression trees*. Belmont, CA: Wadsworth.

Efron, B. (1982). *The jackknife, the bootstrap and other resampling plans* (Society for Industrial and Applied Mathematics CBMS-NSF Monographs, No. 38). Philadelphia: SIAM.

Freund, Y., & Shapire, R. E. (1996). Experiments with a new boosting algorithm. In *Machine Learning: Proceedings of the 13th International Conference* (pp. 148–156). San Francisco: Morgan Kaufman.

Ishwaran, H., Kogalur, U. B., Blackstone, E. H., & Lauer, M. S. (2008). Random survival forests. *Annals of Applied Statistics, 2*(3), 841–860.

Rao, J. S., & Potts, W. J. E. (1997). Visualizing bagged decision trees. In *Proceedings of the 3rd International Conference on Knowledge Discovery and Data Mining* (pp. 243–246). Newport Beach, CA: AAAI Press.

Decision Trees, Construction

A decision model is a mathematical formulation of a decision problem that compares alternative choices in a formal process by calculating their expected outcome. The decision tree is a graphical representation of a decision model that represents the basic elements of the model. The key elements of the model are the possible *choices*, *information* about chance events, and *preferences* of the decision maker. The choices are the alternatives being compared in the decision model. The information consists of an enumeration of the events that may occur consequent to the choice and the probabilities of each of their outcomes. Preferences are

captured by assessing utilities of each outcome that measure the desirability of each outcome. In addition to a utility, each outcome may be associated with a financial cost.

The decision tree is a convenient method, analogous to a high-level graphical language, of specifying the elements of the decision model in a way that leads naturally to a method for quantitatively evaluating the alternative choices, in a process known as averaging out and folding back the tree.

Formulating the Problem

Decision tree construction requires a properly formulated *decision problem*.

Decision Context

The first step is determining the context of the decision. This consists, at a minimum, of the clinical problem (e.g., chest pain), the healthcare setting (e.g., a hospital emergency room), and any characteristics of the patient to which the analysis is restricted (e.g., the age range, gender, or existing comorbid conditions). The context also specifies the timeframe being considered.

Specific Question

The second step is formulating a specific question that is to be answered by the decision analysis. It must be a comparison of specific alternative actions that are available to the decision maker. In healthcare decision making, choices generally involve diagnostic tests and treatments. An example of a clearly formulated decision is whether a patient with a suspected condition should be observed without treatment, given a diagnostic test, or treated empirically. Each choice must be unique. Choices may also contain combinations of actions with later decisions contingent on results of tests or outcomes of observation. These combinations of choices are referred to as *policies*. Typically, decision models involve multiple successive choices, which, in combinations, correspond to alternate policies. These combinations may differ according to the specific elements (e.g., one test or treatment as compared with another) or according to how these elements are applied (e.g., using differing rules for responding to the outcome of a diagnostic test or varying the amount of time before contingent action is taken). For these reasons, the number of decision alternatives that can be considered in a decision model can become very large as the number of combinations of the various factors increases.

Node Types

Standard decision trees contain three basic types of nodes. Decision nodes are typically represented by an open square, chance nodes by an open circle, and terminal nodes by rectangular boxes. Branches are represented as straight lines connecting nodes.

Overall Tree Structure

A simple decision tree is shown in Figure 1. By convention, the root of the tree is a decision node and is represented at the left of the figure, and the terminal nodes (referred to as the "leaves" of the tree) are at the right. According to conventions for drawing decision models that are published in the journal *Medical Decision Making* in the first issue of each year, lines representing branches of the same node are parallel and vertically aligned. *Medical Decision Making* also specifies that the branches should be attached to lines at right angles to nodes, as in Figure 1, but a common variation uses a fan of angled lines from each node leading directly to branches, as in Figure 2.

There can be any number of branches of a decision node as long as they represent distinct alternatives. The branches of a chance node must represent a *mutually exclusive* and *collectively exhaustive* set of events. In other words, the branches must all represent distinct events, and the set of branches must cover all possible outcomes at the chance node. Consequently, the probabilities of the branches must sum to exactly 1. There is no universal convention for the order in which the branches of a node appear. Branches of decision nodes are specified in an order that makes clinical sense to the analyst, keeping in mind that this order will determine the order in which the expected value of each branch is displayed after evaluation. When there are many choices, they may be arranged so that groups of similar strategies are adjacent. Branches of chance nodes are usually arranged so that if there are branches

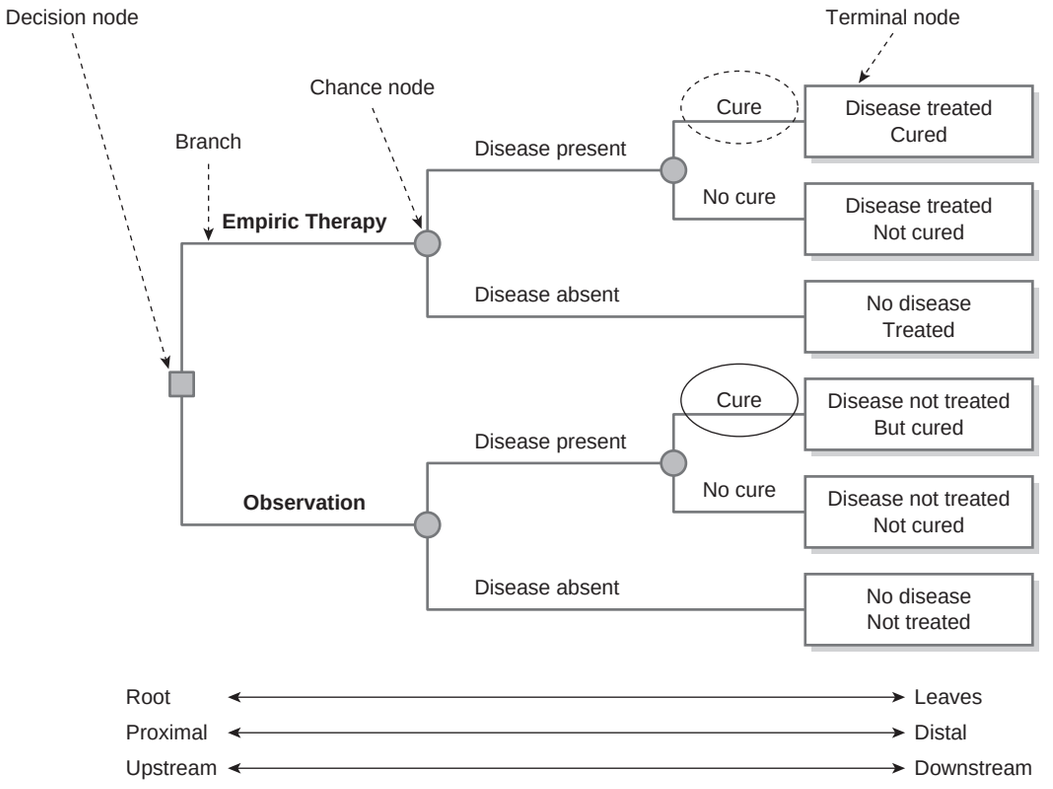


Figure 1 Example decision tree

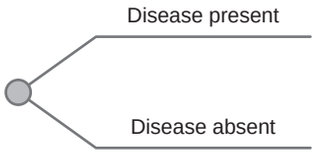


Figure 2 Alternate format for branches

representing occurrence of specific events, they appear on top, and the branch representing nonoccurrence of a specific event is last; however, the order makes no difference to the evaluation when a complete set of probabilities is specified.

Branches are labeled with the names of the choices or events they represent. Terminal nodes may be labeled with a symbolic description of the outcome, as in Figure 1, or with an expression indicating the value or utility of the outcome, as in Figure 3. Branches of chance nodes may also be labeled with the probability of that branch as shown in Figure 4. Note that the probability of "disease present" is the same for both decision branches, but the probability of "cure" is higher for the "empiric treatment" than for "observation." Similarly, utilities can be represented by numbers as shown in Figure 4. The lowest utility is for the worst outcome, which is having the disease, being treated, but not being cured. The highest utility is for the best outcome, which is being observed and not having the disease. Others are intermediate, and their exact values will depend on the specifics of the disease and the treatment. For example, the utility loss due to untreated disease may be worse than the utility loss due to the treatment.

Elements of the model that require assignment of quantitative values (probabilities, utilities, and others) are called model *parameters*.

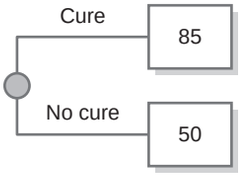


Figure 3 Numerical utilities

Navigation and Orientation

Upstream Versus Downstream

Nodes and branches closer to the root of the tree are said to be *proximal* or *upstream*. Those farther from the root are said to be *distal* or *downstream*. The designation of upstream versus downstream has additional meaning in terms of applying bindings and context-specific variables. A *path* through the tree is defined as the sequence of nodes and branches between any two points in the tree. In general, proximal events occur earlier in time than distal events, but this is not an absolute rule, and nodes at many levels of the tree may represent events that occur simultaneously.

Tree Context

The context of a branch or node in a decision tree is defined as the path from the root of the tree to that branch or node and incorporates all the decisions and consequences that precede them. So, for example, in Figure 1, the context of "cure" indicated by the dotted ellipse is (empiric therapy given, disease present, cured), whereas the context of "cure" indicated by the solid ellipse is (observation, disease present, cured). These often differ in their impact on the probabilities of any downstream events and in determinants of the utilities of the terminal nodes or economic costs.

Variables and Expressions

In the above discussion, probabilities and utilities were expressed either as descriptive labels (Figure 1) or as numerical quantities (Figure 4). It is convenient to represent these quantities symbolically using mathematical expressions composed of variables (Figure 5). There are several reasons for using symbolic variables:

1. To express the model in terms of the meanings of values, allowing alternate values to be specified as input.
2. To facilitate sensitivity analysis by allowing model parameters to vary systematically. For example, the value of pDIS represents the probability of disease and can be varied to determine how the model is affected by changes in disease prevalence.
3. To permit values in specific tree contexts to depend systematically on previous, upstream values. The value of pCure in contexts downstream from "Empiric Treatment" will differ from the value of pCure in contexts downstream from "Observation."
4. To permit the use of subtrees that permit reusing elements of tree structure while allowing values of parameters to change.

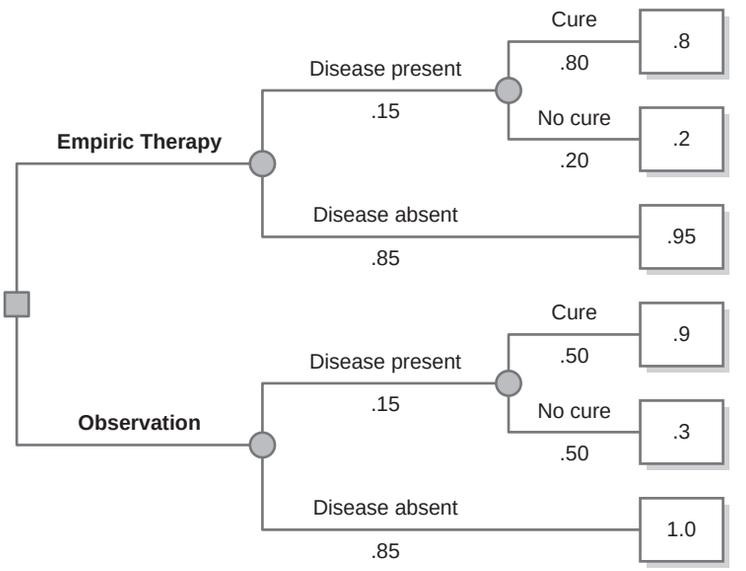


Figure 4 Probabilities on branches

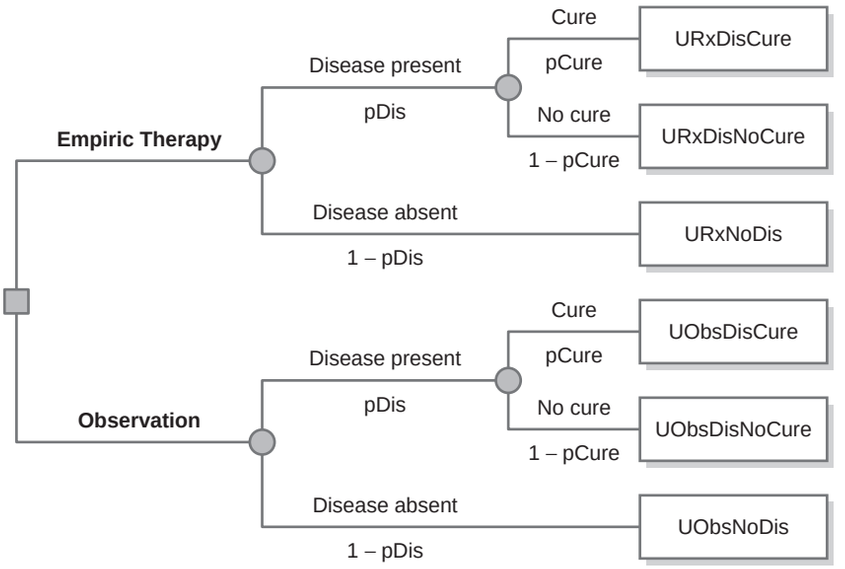


Figure 5 Symbolic probabilities and utilities

5. To express and maintain relationships (*linkage*) between variables during model evaluation using mathematical expressions, thus promoting greater consistency and clarity. This is especially important when parameters are defined functionally, rather than prespecified. For example, the posttest probability of disease may be calculated from the pretest probability in terms of the sensitivity and specificity of a diagnostic test. This not only ensures that the posttest probability is calculated correctly, but linkage of these variables avoids errors during sensitivity analysis. It would be incorrect to vary only the pretest probability of disease or the test sensitivity without also varying the posttest probability. Variables and expressions ensure that these relationships are maintained as models are constructed, modified, and evaluated.
6. To maintain internal statistics of the events that occur at various points in a model.

Expressions

The use of algebraic expressions to express probabilities and utilities permits building them up systematically from more elemental parameters.

More complicated expressions can be constructed in models using a variety of mathematical operators and functions. Application of Bayes's rule is one example. Other examples include the computation of disease prevalence and probabilities in terms of varying factors such as age, and calculating costs as a function of events in specific tree contexts, and employing counting and tracking variables to determine whether and how often specific events occur in a model. Modern decision analysis software implements a full complement of mathematical operators and functions, permitting a great deal of representational power in creating expressions.

The use of variables rather than fixed parameters also facilitates maintenance of the model by enabling the analyst to make lists of parameters. Entire sets of variables can be substituted in the model to represent distinct scenarios or decision contexts.

Utilities

The values of terminal nodes (leaves) of the tree are referred to generically as utilities. The underlying theory and method of assessing and assigning utilities is discussed elsewhere. In practical terms,

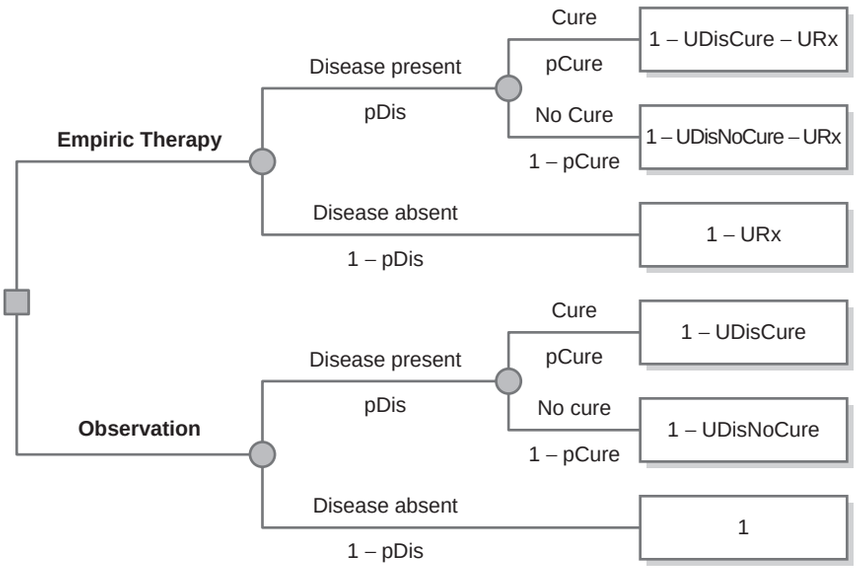


Figure 6 Tree with algebraic expressions

the values of terminal nodes are expressed in terms of health outcomes and financial costs.

The use of algebraic expressions to express utilities is illustrated by the terminal nodes in the tree in Figure 5. There are six unique utilities in this model. While it is feasible to assign each of them a unique variable name as is done in Figure 5, it can be easier to express these utilities in terms of four parameters as in Figure 6. Each utility is calculated by subtracting all applicable disutilities from 1, the value of the "no disease, no treatment" state. When there is a much larger number of terminal nodes, this approach can greatly simplify the assignment of utilities and can greatly reduce the number of parameters the analyst needs to maintain.

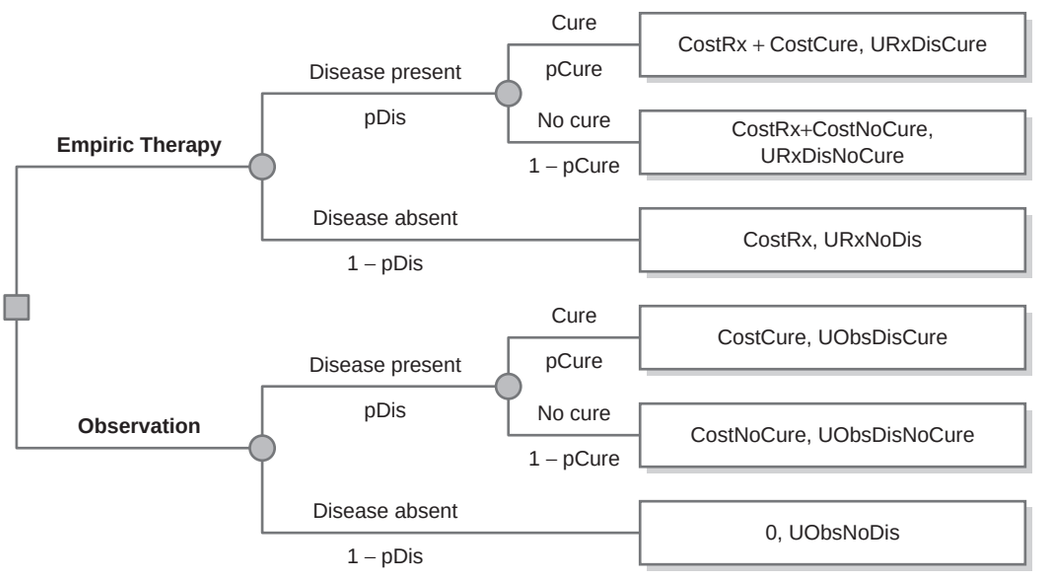


Figure 7 Costs at terminal nodes

Dual Utilities in Cost-Effectiveness Models

In cost-utility analysis, a financial cost must be applied to each path through the tree in addition to its quality measures. Most conveniently, these costs are assigned to the terminal nodes along with the utilities, as shown in Figure 7. Each cost can be calculated as the sum of component costs attributed to treatment, testing, and costs of any of the effects of the disease itself. For the outcome of "observation-disease absent," there are no costs.

Computer Applications for Tree Construction and Evaluation

Several software applications are available for constructing and evaluating decision trees. Using software has many advantages over constructing models manually. By integrating the graphical and mathematical components of the model, such tools greatly speed model construction and minimize errors, allowing much more complicated, clinically realistic models to be considered than would be possible by manual calculation. The ability to load complete sets of variables permits evaluating a model for different scenarios, without manually changing the variables one at a time. Furthermore, the ability to automate the evaluation of models encourages more complete exploration of a model through sensitivity analysis. Graphical representations of models and their results can then be generated, often automatically, for papers and presentations. Models can also be built incrementally and adapted in future applications or work sessions allowing components to be reused, thus providing a systematic means for sharing knowledge and models among analysts.

Frank A. Sonnenberg and C. Gregory Hagerty

See also Cost-Utility Analysis; Decision Trees, Evaluation; Decision Trees: Sensitivity Analysis, Deterministic; Disutility; Expected Utility Theory; Multi-Attribute Utility Theory; Tree Structure, Advanced Techniques

Further Readings

Howard, R. A., & Matheson, J. E. (Eds.). (1984). *The principles and applications of decision analysis, Volume I: Professional collection*. Menlo Park, CA: Strategic Decisions Group.

Information for Authors. (2008). *Medical Decision Making, 28*(1), 157–159.

Pauker, S. G., & Kassirer, J. P. (1981). Clinical decision analysis by personal computer. *Archives of Internal Medicine, 141*(13), 1831.

Sonnenberg, F. A., & Pauker, S. G. (1987). Decision maker: An advanced personal computer tool for clinical decision analysis. *Proceedings of the 11th annual symposium on Computer Applications in Medical Care*. Washington, DC: IEEE Computer Society Press.

Decision Trees, Evaluation

A decision tree is a branched structure used as a tool to support decision making by displaying key elements of the choices among alternatives and the consequences of each choice. This entry uses several examples to illustrate the evaluation of decision trees.

The following examples of surgery versus radiation therapy for Stage 1 (early stage) versus Stage 4 (late stage) disease demonstrate the visual benefits of a decision tree without needing to completely elaborate the tree or to perform any calculations. These examples also explore why the decision analyst in charge of the construction and elaboration of a decision tree needs to be in full control of the key aspects of the decisions that may influence (a) the patient's decision in each stage of this disease process, from Stage 1 (early in life) to Stage 4 (later in life), when the disease is identified early enough in a patient's care, or (b) the patient who presents at the time of diagnosis with Stage 1 disease versus the patient who presents at the time of diagnosis with Stage 2 disease. These examples also explore the difficulties in capturing alternative strategies open to patients in a simple decision tree structure.

Decision Tree for Early Stage 1 Disease

The following decision tree lays out the decision for early Stage 1 disease in a patient for whom physicians believe there are two options open: surgery for Stage 1 disease versus radiation for Stage 1 disease. The decision node (□) represents the decision

for surgery for Stage 1 disease and radiation therapy for Stage 1 disease (Figure 1).

Even at this point in the elaboration of the decision tree in a patient with early Stage 1 disease, many patients would say that this tree is complete. Because both surgery and radiation therapy have outstanding chances for survival in this individual, complications become the focal point. Here, there would be a shift in discussion away from the decision tree representing short-term survival to discussions related to differences in quality of life after surgery and quality of life after radiation therapy. Here, after hearing about the 100% chance of surgical cure of the disease and the quality of life after surgery and the 97% cure rate for radiation therapy and the complications of radiation therapy, the patient may well decide to accept the surgery without any additional exposition of the decision tree or discussion of quality of life. Here, the demonstrated strength of the decision tree is to show the outcomes and features of the comparison between surgery versus radiation therapy for early Stage 1 disease and a simplification for the patient in understanding options.

Laying Out a Decision Tree

The expression "laying out a decision tree" refers to the structuring of a decision tree, with tree growth through the addition of alternatives, outcomes, and their related chances (probabilities) of occurring. We will now structure a set of decision trees to represent a patient's decision problem related to consideration of surgery versus radiation therapy for a progressive disease. Here, we will represent this progression of disease process and disease state in terms of early disease (labeled Stage 1) to the most severe form (labeled Stage 4). We will also consider an intermediate form of progression of this disease (labeled Stage 3). With consideration of these stages, we will examine two strengths of a decision tree: (1) using a decision tree to help visualize the patient's decision problem and (2) using a decision tree in a calculation to determine which of the two treatments would have a survival advantage in a patient with intermediate Stage 3 disease.

In a more complicated patient with Stage 4 disease, the decision tree might take the form as shown in Figure 2.

Here, the complexity of the decision in terms of the questions raised even at the point of the

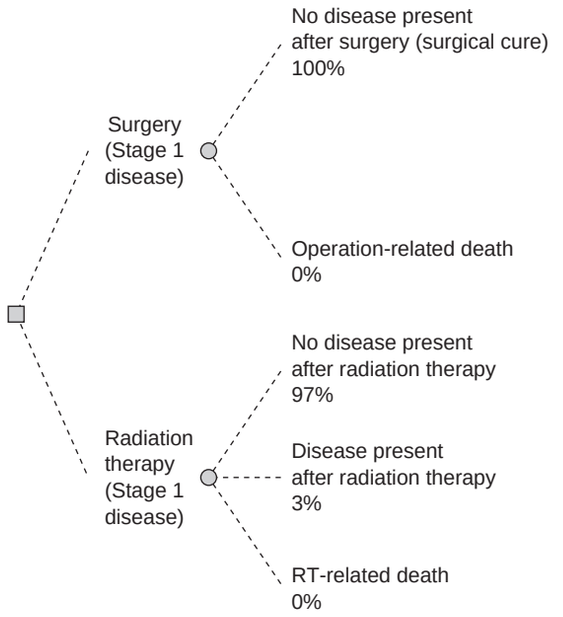


Figure 1 Decision tree for early Stage 1 disease

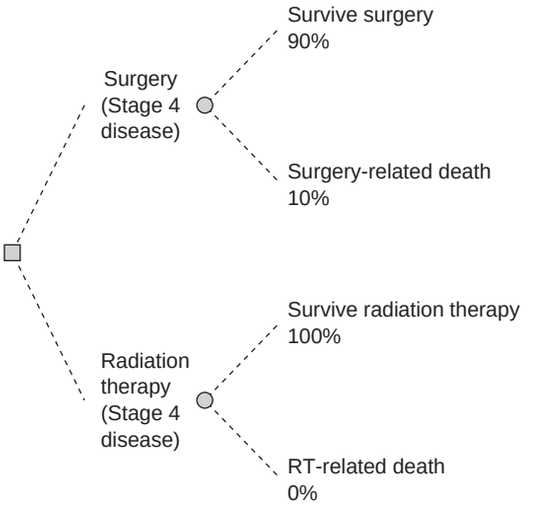


Figure 2 Decision tree for Stage 4 disease

elaboration of the decision tree can be seen. In Stage 4 disease, the patient notes that there is a 10% chance of dying with the surgery and no chance of dying with the radiation therapy and also the suggestion that there is disease still present (which will continue to progress after each therapy).

While one might take this opportunity to further elaborate this tree in terms of the amount of residual disease left after each therapy, another approach would be to see how the disease behaves over time and proceed with an elaboration of a graphical comparison of the 5-year survival curve for surgery and the 5-year survival curve for radiation therapy; the visual comparison of each curve may drive the patient's decision, seeing the difference in 5-year survival at Year 5 after surgery (e.g., 35% of patients sill alive after surgery) in contrast to 5-year survival at Year 5 after radiation therapy (e.g., 25% of patients still alive after radiation therapy).

Here, the patient's decision may be driven by the chance of 5-year survival. However, the 5-year survival curve comparison would also demonstrate the crossover point, which is the point where the shorter-term benefit of survival after radiation therapy is lost and the longer-term benefit of survival after surgery is realized and continues to 5 years, where there would be a 10% 5-year survival benefit at Year 5 with surgery as opposed to radiation therapy (35% – 25% = 10%). Here, a 5-year survival curve comparison between surgery and radiation therapy could be used along with the decision tree to provide the patient fuller information about the decision over time, from Year 0 to Year 5, a time 5 years after the initial treatment.

A decision tree for Stage 3 disease may be more complex because both therapies (a) may not have a clearly defined peer-reviewed medical literature, in contrast to Stage 1 and Stage 4 disease, and (b) there will be more questions about what is going on with survival and quality of life during the time period from Stage 3 to Stage 4 disease (Figure 3).

Given that phase 1 of this decision tree shows the same rates of cure (0%) and disease presence (100%) in this Stage 3 disease, one can simplify the construction by eliminating the first part of the decision tree and move on to Phase 2.

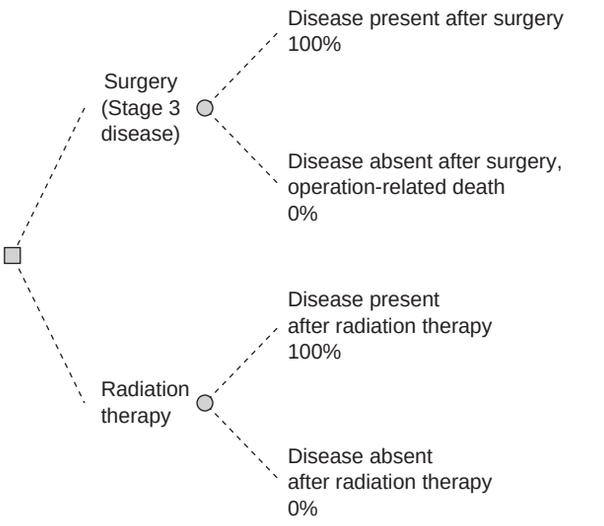


Figure 3 Decision tree for Stage 3 disease, Phase 1

Calculating a Decision Tree

Once a decision tree is laid out, it can be evaluated or calculated. One needs to recognize that for many situations where a specific mathematical calculation is not needed, the process of laying out the decision tree (reviewing the peer-reviewed medical scientific literatures, acquiring expert opinion, eliciting an individual's preferences regarding outcomes, and allowing individuals an opportunity to see the risks and benefits associated with alternatives) is a powerful visual procedure in its own right, without the need for any specific mathematical calculation. This laying out of a decision tree may be very useful in areas of consent and informed consent in medicine and in information disclosure in economic and legal contexts. This said, this entry now discusses how a decision tree is evaluated.

Decision Tree Evaluation

The term *decision tree evaluation* usually refers to the calculation of a decision tree. A decision tree is calculated by *folding back* (averaging out or rolling back) the decision tree.

Referring to the above example, in Phase 2 of the decision tree, we go to the peer-reviewed medical literature and find that there are no studies on Stage 3 disease, so we go to local experts (the physicians who actually are doing the surgery and

performing the radiotherapy). These local experts may rely on their own data collection on the patients that have been treated in both departments (surgery and radiation therapy), and we need to rely on these data.

From these data, derived from the database in both departments, we see that patients with Stage 3 disease who had surgery on average have 15 years of life expectancy, or 15 life years (LY), and that patients with Stage 3 disease who underwent radiation therapy have 10 years of life expectancy, or 10 LY (Figure 4).

Folding Back the Decision Tree

Once reliable baseline probabilities and outcome values are attained from the peer-reviewed medical scientific literature, expert opinion, and patient preferences (through the elicitation of patient preferences from a standard gamble), the tree is ready to be folded back or rolled back. Theoretically, the expression *folding back* (averaging out or rolling back) the decision tree is an overall calculation that is executed at a particular point in time, when all outcomes are enumerated and listed, all probabilities have been gathered, and all preferences have been elicited. However, with any expressions where the term *all* is used, as in the above expressions, including *all outcomes, all probabilities,* and *all preferences,* caveats are in order and must be examined.

We will now perform the calculations based on the data set obtained from the hospital that is providing the patient's care (Figures 5, 6, and 7).

Based on this decision tree, surgery (13 years – 10 years = 3 years) would offer the patient a better survival than radiation therapy, and for the patient whose primary preference is survival, surgery would be the dominant choice given the above numbers.

Pruning

The above example of radiation therapy versus surgery for early Stage 1 disease did not consider chemotherapy as one of the alternative treatments. Here, the medical-scientific point may be that this early-stage disease does not respond well to existing chemotherapies. And even if chemotherapy did exist for Stage 1 disease, the patient may not want to consider any therapeutic options. In both cases, the chemotherapy alternative was pruned away from the decision structure in Tree 1 above.

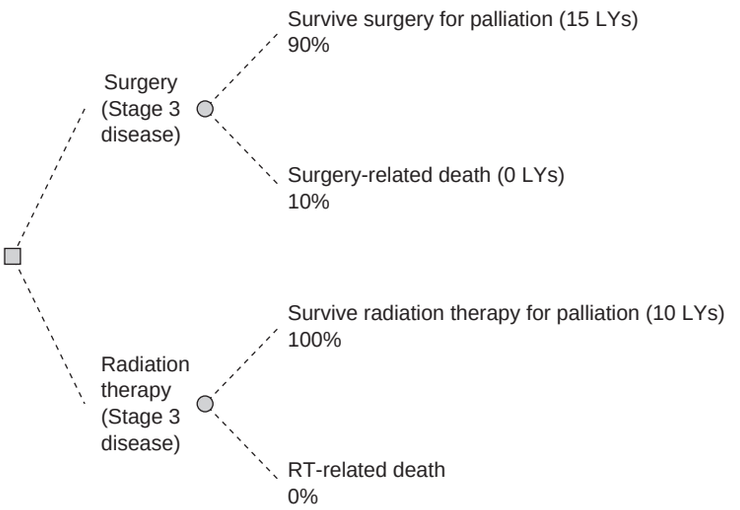


Figure 4 Decision tree for Stage 3 disease, Phase 2

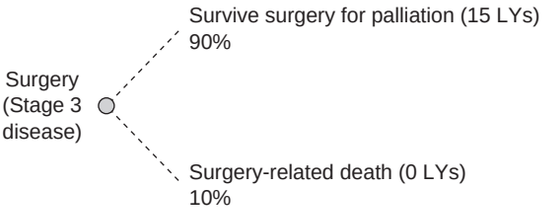


Figure 5 Calculation of the average life years for surgery

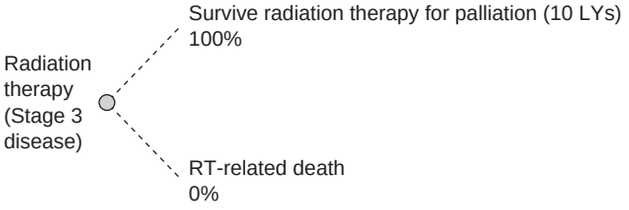


Figure 6 Calculation of the average life years for radiation therapy

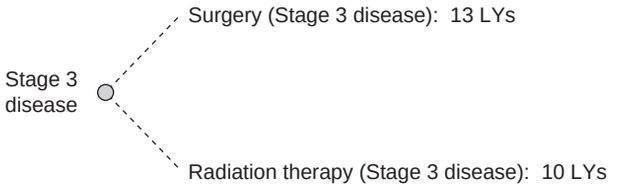


Figure 7 Calculations of the average life years for surgery and radiation therapy

Future Treatment Options

In later stages of the disease under consideration after either surgery or radiation therapy (or both surgery and radiation therapy) have been exhausted, there may be a role for palliative chemotherapy, that is, therapy intended to palliate, not cure, the disease. Early surgery, for example, may well have been intended to offer an option for cure for the patient, based on review of the peer-reviewed medical scientific literature. However, if a cure was not secured and the disease returned, radiation therapy could be offered. And when the disease recurs after both surgery and radiation therapy, there may be a role for palliative chemotherapy in a patient whose main goal is to survive as long as possible.

Certain patients may want to see how surgery followed by chemotherapy versus radiation therapy followed by chemotherapy look in a decision tree. In Figure 8, one can see that when chemotherapy for palliation is considered as an option after surgery and after radiation therapy as requested by the patient, the surgery alternative becomes stronger in terms of survival over radiation therapy because the palliative chemotherapy after surgery provides a longer survival than palliative chemotherapy after radiation therapy. Thus, the addition of a future treatment option (palliative chemotherapy) may change a patient's mind toward surgery as an initial therapy after seeing the tree in Figure 8.

Addition of a Wait-and-See Alternative

It is important to recognize that decision trees are not optimal for structuring all types of decisions. One of the key alternatives in patient care is the wait-and-see alternative (or watchful waiting). Here, no intervention is made in a disease state. Rather, the patient elects to wait and see how his or her disease acts over time and then decides to act at the time when there is an increase in tumor activity noted on the basis of a worsening of symptoms, a change in physical examination suggesting an increase in growth of tumor mass, a change in laboratory testing measurement, or a change in biopsy results suggesting a move from a lower-stage tumor to one that is more aggressive.

In this case, one can construct a decision tree that considers surgery for palliation versus radiation therapy for palliation versus watchful waiting (wait and see) and then palliative chemotherapy for survival for Stage 3–4 disease in a patient whose main goal is to live as long as possible regardless of quality of life. Given the added emphasis that is being placed by oncologists on the offering of palliative care options to patients with oncologic diseases, one can add a wait-and-see palliative care decision option with palliative chemotherapy to the tree in Figure 8, creating the tree shown in Figure 9.

Figure 9 illustrates the pruning of the tree in Figure 8 and the elimination of branches from the tree structure.

Figure 10 shows the pruning of the tree in Figure 9 and the LY expectancy with the three approaches that can be offered to the patient. Figure 11 shows the comparison of the LY expectancy of three

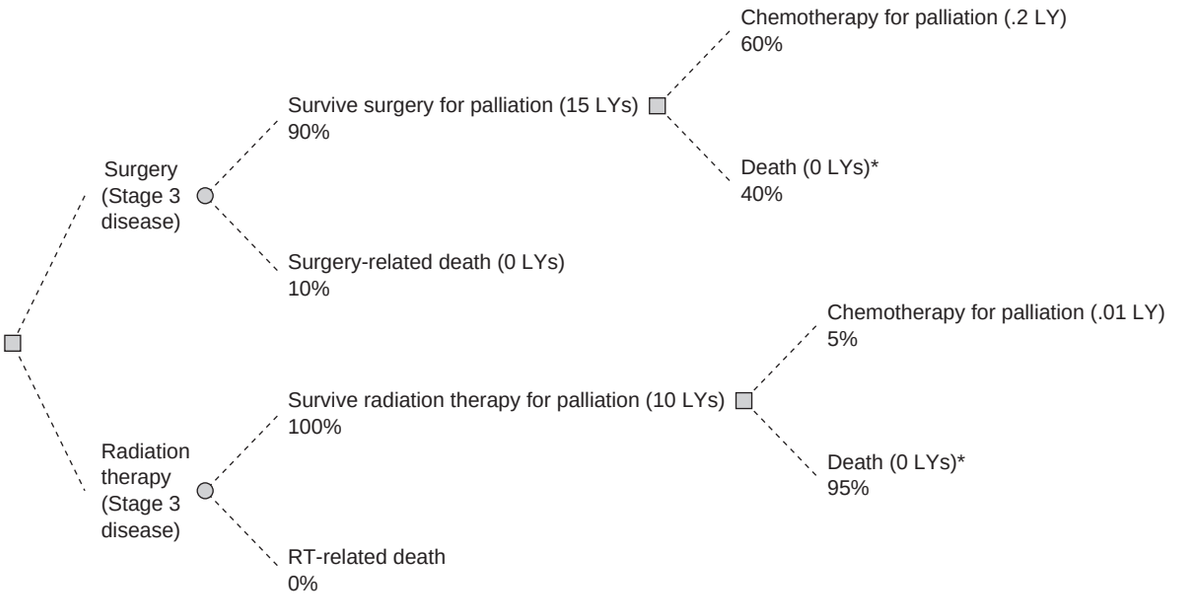


Figure 8 Decision tree with the addition of palliative chemotherapy

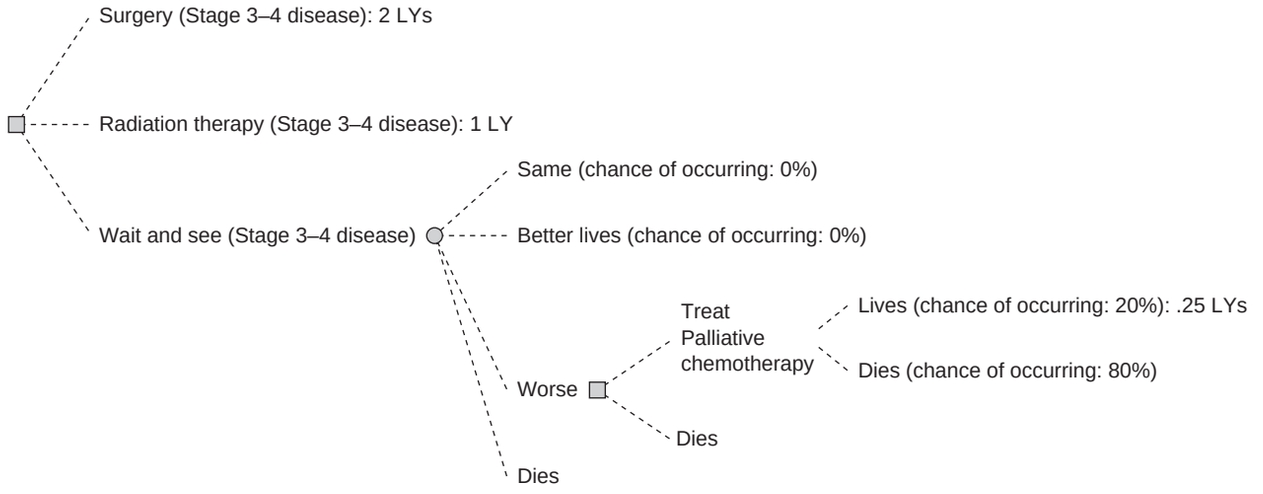


Figure 9 Decision tree with addition of wait-and-see alternative

treatments—palliative surgery versus palliative radiotherapy versus wait-and-see and then treat with palliative chemotherapy—for a patient who is focused on survival rather than quality of life in his or her decision making.

The problem with the trees in Figures 9 through 11 is that by displaying wait-and-see as simply another alternative similar to surgery and radiation therapy, there is no reflection of the fact that there may be vast differences in time that is accorded to a wait-and-see state, such that a patient may spend various times (from days, weeks, months to years) in a wait-and-see state, and this time variability is not reflected in the basic decision tree structure.

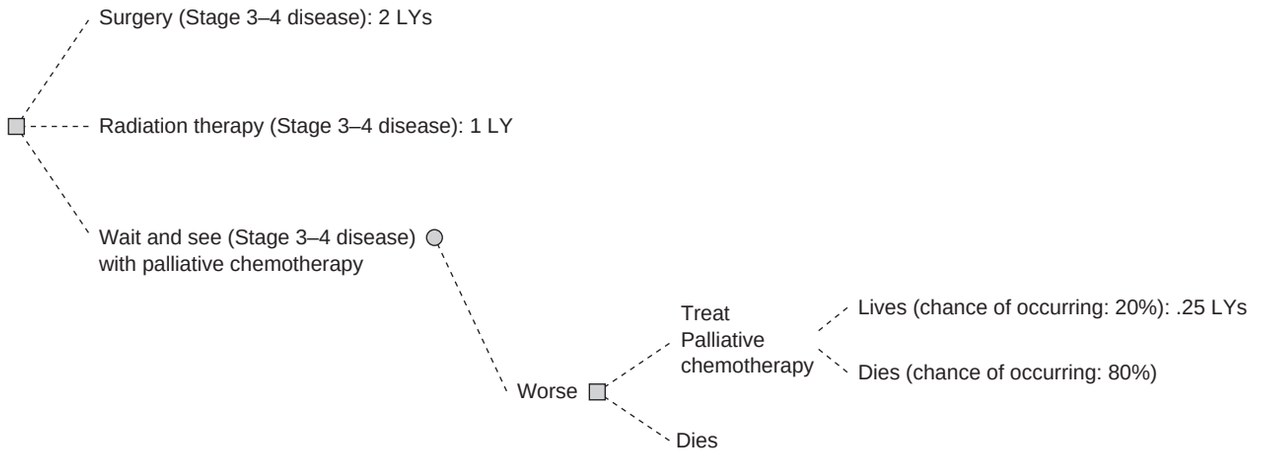


Figure 10 Decision tree showing life years expectancy of three treatments

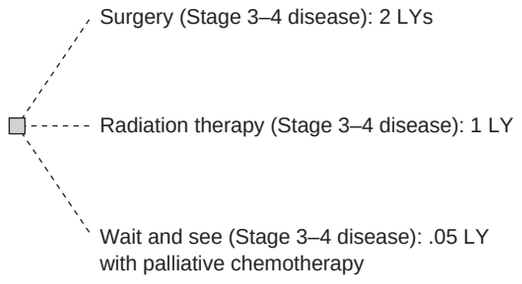


Figure 11 Decision tree showing the years of life expectancy for three treatments

Here, there is a call for procedures and analyses, such as Markov models, to portray wait-and-see states, to allow for a more real and dynamic interpretation of wait-and-see attitudes and alternatives in patient decision making.

Caveats

The fact of the matter is that in any decision analysis, all the outcomes, all the probabilities, and all the preferences will not have been derived and enumerated as extensively as required by physicians and providers. And there will be missing, problematic, or illusory numbers appearing in the tree as considered by physicians and other providers. Decision scientists may examine the same tree and find it to be acceptable. Viewed from another perspective, a physician may view a decision tree and see all the unanswered questions that exist regarding outcomes, probabilities, and preferences and say that more work is needed before the tree can be folded back.

Decision to Stop Building a Decision Tree

Folding back a decision tree is a procedure that can begin only when one stops building a decision tree or pauses in the building of a decision tree. Some authors will say that one can stop building a decision tree at any point, with one provision: that one is able to describe the terminal outcome at each endpoint in such a way that the unmodeled future from that endpoint onward in time can be accounted for and approximated. Other authors will argue that one can elaborate the tree to the point where one feels comfortable with the approximations made and their analytical implications for the immediate choice of action that needs to be made.

Each attempt at approximation will be made by those who will agree with the approximation or those who will accept the approximation to see where it leads (how it performs), while others will reject the approximation as outlandish from its outset of construction to its conclusion. At some point, some will be happy with the probabilities as gathered; others will be insistent that the peer-reviewed medical literature, expert opinion, and patient preference must be more thoroughly searched for. The only limitation on continued folding back of a decision tree in the case of an

individual's decision making is the estimated length of time a decision can be postponed or delayed without significant impact on the patient's survival and quality of life.

Dennis J. Mazur

See also Decision Trees, Evaluation With Monte Carlo; Expected Utility Theory; Markov Models

Further Readings

Audrey, S., Abel, J., Blazeby, J. M., Falk, S., & Campbell, R. (2008). What oncologists tell patients about survival benefits of palliative chemotherapy and implications for informed consent: Qualitative study. *British Medical Journal, 337,* a752.

Pratt, J. W., Raiffa, H., & Schaifer, R. (1995). *Introduction to statistical decision theory*. Cambridge: MIT Press.

Pratt, J. W., & Schlaifer, R. (1988). On the interpretation and observation of laws. *Journal of Econometrics, 39,* 23–52.

Raiffa, H. (1968). *Decision analysis: Introductory lectures on choices under uncertainty*. Reading, MA: Addison-Wesley.

Raiffa, H., & Schaifer, R. (2000). *Applied statistical decision theory*. New York: Wiley. (Original work published 1961)

Savage, L. J. (1972). *The foundations of statistics* (2nd rev. ed.). New York: Dover. (Original work published 1954)

Decision Trees, Evaluation With Monte Carlo

Monte Carlo simulations are based on Monte Carlo methods. *Monte Carlo method* refers to a method of solving sets of equations using an algorithm dependent on repeated random sampling. The Monte Carlo method is used in the process of simulating (approximating) a system. Monte Carlo methods are computational algorithms that rely on repeated random sampling to compute their results. *Monte Carlo simulation* involves repeated random sampling from input distributions and subsequent calculation of a set of sample values for the output distributions with the repeating of the process over several iterations.

The term *Monte Carlo method* was used in the 1940s in the more rapid solving of equations and algorithms possible on the first electronic digital computer, the ENIAC computer. The term was used by Nicholas Metropolis and Stanislaw Ulam in 1949. Metropolis attributed the initial insights on the use of the method to Enrico Fermi. The reference to the gaming tables of Monte Carlo, Monaco, shows the importance of randomness and chance events in the entities that are being simulated.

Today, the major uses of the Monte Carlo method involve examining real-life phenomena that need to be approximated or simulated rather than tested in the sense of real-world testing of scientific hypotheses. In testing, for example, in research on humans, there would be the tasks of developing scientific protocols, collecting data in trials, and conducting research on humans that in turn would need to be derived in conjunction with existing federal laws and approved by an institutional review board. If alternative strategies can be effectively modeled, sparing humans lives and costs, then real-world testing may not be needed as extensively as it is needed today.

Use of Monte Carlo simulation has expanded exponentially into many areas where random behavior, uncertainty, and chance events characterize the system being simulated in a diverse range of real-world endeavors: economics; finance (interest rates and stock prices); business (inventory, staffing needs, and office tasks); the sciences; and medical decision making with economic implications (e.g., impact of colonoscopic referral for small and diminutive polyps detected on CT colonography screening).

Monte Carlo Simulation

Monte Carlo simulation selects value variables at random in the attempt at simulating a real-life situation whose outcome needs to be estimated or predicted. The variables of interest will have a known range or at least a range that can be estimated.

A variable may be uncertain, but if that variable is known to have a range of values (or estimated to have a range of possible values), this range of possible values can define a probability distribution. A simulation calculates multiple scenarios by repeatedly sampling values from the probability distributions for the uncertain variables.

Monte Carlo simulations depend on the computational tools available at the time a simulation is run. Simulations run during the days of the Manhattan Project in the 1940s are dwarfed by computations performed on a laptop computer today.

Deterministic Models Versus Iterative Models

When a model is created with a spreadsheet, one has a certain number of input parameters and a few equations that use those inputs to give a set of outputs (or response variables). This type of model is usually termed *deterministic* in that one gets the same result no matter how many times one performs a recalculation.

A basic decision tree is an example of a deterministic model. Although inputs may differ in terms of the chance success of a surgery versus a radiotherapy intervention on a cancer at different medical centers based on the specific patient data found in one medical center versus another medical center, when one calculates the model based on the same data, everyone who performs the calculation should come up with the same result.

One sense of Monte Carlo simulation is as an *iterative model* for evaluating a deterministic model. Here, the Monte Carlo simulation uses a set of random numbers as the input numbers for the model.

The Monte Carlo method is one of many methods that can be used to understand how (a) random variation, (b) lack of knowledge, or (c) error rates affect the sensitivity, performance, or reliability of the system being modeled. Monte Carlo simulation is categorized as a sampling method because the inputs are randomly generated from probability distributions used to simulate the process of sampling from an actual population. Hence, there must be a choice of a distribution for the inputs that most closely matches the data that is available on the question about which an answer is sought. The data generated from the simulation can be represented as probability distributions and in turn converted to confidence intervals.

Input and Output Variables

A simulation begins with the development of a model of a system that one wishes to test. The model comprises mathematical equations describing relationships between one or more input (independent) variables and one or more output (dependent) variables. By selecting specific values for the input variables, corresponding output values may be calculated for the output variables. In this manner, one can determine how the system, to the extent that it is accurately represented by the model, will respond to various situations represented by the input values. Note that, as used herein, a "system" may comprise virtually anything that can be represented by an appropriately constructed mathematical model, for example, the impact of referral to a colonoscopist for direct visualization of small and diminutive polyps detected on indirect imaging, for example, visualization on a CT scan.

In Monte Carlo simulations, a range of plausible input values is designated for each input variable. Likewise, a distribution for each input variable (i.e., a probability distribution function) is also designated. Thereafter, the Monte Carlo simulation generates random inputs for each input variable based on the designated range of values and distributions for the corresponding variables. The random input values are then used to calculate corresponding output values. This process is repeated many times, typically numbering in the hundreds, thousands, ten thousands, or more, and is used to create statistically meaningful distributions of one or more of the output variables. In this manner, the analyst performing the Monte Carlo simulation can develop insight into how the model will perform under certain sets of assumed input conditions. The analyst needs to have intimate knowledge of the underlying system and its simulation model.

Incorporation of Monte Carlo Simulation Into a Decision Tree

The incorporation of Monte Carlo simulation into a decision tree allows examination of "probability distributions" rather than "single expected values" or "ranges of expected values." Some describe Monte Carlo simulation as replacing the analysis of point estimates with fuzzy values (or better, ranges of fuzzy values).

For example, monetary values can be replaced with normal distribution functions (e.g., a normal

distribution with a specified mean and a standard deviation). One can present the distribution of results for the expected value after a Monte Carlo simulation with 10 trials, 100 trials, 1,000 trials, 10,000 trials, and so on.

Probability Distributions

The probability distributions selected must describe the range of likely values for each parameter. This is a selection problem for the analyst, who must be able to represent the best probability distribution for a particular setting.

Probability distributions may be of standard form (normal or lognormal distributions) or may have empirical forms (rectangular, triangular, among others). Here, an analyst can start with the historical data of the parameters being considered and attempt a "best-fit" approach of a distribution to the historical data.

The parameters of the distribution (mean and standard deviation in the case of normal distributions) may be based on data derived from (a) the peer-reviewed medical scientific literature (if present and available), (b) historical data as is contained in the databases of surgery departments or radiation therapy departments in medical centers (if accessible), or (c) the experience of experts.

In absence of specific knowledge about the form of a distribution, assumptions are made about what the distributions should look like, and certain distributions may be selected, for example, normal or lognormal distributions. Some properties may need to be bounded, as it may not be possible to have specific properties outside of specific ranges.

Statistics Obtained From a Monte Carlo Simulation

The statistics obtained from any simulation are estimates of the population parameters; the exact values of the population parameters will never be known. The assumption is that as the number of iterations increases, the probability that an estimate of a population parameter is within a specific amount of the actual population also increases.

The analyst himself or herself selects the number of iterations, the accuracy required from the procedure. The analyst's assumptions regarding number of iterations and accuracy required in a task are issues that can be argued about and taken up with the analyst. Nonanalyst-related impacts on Monte Carlo simulation include the complexity of the initial problem being modeled and cost of the procedure (e.g., analyst's time, computing time).

Simulation Analysis

In simulation analysis, a decision tree is "rolled forward." A bank of data is generated by the simulation analysis that, if interpreted correctly, can give a probabilistic picture of the consequences of a decision strategy.

Example

Let us take a medical example using a Monte Carlo simulation. A patient with adult respiratory distress syndrome (ARDS) has a diffuse injury to lung tissue due to diffuse damage to the smallest air sacs of the lungs (alveoli) in the absence of congestive heart failure. (The fluid in the lung of the ARDS patient is not due to heart failure.) ARDS is a serious medical condition of acute onset with infiltrates found in both lungs on chest X-ray and has as its origin a diverse array of predisposing conditions causing fluid buildup in the lungs, including direct pulmonary injury (lung infection or aspiration of materials into the lung) and indirect injury (blood infection, pancreatitis, moderate to severe trauma). Here, a patient with ARDS may undergo care in the following states:

- Patient intubated in the intensive care unit
- Patient nonintubated on a hospital ward
- Patient in offsite long-term care
  - In an offsite long-term care facility that accepts respirators
  - In an offsite long-term care facility that does not accept respirators
- Patient in home care

The patient will not stay in any one state but will transition between states (home care, long-term care facility without respirator, long-term care facility with respirator, medical ward extubated, intensive care unit intubated until death) with time at home decreasing and time in all care states increasing until death.

In this setting, neuromuscular blocking (NMB) has potential benefits (NMB drugs may facilitate mechanical ventilation and improve oxygenation) and potential risks (NMB drugs may result in prolonged recovery of neuromuscular function and acute quadriplegic myopathy syndrome [AQMS]). The researchers attempted to answer the question whether a reduction in intubation time of 6 hours and/or a reduction in the incidence of AQMS from 25% to 21% provide enough benefit to justify an NMB drug with an additional expenditure of $267 (the difference in acquisition cost between a generic and brand name NMB drug, the neuromuscular blocker). They performed this task by (a) constructing a Markov computer simulation model of the economics of NMB in patients with ARDS (b) using Monte Carlo simulation to conduct a probabilistic sensitivity analysis considering uncertainties in all probabilities, utilities, and costs.

If one attempted to model ARDS in terms of decision trees (a deterministic approach), these trees and their analysis would be limited in their abilities to model the events of ARDS because of the need to model "multiple times" in the care of a patient who transitions from state to state, moving from one extreme, the most severe state with most intensive care (patient intubated), to the least severe state, the stable state of having resolved ARDS and being now with minimal care at home.

The researchers, Macario and colleagues, used probabilistic sensitivity analysis to consider uncertainties in all probabilities, utilities, and costs simultaneously. In their model, mean values of the net monetary benefit were calculated for results of $N = 10{,}000$ Monte Carlo simulations, where triangular distributions were used for parameter values, with the mode being the case and the 5th and 95th percentiles of the lower and upper limits of the ranges reported. They reported all costs in year 2004 U.S. dollars and discounted all future costs and quality-adjusted life years at 3% per annum.

This report of the results of a Monte Carlo simulation followed the recommendation of Doubilet and colleagues that the following results be recorded:

- The mean and standard deviation of the expected utility of each strategy
- The frequency with which each strategy is optimal
- The frequency with which each strategy "buys" or "costs" a specified amount of utility relative to the remaining strategies

Macario and colleagues reported the results of the simulation by noting that the net monetary benefit was positive for 50% of simulations with a ceiling ratio of $1,000 versus 51% if the ceiling ratio was increased to $100,000. They argued that lack of sensitivity was caused by the mean changes in quality-adjusted life year (QALY) and cost being small relative to their standard deviations.

Their Markov model noted that the following variables had the largest influence on their results: (a) probability from ICU intubated to death, (b) probability from ICU intubated to extubated, and (c) probability from ICU extubated to ward. The model showed that the better the patients do overall, the larger the net monetary benefit of a drug that reduces AQMS and/or intubation times.

First-Order and Second-Order Uncertainty

There are two categories of uncertainty related to the ARDS model above and similar models. *First-order uncertainty* refers to variability among individuals. *Second-order uncertainty* refers to parameter uncertainty. First-order uncertainty can be captured in the phrase *overall variability between patients* and is reflected in standard deviation associated with a mean value. Second-order uncertainty is *parameter uncertainty,* where uncertainty exists in mean parameter values and is reflected in standard error of the mean.

To understand the uncertainty within a model, Monte Carlo simulation techniques can be applied using both first-order and second-order simulations. A first-order simulation is also called a *run of a random trial,* a *microsimulation,* or a *random walk*. A first-order simulation is performed by running each patient in the hypothetical cohort through the model, one at a time. First-order simulation trials can be used to model the variability in individual outcomes. First-order simulation reflects what can be described as first-order uncertainty involving the variability among individuals.

Variability between individuals can be modeled using first-order Monte Carlo microsimulation. But what about questions of second-order uncertainty? In practice, the most commonly used measures are

those that are based on formulating uncertainty in the model inputs by a joint probability distribution and then analyzing the induced uncertainty in outputs, an approach which is known as probabilistic sensitivity analysis. Probabilistic sensitivity analysis is more readily applied to an aggregate cohort of patients.

Probabilistic Sensitivity Analysis

Probabilistic sensitivity analysis uses a probabilistic approach where all the input parameters are considered random variables, endowed with known prior probability distributions. Why is probabilistic sensitivity analysis needed? First, there are numerous parameters in decision models. Second, each parameter has an estimated uncertainty. There is a need to "propagate" parameter uncertainty. The use of an analysis approach to estimate the effect of uncertainties on model prediction is referred to as *uncertainty propagation*. Second-order simulation—as an analysis approach—relies on sampling parameter values to estimate the effect of uncertainties on model prediction.

Probabilistic sensitivity analysis requires one to identify sources of parameter uncertainty, to characterize uncertain parameters as probability distributions, and to propagate uncertainty through the model using Monte Carlo simulation.

When applied to groups of patients rather than individual patients, Halpern and colleagues note that implementing a probabilistic sensitivity analysis may lead to misleading or improper conclusions. The authors argue that the practice of combining first- and second-order simulations when modeling the outcome for a group of more than one patient can yield an error in marginal distribution, thus underrepresenting the second-order uncertainty in the simulation. It may also distort the shape (symmetry and extent of the tails) in any simulated distribution, resulting in premature or incorrect conclusions of superiority of one strategy over its alternatives being modeled.

The complexity of Monte Carlo simulations—how they are conducted and how they are interpreted—is still being unraveled in relation to first- and second-order effects.

Dennis J. Mazur

See also Decision Trees, Evaluation; Expected Utility Theory; Markov Models; Quality-Adjusted Life Years (QALYs)

Further Readings

Doubilet, P., Begg, C. B., Weinstein, M. C., Braun, P., & McNeil, B. J. (1985). Probabilistic sensitivity analysis using Monte Carlo simulation. A practical approach. *Medical Decision Making, 5,* 157–177.

Eckhardt, R. (1987). Stan Ulam, John von Neumann, and the Monte Carlo method. *Los Alamos Science* (Special issue), *15,* 131–137.

Halpern, E. F., Weinstein, M. C., Hunink, M. G., & Gazelle, G. S. (2000). Representing both first- and second-order uncertainties by Monte Carlo simulation for groups of patients. *Medical Decision Making, 20,* 314–322.

Macario, A., Chow, J. L., & Dexter, F. (2006). A Markov computer simulation model of the economics of neuromuscular blockade in patients with acute respiratory distress syndrome. *BMC Medical Informatics and Decision Making, 15*(6), 15.

Metropolis, N., & Ulam, S. (1949). The Monte Carlo method. *Journal of the American Statistical Association, 44*(247), 335–341.

Pickhardt, P. J., Hassan, C., Laghi, A., Zullo, A., Kim, D. H., Iafrate, F., et al. (2008). Small and diminutive polyps detected at screening CT colonography: A decision analysis for referral to colonoscopy. *American Journal of Roentgenology, 190,* 136–144.

Student. (1908). Probable error of a correlation coefficient. *Biometrika, 6,* 302–310.

Student. (1908). The probable error of a mean. *Biometrika, 6,* 1–25.

Sullivan, P. W., Arant, T. W., Ellis, S. L., & Ulrich, H. (2006). The cost effectiveness of anticoagulation management services for patients with atrial fibrillation and at high risk of stroke in the US. *PharmacoEconomics, 24,* 1021–1033.

Decision Trees: Sensitivity Analysis, Basic and Probabilistic

Sensitivity analysis is defined as systematically varying one or more parameters in a decision model over a specified range and recalculating the

expected utility of the model for each value. There are four reasons to employ sensitivity analysis:

1. to determine the effect of reasonable variations in the estimates of parameters on the results of the analysis;
2. to determine which variables are most critical to the analysis—and, therefore, may justify further efforts to estimate them more precisely;
3. to determine what the analysis would recommend for various *scenarios* (combinations of parameters); and
4. to explore the model for bugs or anomalies.

The best estimate of the value of each parameter in a model is called the *baseline* value. When all parameters are at their baseline values, the model is said to be the *base case* model. Without sensitivity analysis, one can say only that the results of the analysis apply to the base case model. When small changes in a parameter affect the recommended choice or cause significant changes in the results, the analysis is said to be *sensitive* to that parameter. Some changes in parameters affect the decision only if they are combined with specific changes in one or more other variables. Therefore, complete sensitivity analysis must examine more than one variable at a time. Sensitivity analyses may examine any number of variables at a time, although in practical terms, only one-, two-, or three-way sensitivity analyses can be illustrated graphically. Examination of more than three variables simultaneously requires probabilistic sensitivity analysis (PSA).

One-Way Sensitivity Analysis

The decision tree shown in Figure 1 models a simple decision between Observation and Treatment. The prior probability of disease (.3) and the utilities of each combination of treatment and disease state are indicated. The expected utilities are 21.98 for the Treatment strategy and 22.38 for the Observation strategy. Figure 2 shows a one-way sensitivity analysis on the probability of disease. When pDis = 0, the difference between the Observe and Treatment strategies represents the "cost" of treatment (in this case the morbidity cost). When pDis = 1, the difference in utility between the Observe and Treatment strategies represents the net benefit of treatment. It is intuitive that when pDis = 0, Observation must be the preferred strategy and when pDis = 1, Treatment must be the preferred strategy. If not, then the treatment is worse than the disease and the analysis makes no sense.

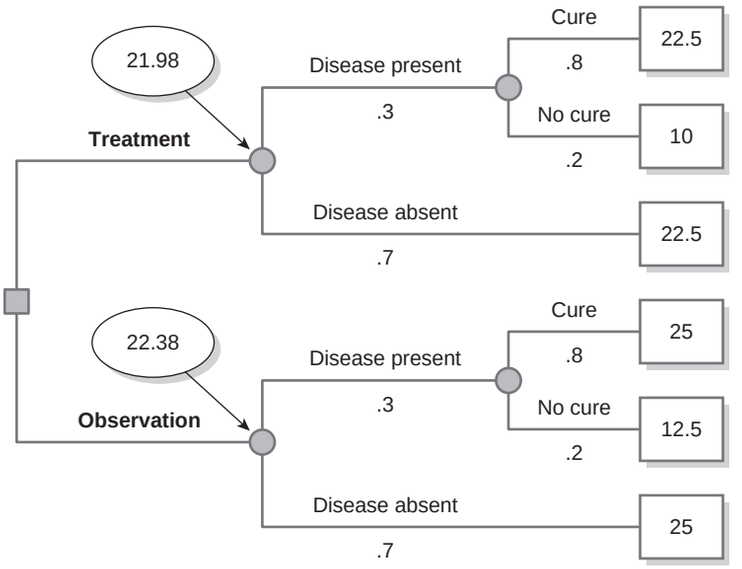


Figure 1 Empiric therapy versus observation decision tree

Threshold Approach

It is apparent from Figure 2 that the lines representing the two strategies cross at a point. This point is called a *threshold* because it represents the value of the independent variable above or below which the preferred strategy changes. In the case of the Observe/Treatment choice, the threshold is referred to as the *treatment threshold*. Figure 3 shows a simple geometric way of calculating the treatment threshold. Assuming that the expected utilities of both strategies are straight lines (i.e., they vary linearly with the independent variable), the combination of the expected utility lines forms a set of similar triangles. The "height" of the left-ward triangle is the threshold value. The width of the base of the left triangle is the cost of treatment. The width of the base of the right triangle is the benefit of treatment. For similar triangles, the ratios of the heights are equal to the ratios of the bases:

$$\frac{C}{B} = \frac{t}{1-t}.$$

Solving for t,

$$t = \frac{C}{C+B} \quad \text{or} \quad \frac{1}{1+\frac{B}{C}}.$$

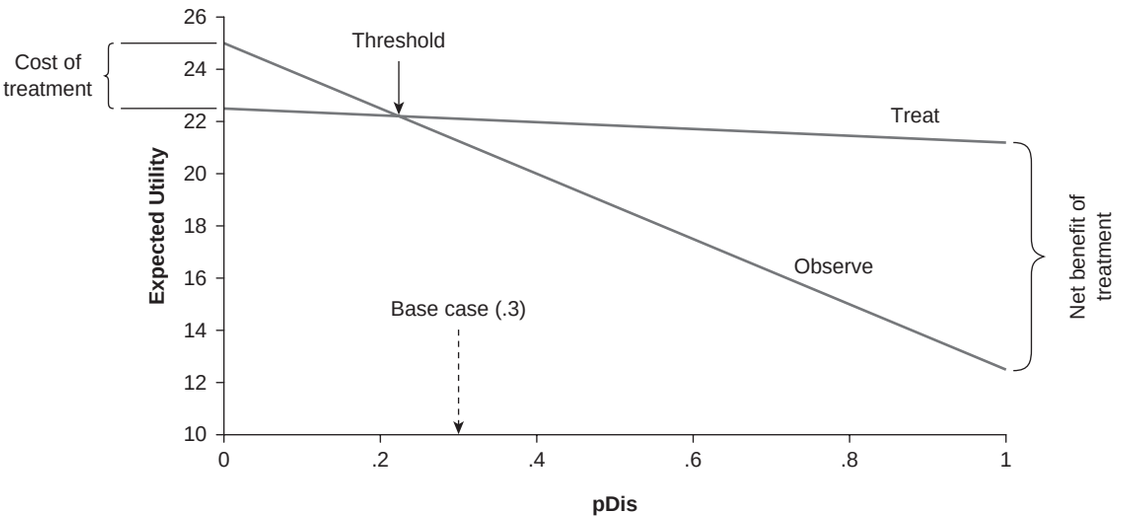


Figure 2 One-way sensitivity analysis on probability of disease

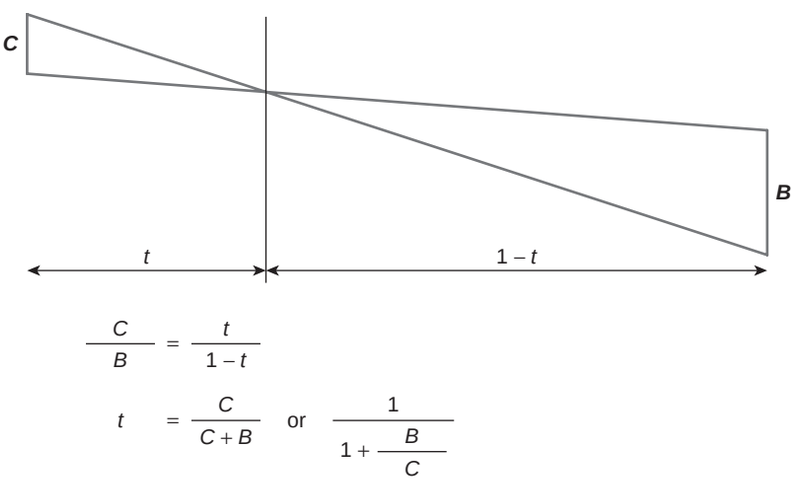


Figure 3 Geometric calculation of treatment threshold

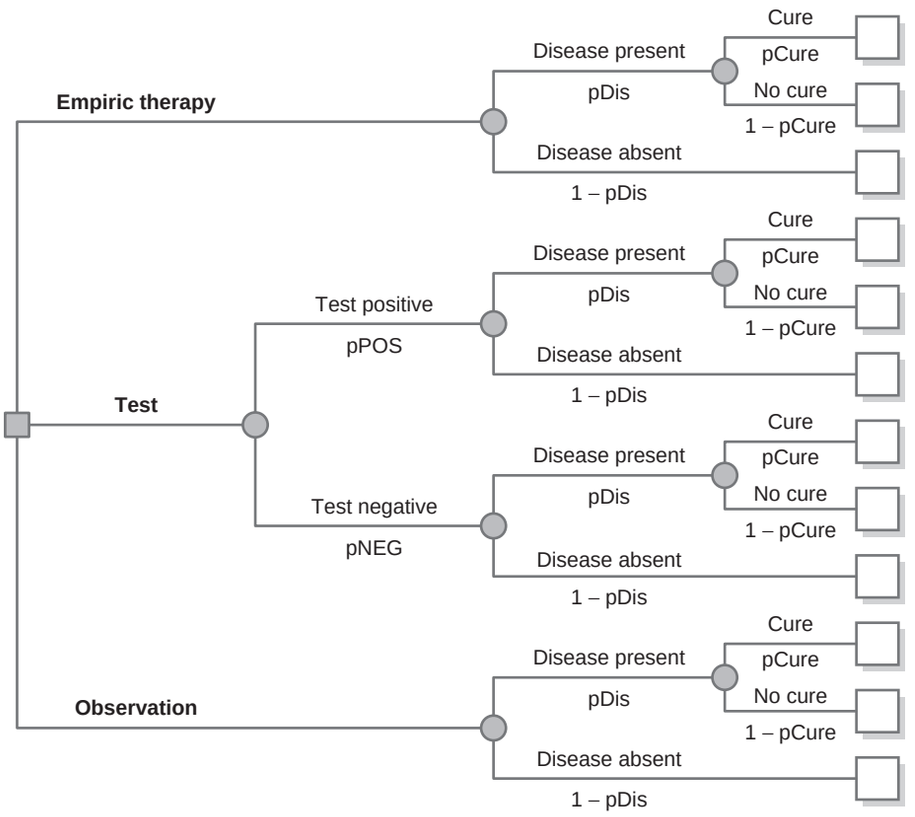


Figure 4 Tree with diagnostic test

C and *B* can be calculated by evaluating the model for pDis = 0 or pDis = 1, respectively.

Consider the more complicated tree shown in Figure 4, which adds a third Test strategy. Figure 5 shows a one-way sensitivity analysis on pDis, which now has three lines. There are now two new thresholds. The Testing threshold is the value of pDis above which Test is favored over Observe. The Test-Treatment threshold is the value of pDis above which Treat is favored over Test.

Two-Way Sensitivity Analysis

A two-way sensitivity analysis looks at variations in two independent variables simultaneously. Since a one-way sensitivity analysis requires a two-dimensional graph, as in Figure 2, a two-way analysis would require a three-dimensional graph, plotting one independent variable on each horizontal axis and the expected utilities on the vertical axis. However, a more convenient way has been devised of representing a two-way analysis on a two-dimensional graph.

Figure 6 illustrates a two-way sensitivity analysis considering simultaneously pDis and SENS (test sensitivity). For each value of pDis, the Test-Treatment threshold is calculated and plotted on the vertical axis. The resulting points define a curve that divides the plane of the graph into two regions. Points above the curve represent combinations of pDis and SENS for which Test is favored. Points below the curve represent combinations of pDis and SENS for which Treat is favored. Figure 7 illustrates the same kind of two-way analysis in which all three strategies are considered. There is an additional curve representing the Testing threshold, thus dividing the plane of the graph into three areas, each favoring one strategy. Note that below the test sensitivity at which the Testing threshold equals the Test-Treatment threshold, testing is not favored regardless of the value of pDis.

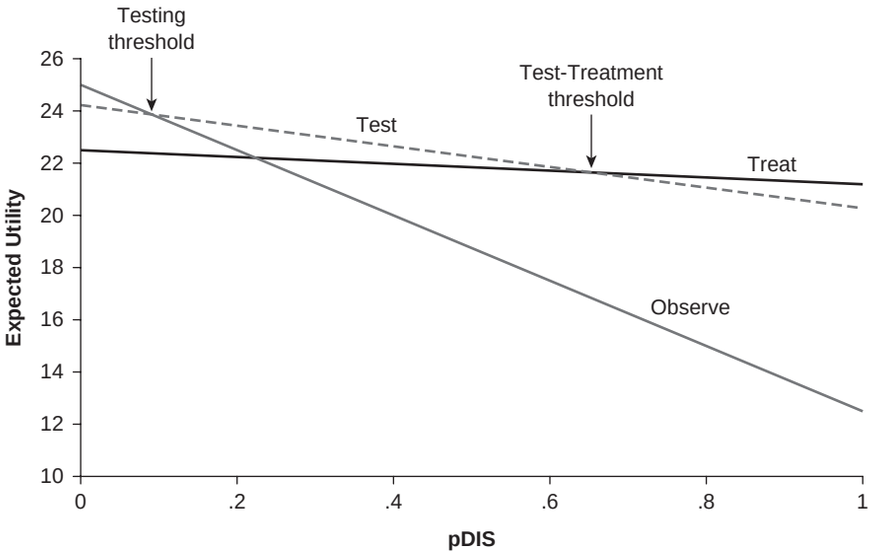


Figure 5 One-way sensitivity analysis with three strategies

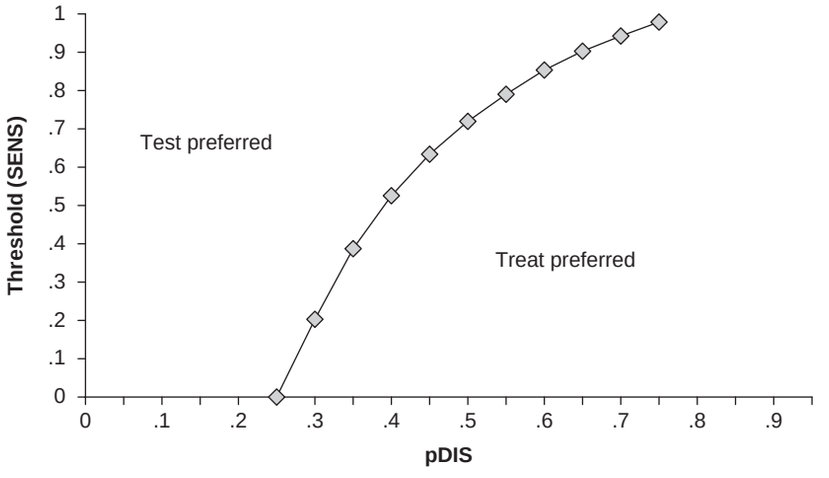


Figure 6 Two-way sensitivity analysis

Three-Way Sensitivity Analysis

As with the two-way analysis, a three-way sensitivity analysis can be represented on a two-dimensional graph by using threshold curves. Figure 8 shows a series of Test-Treatment threshold curves, one for each value of a third variable (test specificity, or Spec). Each curve divides the plane of the graph into a different pair of regions.

Probabilistic Sensitivity Analysis

Sensitivity analyses, as illustrated above, perform deterministic calculations on the model. While they explore variations in key parameters, they do not represent actual uncertainty in the parameters since nothing in the results indicates which scenario is more likely. In PSA, uncertainty in parameters is represented by using probability distributions to represent the values of parameters.

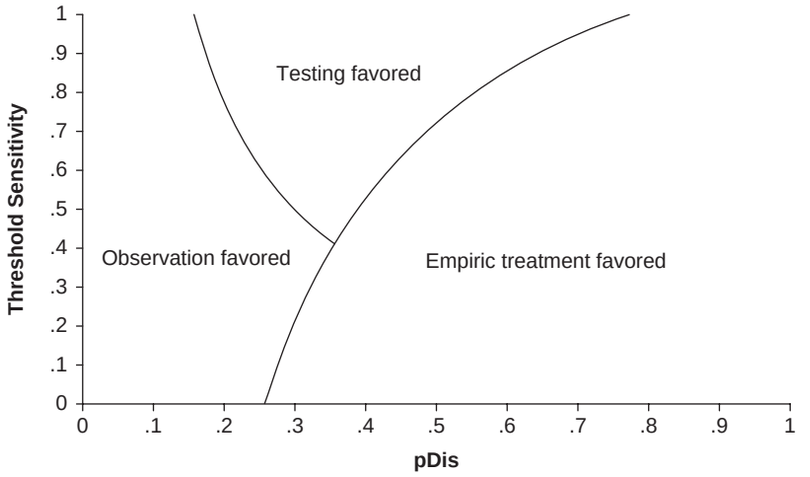


Figure 7 Two-way sensitivity analysis showing all three strategies

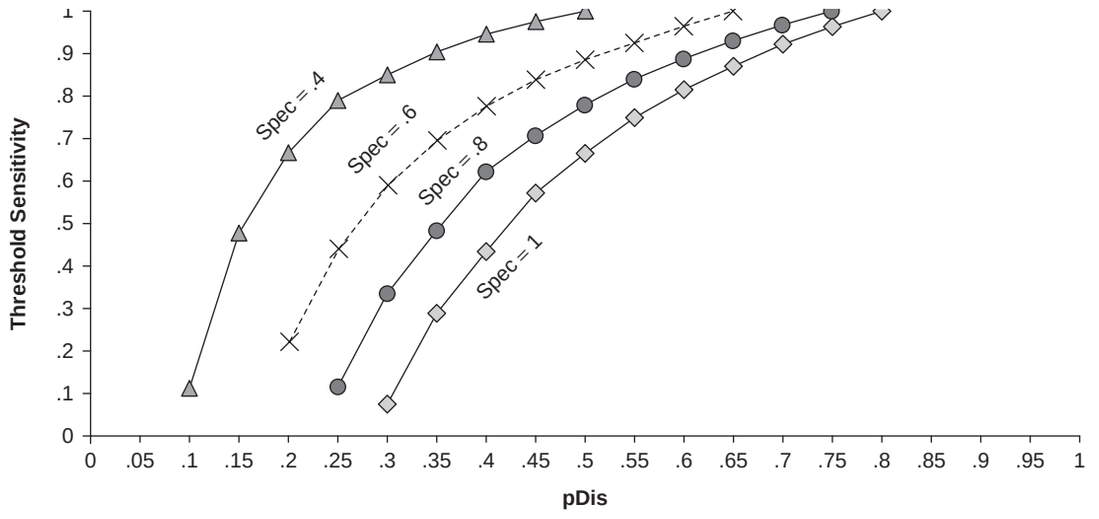


Figure 8 Three-way sensitivity analysis

Mathematical Distributions

A mathematical distribution describes the likelihood that the value of a parameter will be in a certain range. It is usually represented by a probability density function (PDF), as illustrated in Figure 9. The height of each bar (or point on the curve) represents the relative likelihood of the corresponding value (on the horizontal axis) occurring. Probability distributions are characterized by bounds (e.g., 0 to 1 or unbounded), mean value, and shape. A complete discussion of probability distributions is beyond the scope of this entry but may be found elsewhere.

Useful Distributions

The most important distributions in PSA are those representing probabilities. Thus, they must be bounded between 0 and 1. The *beta distribution* has many desirable characteristics for representing probabilities and is therefore commonly used. Parameters for determining the parameters of the distribution (mean and shape) may be determined by analyzing sets of data or by estimating the range of likely values.

In PSA, any number of variables may be represented by distributions. During evaluation, each value is drawn from its distribution according to

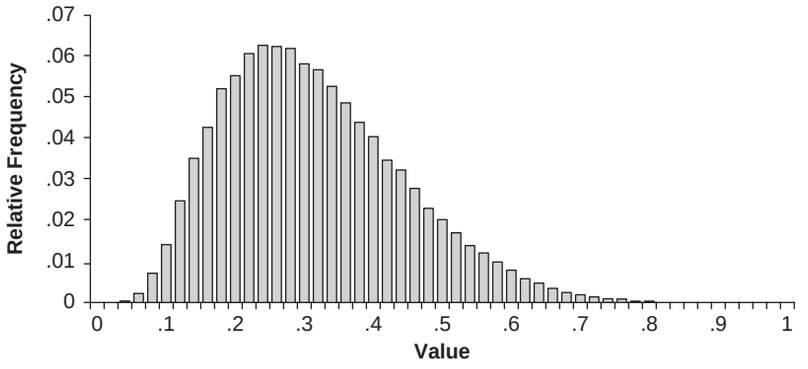


Figure 9 Probability density function for a distribution

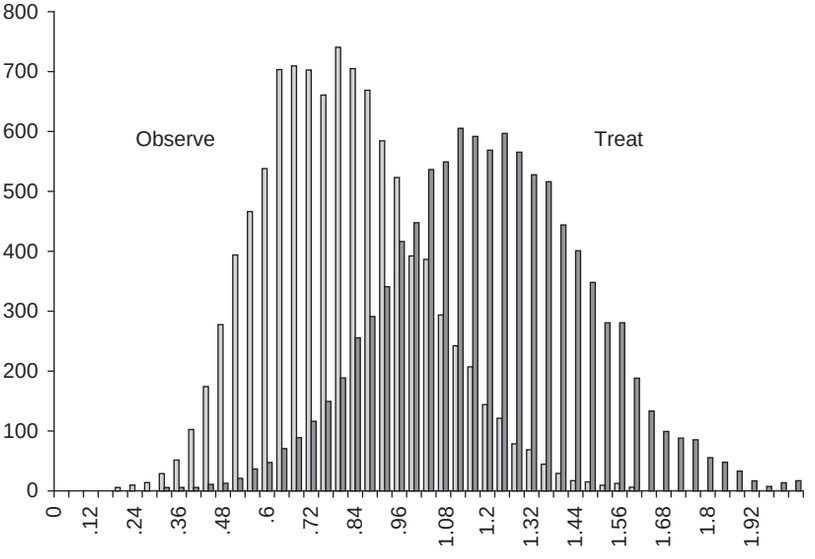


Figure 10 Results of probabilistic sensitivity analysis

its PDF, and the model is evaluated with the resulting set of parameters. This is repeated a large number of times, typically 1,000 to 10,000 times. Because each iteration has a different set of parameters and a different result, the process is said to be *stochastic* rather than *deterministic*. The resulting expected utilities of each of the model's strategies are themselves combined into a results distribution (Figure 10) and thus provide measures of the uncertainty of the results (e.g., variance). The results may be interpreted as the difference in the means of the distributions and also in terms of the percentage of iterations for which one strategy is favored over the other.

Detecting Model Bugs and Errors With Sensitivity Analysis

Another important purpose of sensitivity analysis is detecting errors in the model. The sensitivity analysis illustrated in Figure 11 shows the expected utilities of medical management (MedRx) and cardiac catheterization (Cath) as a function of the probability of left main disease (pLeftMain). Because the

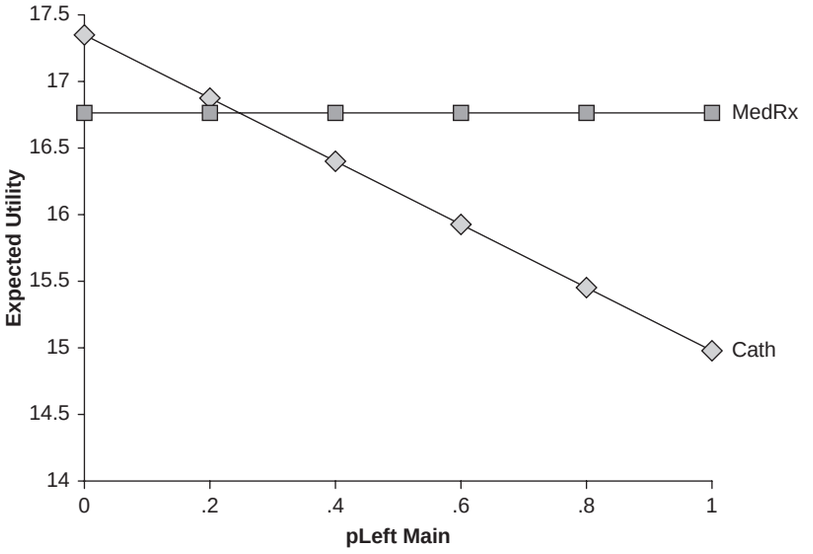


Figure 11 Model "bug" revealed by sensitivity analysis

analyst has neglected to include pLeftMain in the model for medical therapy, it appears that MedRx is favored more when pLeftMain is higher. The opposite is true. This is an example of *asymmetry* error, in which different strategies model the underlying disease or outcomes differently.

Frank A. Sonnenberg

See also Cost-Effectiveness Analysis; Decision Trees: Sensitivity Analysis, Basic and Probabilistic; Decision Trees: Sensitivity Analysis, Deterministic; Test-Treatment Threshold; Threshold Technique

Further Readings

Briggs, A. H. (2000). Handling uncertainty in cost-effectiveness models. *PharmacoEconomics, 17*(5), 479–500.

Doubilet, P., Begg, C. B., Weinstein, M. C., Braun, P., & McNeill, B. J. (1985). Probabilistic sensitivity analysis using Monte Carlo simulation. A practical approach. *Medical Decision Making, 5*(2), 157.

NIST/SEMATECH. (2003). Probability distributions. In *NIST/SEMATECH e-handbook of engineering statistics*. Retrieved January 27, 2009, from http://www.itl.nist.gov/div898/handbook/eda/section3/eda36.htm

Pauker, S. G., & Kassirer, J. P. (1980). The threshold approach to clinical decision making. *New England Journal of Medicine, 302*(20), 1109.

Decision Trees: Sensitivity Analysis, Deterministic

All decision analyses have to deal with various forms of uncertainty in a manner that informs the decisions being made. In particular, it is essential to establish the degree to which the results of an analysis are sensitive to a change in a parameter or an assumption and the extent to which the conclusions of the analysis are robust to such changes. The assessment of sensitivity or robustness is known as sensitivity analysis. Such an analysis would consider, for example, the fact that the mean length of inpatient hospital stay associated with a particular clinical event is estimated with uncertainty (reflected in its standard error) and would consider how the results of the study would change if a higher or lower value were used for this parameter. Two different forms of sensitivity analysis are used in this situation: (1) deterministic analysis, which varies the parameter (or assumption) in one or a small number of stages and assesses the implications for results, and (2) probabilistic analysis, which uses simulation methods to simultaneously vary a number of parameters in terms of a large number of possible alternative values they could take. This entry considers deterministic sensitivity analysis.

Different Types of Uncertainty in Decision Analysis

The uncertainties relevant to a decision model have been categorized in various ways in the literature. The main distinction is between parameter and model (or structural) uncertainty. The former refers to the uncertainty that exists in the parameter inputs that are incorporated into models—for example, the baseline risk of a clinical event in a particular patient group under current treatment, the risk reduction in the event associated with a new intervention relative to current practice, the mean cost of the event, or the mean decrement in health-related quality of life associated with the event. Model uncertainty relates to a range of possible assumptions that are made in developing a model. These could include the extent to which the baseline risk of an event changes over time, the duration of the risk reduction associated with a new intervention, or whether or not to include a particular study in a meta-analysis to estimate the relative treatment effect. The distinction between parameter and model uncertainty is blurred in that many forms of model uncertainty could be expressed in terms of an uncertain parameter.

Deterministic sensitivity analysis can also be used to address heterogeneity rather than uncertainty—that is, to assess the extent to which the results of an analysis change for different types of patients. For example, a treatment may be more effective in females than in males, so the results of the analysis could be separately reported for the two genders. This is probably more correctly labeled as a subgroup, rather than a sensitivity, analysis and is not further discussed here.

Different Forms of Deterministic Sensitivity Analysis

Deterministic sensitivity analysis can be characterized in a number of ways. One is whether a parameter is varied across a range or simply takes on discrete values. In the case of model assumptions that have not been formerly parameterized, the use of discrete values is usually required. Table 1 shows an example of this form of sensitivity analysis (which can also be described as a scenario analysis) in the context of a cost-effectiveness model of endovascular abdominal aortic aneurysm repair (EVAR) compared with open surgery for abdominal aortic aneurysm. It shows the impact of variation on the difference in costs, quality-adjusted life years (QALYs), and the incremental cost-effectiveness ratio relative to the "base-case" or primary analysis. It also shows the results of a probabilistic sensitivity analysis in terms of the probability that EVAR is more cost-effective conditional on a threshold cost-effectiveness ratio. The table mostly includes assessment of uncertainty in the parameter estimates used in the model. However, there are also examples of modeling assumptions that have been varied, for example, Scenario 6 (source of a parameter); and some subgroup analyses are reported (e.g., Scenarios 10 and 11).

An alternative form of deterministic sensitivity analysis is to vary a parameter along a continuous scale and to present this diagrammatically. An example of this is presented in Figure 1, which shows how the incremental cost per QALY gained of primary angioplasty, relative to the use of thrombolysis, in patients with ST-elevation myocardial infarction varies with the additional capital cost per patient required for the angioplasty service. The results are shown for two assumptions regarding the time delay to provide angioplasty compared with thrombolysis.

A second way in which deterministic sensitivity analysis can be characterized is in terms of the number of uncertain parameters/assumptions that are varied simultaneously. Table 1 generally shows analyses that vary one parameter/assumption at a time (one-way sensitivity analysis). There are, however, examples of analyses where two parameters are varied at a time (e.g., Scenarios 12 and 13) (two-way sensitivity analysis). Figure 1 also represents an example of two-way sensitivity analysis in that two uncertain parameters are being varied together: the additional capital cost of angioplasty per patient (as a continuous variable) and the time delay associated with angioplasty (as a categorical variable) compared with thrombolytics.

It becomes very difficult to present deterministic sensitivity analyses when more than two variables are being varied at a time—this is one of several reasons why probabilistic sensitivity analysis might be preferred. One way of looking at multiple

Table I Example of deterministic sensitivity analysis

						Probability EVAR is cost-effective[†]	
Scenario	*Base-case assumption*	*Secondary analysis*	*Difference in cost (£)*	*Difference in QALYs*	*ICER for EVAR versus open**	Λ = *£20,000*	Λ = *£40,000*
1	Base case		3758	–0.020	EVAR dominated	0.012	0.080
2	Hazard of cardiovascular death is twice that of the general population	Baseline hazard of cardiovascular death is the same as the general population	4105	0.017	239,000	0.028	0.161
3	Lower rate of cardiovascular death following open surgery	Same hazard of cardiovascular death following each treatment strategy	3687	0.087	42,000	0.098	0.481
4	1 CT and 1 outpatient visit per year after EVAR	Same cost of monitoring following each treatment strategy	2613	–0.020	EVAR dominated	0.045	0.145
5	Cost of EVAR device is £4800	Cost of EVAR device is £3700	2669	–0.020	EVAR dominated	0.048	0.147
6	Odds ratio of 30-day mortality from EVAR 1 only	Odds ratio from a meta-analysis of DREAM[2] and EVAR trials	3765	–0.015	EVAR dominated	0.012	0.084
7	Discount rate of 3.5%	No discounting of costs nor health benefits	4103	–0.041	EVAR dominated	0.016	0.084

8	Odds ratio of AAA-related death during follow-up from EVAR 1	No difference between EVAR and open repair of the long-term rate of AAA-related death	3859	0.080	48,000	0.076	0.419
9	5% die within 30 days of open repair	8% die within 30 days of open repair	3795	0.090	42,000	0.147	0.463
10	Age 74 years	Age 66 years	4513	–0.144	EVAR dominated	0.001	0.025
11	Age 74 years	Age 82 years	3072	–0.015	EVAR dominated	0.047	0.138
12	Age 74 years and lower long-term rate of cardiovascular death after open surgery	Age 66 years and no difference in rate of cardiovascular death after open repair or EVAR	4468	–0.075	EVAR dominated	0.006	0.068
13	Age 74 years and lower long-term rate of cardiovascular death after open surgery	Age 82 years and no difference in rate of cardiovascular death after open repair or EVAR	2960	0.110	27,000	0.262	0.670

Source: Modelling the long-term cost-effectiveness of endovascular or open repair for abdominal aortic aneurysm. Epstein, D. M., Sculpher, M. J., Manca, A., Michaels, J., Thompson, S. G., Brown, L. C., et al. *British Journal of Surgery, 95,* 183–190.

Note: AAA, abdominal aortic aneurysm; CT, computed tomography; EVAR, endovascular abdominal aortic aneurysm repair; ICER, incremental cost-effectiveness ratio (difference in mean cost divided by difference in mean health benefits); QALY, quality-adjusted life year.

*"EVAR dominated" means EVAR, on average, costs more and has fewer QALYs than open repair and is not expected to be cost-effective.

†The probability EVAR is cost-effective is evaluated at threshold ICERs (λ) of £20,000 and £40,000 per additional QALY20. The National Institute for Health and Clinical Excellence in the United Kingdom has not to date funded interventions with an ICER above £40,000. Given the uncertainty in the model parameters, this represents the probability that a decision to implement EVAR will be better than open repair.

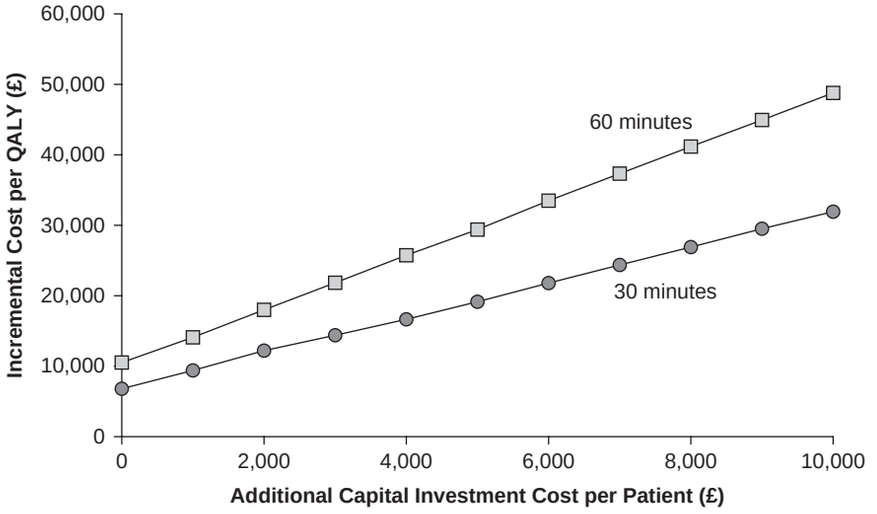


Figure 1 Example of a graphical deterministic sensitivity analysis

Source: Bravo Vergel, Y., Palmer, S., Asseburg, C., Fenwick, E., de Belder, M., Abrams, K., et al. (2007). Results of a comprehensive decision analysis. Is primary angioplasty cost effective in the UK? *Heart, 93,* 1238–1243. Reprinted with permission of BMJ Publishing Group Ltd.

sources of uncertainty is to undertake threshold analysis, a variant on sensitivity analysis. This involves identifying a particular threshold in the results of an analysis that is expected to trigger a policy shift—for example, a point where the incremental cost-effectiveness ratio is equal to a policy maker's cost-effectiveness threshold or when an intervention is expected to generate a net cost saving. The uncertain parameters/assumptions are then varied across a range until the threshold in results is reached, indicating the value(s) of the uncertain variable(s) that, if true, would potentially change a policy decision.

Figure 2 presents an example of a threshold analysis. The context is a cost-effectiveness study of alternative hip prostheses. The analysis provides a general framework for addressing the question of how effective a particular new prosthesis needs to be (in terms of a reduction in the rate of revision procedures) for a given additional cost (compared with a standard prosthesis) to be cost-effective. The example defines *cost-effectiveness* in terms of combinations of additional cost and effectiveness that result in the new prosthesis meeting three alternative thresholds: cost neutrality (including the cost of the prosthesis and other costs of care), an incremental cost per QALY gained of £6,500, and an incremental cost per QALY gained of £10,000.

Mark Sculpher

See also Applied Decision Analysis; Decision Trees: Sensitivity Analysis, Basic and Probabilistic; Managing Variability and Uncertainty; Uncertainty in Medical Decisions

Further Readings

Bravo Vergel, Y., Palmer, S., Asseburg, C., Fenwick, E., de Belder, M., Abrams, K., et al. (2007). Results of a comprehensive decision analysis. Is primary angioplasty cost effective in the UK? *Heart, 93,* 1238–1243.

Briggs, A. H. (2000). Handling uncertainty in cost-effectiveness models. *PharmacoEconomics, 17*(5), 479–500.

Briggs, A., Sculpher, M., Britton, A., Murray, D., & Fitzpatrick, R. (1998). The costs and benefits of primary total hip replacement. How likely are new prostheses to be cost-effective? *International Journal of Technology Assessment in Health Care, 14*(4), 743–761.

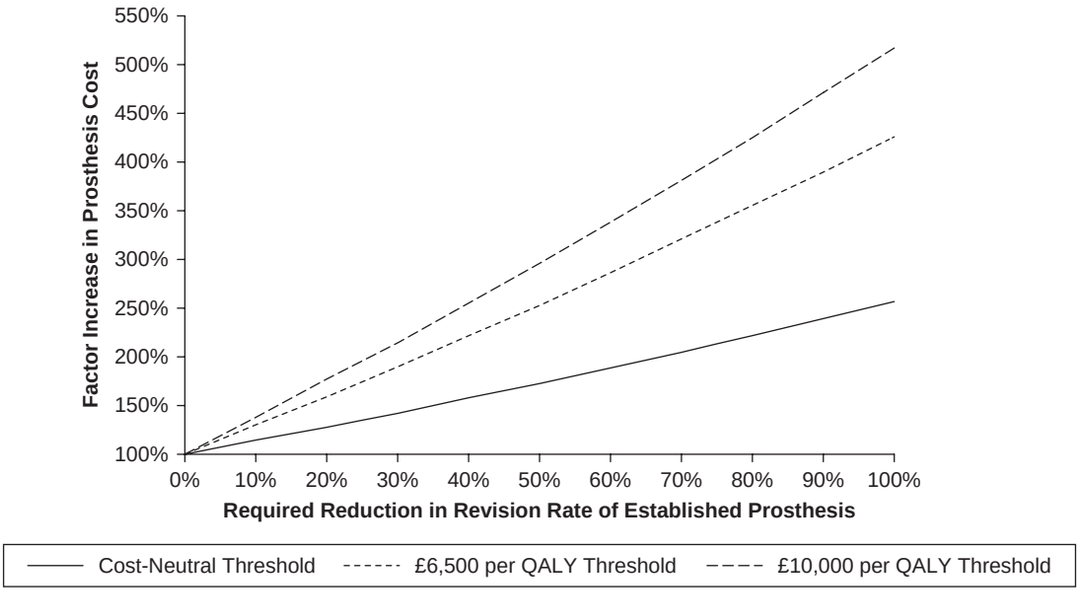


Figure 2 Example of a threshold analysis

Source: Briggs, A., Sculpher, M. J., Britton, A., Murray, D., & Fitzpatrick, R. (1998). The costs and benefits of primary total hip replacement: How likely are new prostheses to be cost-effective? *International Journal of Technology Assessment in Health Care*, *14*, 743–761.

Claxton, K., Sculpher, M., McCabe, C., Briggs, A., Akehurst, R., Buxton, M., et al. (2005). Probabilistic sensitivity analysis for NICE technology assessment: Not an optional extra. *Health Economics, 14,* 339–347.

Epstein, D. M., Sculpher, M. J., Manca, A., Michaels, J., Thompson, S. G., Brown, L. C., et al. (2008). Modelling the long-term cost-effectiveness of endovascular or open repair for abdominal aortic aneurysm. *British Journal of Surgery, 95,* 183–190.

Decision Weights

A decision weight reflects a person's subjective interpretation of an objective probability. Almost all medical decisions involve probabilistic outcomes. For example, there is some chance that a treatment will cure a disease and some chance that the treatment will have a side effect. Data are often available to help patients and providers know the probability that an outcome, such as a serious side effect, will occur. When people face decisions involving uncertain outcomes, how do they use these probabilities?

Theories of rational decision making recommend using the exact value of the probability in evaluating a decision. For example, in expected utility theory, a rational decision maker should evaluate the overall worth of an option by (a) multiplying the probability of each possible outcome by the utility of that outcome and (b) summing the products across all possible outcomes. However, people making actual decisions do not use the real, or "objective," probability when making decisions; the subjective sense of a given probability p is not necessarily the same as p. This phenomenon is analogous to the psychophysics of light perception, in which the brightness a person perceives does not have a 1:1 relationship with the actual luminous energy in the environment.

In the most well-known descriptive theory of decision making, prospect theory, the subjective sense of a probability is known as the *decision weight* corresponding to that probability, denoted by π. Understanding how a person uses objective probabilities in decision making requires knowledge

of that person's *decision weight function*, which describes how probabilities are related to decision weights.

Figure 1 shows a typical decision weight function and illustrates some typical findings from research on decision weights.

First, people tend to overweight small probabilities. Because people have difficulty conceptualizing small probabilities, they translate them into decision weights that are greater than the actual probabilities. This finding might help explain why, for example, both patients and investigators overestimate the small chances of benefit and harm associated with participation in early-phase oncology trials.

Second, people tend to be less sensitive to the differences among probabilities near the middle of the probability scale. Theories of rational decision making state that changes in objective probabilities should make a difference to people. However, actual decision weight functions are relatively flat for intermediate objective probabilities. Thus, a patient might appear to disregard information about the probabilities of success or failure when those probabilities are in the intermediate range (e.g., $p = .25$ to .75). In fact, the patient might be attending to the probabilities presented but assigning them similar decision weights.

Third, the decision weight function is usually steepest as it approaches 0 and 1.00. People tend to prefer changes in probabilities that will result in a state of certainty, something known as the *certainty effect*. Consider a patient deciding between medical and surgical therapies for a heart condition. If the probabilities of success are .80 and .90, respectively, there is a .10-point difference between the treatments. Now imagine that the .10-point difference arises from the probabilities of .90 and 1.00. In expected utility theory, these two scenarios should not be different, because the difference between the options is .10 in both. Yet people do not typically experience these scenarios as equivalent. The decision weight function shows that this is the case because people assign greater weight to the elimination of uncertainty.

Another feature of decision weights is that they are not necessarily additive. Consider a treatment in which only one of two outcomes can occur: (1) a .40 probability of cure and normal life expectancy and (2) a .60 probability of immediate death. Because these are the only possible outcomes, the probabilities sum to 1.00. However, a patient with the decision weight function shown in Figure 1 would convert the probabilities to decision weights that do not sum to 1.00. When people operate according to nonadditive decision weights, their behavior may be contrary to most tenets of rational decision making.

Decision weight functions make it possible to describe many types of situations. For example,

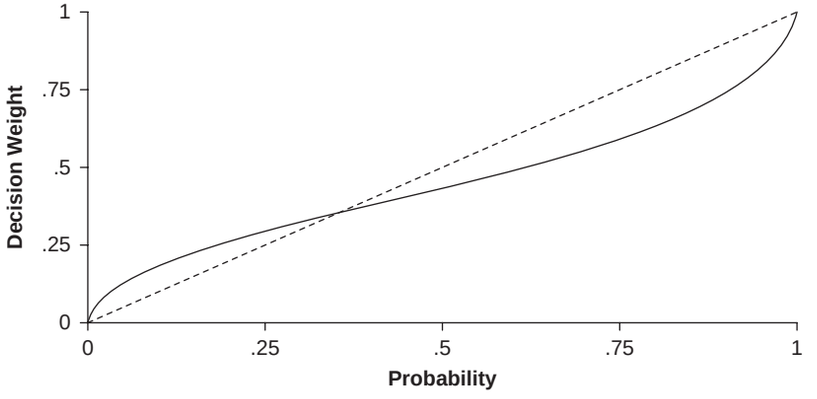


Figure 1 Decision weight function for a hypothetical person

Note: Dotted reference line denotes 1:1 relationship; solid line denotes actual decision function.

some people appear to use a decision weight function that has only three regions. They interpret a probability of 0 as a decision weight of 0 (i.e., there is no chance that the outcome will occur), probabilities between 0 and 1.00 as decision weights equal to .50 (i.e., there is some chance that the outcome will occur), and a probability of 1.00 as a decision weight of 1.00 (i.e., the outcome will occur). Alternatively, some people use a *threshold function*. For example, in deciding which potential side effects to discuss with a patient, a physician might regard all side effects with an objective probability less than .001 to be essentially 0. Because the physician's decision weights for the outcomes are 0, the physician might not mention the side effects to the patient.

Decision weights have implications for the *standard gamble method* of eliciting health utilities. In a simple standard gamble, a patient might be asked to choose between two treatments. Treatment 1 will produce Health state A with a probability of 1.00. Treatment 2 will produce either perfect health (utility = 1.00) with probability p or instant death (utility = 0) with a probability of $1 - p$. The value of p is the point at which the patient is indifferent between Treatments 1 and 2. Assume that $p = .60$. In expected utility theory, the utility of Health state A is calculated as .60. However, this conclusion is only correct if the patient's decision weight for the probability is .60; that is, there is no subjective distortion in the underlying probability. Because the patient's decision weight is generally not known, most researchers interpret standard gamble results as though there is a 1:1 relationship between probabilities and decision weights.

Approaches other than prospect theory extend the use of decision weights to more complex situations. For example, *rank-dependent* models can order multiple possible outcomes in terms of how good or bad they are for the person. In these approaches, it is desirable to understand how people interpret the *cumulative probabilities* of the outcomes. Imagine that a patient with advanced cancer is examining different treatment options. The possible outcomes of treatment are disease progression, stable disease, partial tumor response, and complete tumor response. Here, the patient is less likely to think about the probabilities of each outcome one at a time. Rather, the patient might think about the chance that a treatment will result in an outcome "at least as good as," say, stable disease. A model of such ranked outcomes posits a cumulative decision weight function to correspond to the cumulative probabilities of the outcomes. This promising approach has yet to take hold in studies of medical decision making.

Kevin Weinfurt

See also Expected Utility Theory; Probability; Prospect Theory

Further Readings

Birnbaum, M. H., & Chavez, A. (1997). Tests of theories of decision making: Violations of branch independence and distribution independence. *Organizational Behavior and Human Decision Processes, 71,* 161–194.

Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica, 47,* 263–291.

Tversky, A., & Kahneman, D. (1992). Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and Uncertainty, 5,* 297–323.

Declining Exponential Approximation of Life Expectancy

The declining exponential approximation of life expectancy (DEALE) is a model that simplifies the problem of handling life expectancy calculations in clinical decision analyses. During the early years of clinical decision analysis, much of the focus was on tree construction and probability estimation. Utility or outcome measures were less of a focus; measures such as "percent chance of cure" or "5-year survival" were commonly used in lieu of life expectancy values. In large part, this was because the clinical literature reported results that way. Combining medical risks to estimate survival was rarely done. As decision modelers focused on chronic diseases over short-term

problems, the need arose to model life expectancy for healthy persons and those battling disease.

The Mathematical Formulation

Life expectancy has been studied for 180 years. Benjamin Gompertz, a self-educated English mathematician, published a demographic model in 1825. The Gompertz function is a sigmoid curve, shallow at the beginning and at the end, that represents general-population survival with a fair degree of accuracy. The Gompertz survival function is

$$S(t) = e^{-be^{ct}},$$

where

b is the base rate (i.e., initial mortality) and is negative (decreasing survival),

c is the growth rate (i.e., accelerating mortality), and

e is Euler's constant (= 2.71828 …).

Figure 1 shows the Gompertz survival curve for a healthy population near 70 years. The curve falls slowly at the beginning, with 90% of the population alive after 7 years. By 10 years, however, only 80% of the population is alive, and after 20 years, less than 15% of the cohort is surviving. Thereafter, the curve flattens out, as the still increasing force of mortality acts on the fewer people remaining alive.

Although the Gompertz curve reflects the survival of a healthy population rather well, it offers a basic mathematical challenge: Its integral does not have a closed-form solution. Therefore, the expected value of t in Figure 1 (expected survival time, or life expectancy) cannot be solved exactly. Of course, with modern computational assistance, the area under the survival curve can be calculated to any degree of precision, which would be fine if the only issue were to calculate life expectancy for the general population.

The problem faced in medical decision making adds complexity to this mathematical issue. In a clinical decision analysis, the mortality attached to a disease, or disease-specific mortality, needs to be considered. In many cases, however, the mortality attached to a disease can be estimated from the literature. For many chronic illnesses, a constant specific mortality force can be applied. Assuming that disease-specific mortality rate is independent and additive, the survival function for a person with a chronic disease with constant-mortality rate m would be

$$S(t) = e^{-(be^{ct} + m)}.$$

This additional mortality force would depress the Gompertz curve, more at the beginning than later, as the constant additive risk acts on a larger population early. Of course, this function also cannot be integrated directly, so an expected survival cannot be calculated exactly.

However, if the population mortality were a constant M, then the joint survival function would be

$$S(t) = e^{-(M + m)t},$$

which would be easy to calculate and simple to integrate. The expected value of a probability function is

$$\int_{-\infty}^{\infty} tf(t)dt.$$

For the joint mortality function, which is a probability, the expected value (i.e., life expectancy) is $\int_0^{\omega} te^{-(M + m)t}dt$. The value of this integral is $1/(M + m)$; that is, the life expectancy associated with a constant mortality μ is $1/\mu$.

Of course, the population mortality is not constant. The conceptual attractiveness of the mathematics led Beck, Kassirer, and Pauker to model Gompertz mortality with various clinically plausible constant excess mortality rates, to determine how this constant-mortality assumption would affect overall-survival calculations. They discovered that this DEALE tended to overestimate mortality, especially in later years, and underestimated survival. For diseases with overall life expectancy at or below 10 years, the DEALE model proved a good approximation to detailed calculations using the "correct" formulation.

The DEALE in Medical Decision Making

The first application of the constant-mortality model was in traditional clinical decision analyses, where life expectancy was the desired outcome

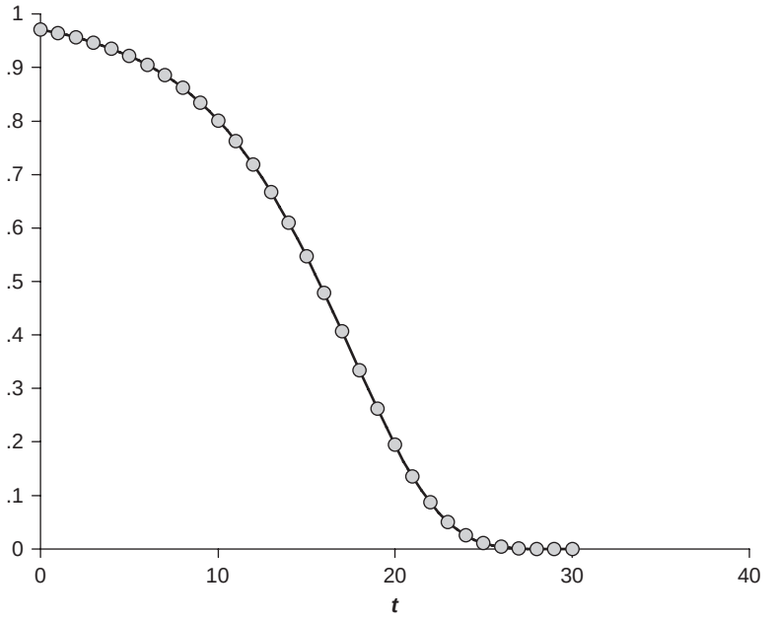


Figure 1 Gompertz survival function

measure. General-population mortality was subdivided by many authors into age-, gender-, and race-specific rates, taken from *Vital Statistics of the United States*. One to three competing disease-specific mortalities, modified by surgical and medical therapies, were added to the population mortality, and reciprocals were taken to generate outcome measures. Until these models were superseded by Markov cohort software, they were standard practice for medical decision analyses where life expectancy was the clinically relevant outcome. Over 30 papers employed this strategy in the 5 years after the DEALE was first published, and the method diffused into many areas of clinical medicine over the ensuing 20 years. Somewhat surprisingly, articles continue to appear in the literature that use the DEALE model as an outcome measure, although with the worldwide availability of personal computers that run decision analysis software, stochastic modeling approaches should be in routine use.

The DEALE as a Bedside Approximation of Mortality

The fact that mortality and life expectancy are reciprocals under the assumption of constant-mortality rate (the negative exponential function) led to another early use of the DEALE, one that has persisted. Suppose a male patient is 65 years of age. According to a life table, his life expectancy is 14.96, or approximately 15 years. The reciprocal of this is .067, or 6.7%. If this patient has a malignancy that has a 10% excess-mortality rate, then his risk of death due to cancer is 1.5 times as great as his general-population mortality. Thus, he has a 60% lifetime risk of death from cancer versus a 40% risk of death from other causes. This approach can be extended to multiple risk factors. Over the past several years, this comparison has been used in oncology to compare therapeutic regimens for patients of varying ages.

The DEALE as a Technique for Probability Estimation

As clinical decision models became more complex, and as trees were supplanted by Markov models for chronic diseases, the DEALE's role as an outcome estimator waned, to be replaced by its enduring value as an aid to probability calculation. The approach has several steps:

1. Obtain a life-expectancy-related value from a study or the literature.
2. Determine what form the value takes: overall-mortality rate, excess mortality, 5-year survival, median survival, and so on.
3. Transform the value into an excess-mortality rate.
4. Transform the excess-mortality rate into a probability.

Table 1 illustrates the first three of these steps (adapted from Beck et al., 1982). Four types of data are commonly found in the literature. Mortality rates are often presented as overall values, which include the underlying population mortality as well as disease-specific excess mortality. These rates are transformed into excess mortality (μ_D) by simply subtracting the population mortality (μ_{pop}) from the overall or compound mortality (μ_C). Life expectancy values are reported as time units, most often years. Taking a reciprocal gives the corresponding constant-mortality rate (an approximation given the Gompertz behavior of general-population mortality). From this, one proceeds as above to obtain μ_D.

Five-year survivals require a bit more calculation. From the survival function $S(t)$ above, some algebra transposes it to $\mu = -(1/t)\ln S$. Substituting 5 for t (5-year survival), the reported value for S (38% in Table 1) will yield μ_C. Similarly, median survival is transformed into mortality by substituting the survival time for t and .5 for S (median survival is the time at which half of the cohort, or 50%, has died).

The final step in using the DEALE to generate transition probabilities is to use the equation

$$p = 1-e^{-rt},$$

where r is the rate, in this case a mortality rate, t the time period (in most cases 1 year or one unit, but not necessarily so), and e the natural logarithm base. Of the nearly 150 articles from 2000 to 2008 that cite the original DEALE papers (and several hundred more that do not), most use probability transformation techniques.

Extensions to the DEALE

Although the DEALE was developed to simplify the problem of handling life expectancy calculations in clinical decision analyses, the "fun" mathematics of the model led to refinements and extensions. Stalpers, van Gasteren, and van Daal extended the model to handle multiple time periods, each with different partial DEALE calculations. Durand-Zaleski and Zaleski showed that the DEALE model could admit discounting of present values as a pseudomortality. Keeler and Bell, and van den Hout looked at other mortality functions and showed how some could admit direct or approximate closed-form solutions that would improve the fidelity of the model. These extensions have found uses in clinical decision analyses. Other refinements essentially put the cart before the horse: The math involved in some sophisticated remodeling was so complex that computer assistance was required to use it.

Gompertz functions and mortality modeling have helped increase the rigor of formal clinical

Table 1 Examples of excess mortality rates

Source	*Study Population*	*Reported Data*	*Compound Rate (μ_C)*	*Baseline Rate (μ_{pop})*	*Excess Rate (μ_D)*
Mortality rate	66-year-old men	.230 per year	.230	.070	.160
Life expectancy	55-year-old women	4.5 years	.222	.037	.185
5-year survival	60-year-olds	38%	.194	.045	.148
Median survival	44-year-old men	7.2 years	.096	.032	.065

decision analyses and risk analyses. Despite the limitations of the approximation, over the past 25 years the approach has meant a good "DEALE" for medical decision making.

J. Robert Beck

See also Decision Tree: Introduction; Life Expectancy; Markov Models

Further Readings

Beck, J. R., Kassirer, J. P., & Pauker, S. G. (1982). A convenient approximation of life expectancy (The "DEALE"). I. Validation of the method. *American Journal of Medicine, 73,* 883–888.

Beck, J. R., Pauker, S. G., Gottlieb, J. E., Klein, K., & Kassirer, J. P. (1982). A convenient approximation of life expectancy (The "DEALE"). II. Use in medical decision making. *American Journal of Medicine, 73,* 889–897.

Durand-Zaleski, I., & Zaleski, S. (1994). DEALE-ing and discounting: A simple way to compute the accrued costs of preventive strategies. *Medical Decision Making, 14,* 98–103.

Gompertz, B. (1825). On the nature of the function expressive of the law of human mortality. *Philosophical Transactions of the Royal Society of London, 115,* 513–585.

Keeler, E., & Bell, R. (1992). New DEALEs: Other approximations of life expectancy. *Medical Decision Making, 12,* 307–311.

Life expectancy tables. Retrieved February 10, 2009, from http://www.annuityadvantage.com/lifeexpectancy.htm

Stalpers, L. J. A., van Gasteren, H. J. M., & van Daal, W. A. L. (1989). DEALE-ing with life expectancy and mortality rates. *Medical Decision Making, 9,* 150–152.

van den Hout, W. B. (2004). The GAME estimate of reduced life expectancy. *Medical Decision Making, 24,* 80–88.

Decomposed Measurement

The decomposed approach to the measurement of preferences for health states, services, or treatments expresses the overall preference as a decomposed function of the attributes of the health state, service, or treatment. It requires the systematic decomposition of the decision problem into smaller parts. It enables the investigator to obtain values for all health states, services, or treatments without requiring the judge to assign values to every one. Decomposition of complex decisions has been shown to aid the decision-making process and its outcomes.

Valuing Health States, Services, or Treatments

Basically, there are two different approaches to measuring preferences for health states, services, or treatments. The holistic approach requires the rater to assign values to each possible health state or treatment, where a state or treatment represents a combination of many attributes. The rater is thus required to simultaneously consider all the relevant attributes during the assessment. The decomposed approach expresses the overall value as a decomposed function of the attributes. The decomposed approach can also be used to simply obtain values for aspects (attributes) of health states or treatments.

As an example, preoperative adjuvant radiotherapy for rectal cancer may increase survival and local control over surgery alone, but at the expense of continence and sexual functioning. The relative value patients place on each of these attributes will determine whether they are prepared to undergo radiotherapy as an adjunct to surgery.

The decomposed models that reveal how a patient values different attributes can be based on statistical inference or explicit decomposition. They have several purposes. First, as in the case of multi-attribute utility theory (MAUT), discussed below, relative importance ratings for attributes can be used to identify global preferences for health states or treatments. Second, where there are individual differences in preferences, the values underlying those preferences can be identified. Such an analysis can highlight the key issues that carers should raise when discussing treatments with patients. For example, conjoint analysis may reveal that lack of energy is an important determinant of preferences for the management of non-metastatic prostate cancer. With this in mind, patient treatment could focus on increasing the energy levels. Such analysis may thus identify new treatment packages that, with minimum cost or

effort, create a much preferred alternative. Third, knowledge of other patients' preference patterns may aid individuals in making choices about their own treatment.

Multi-Attribute Utility Theory

The best-known application of a decomposed method is that based on MAUT, which uses explicit decomposition. Each attribute of a health state (or similarly of a treatment) is given an importance weight. Next, respondents score how well each health state (or treatment) does on each attribute. These scores are weighted by the importance of the attributes and then summed over the attributes to give an overall multi-attribute score for each state (or treatment). For this summation, the theory specifies utility functions and the independence conditions under which they would be appropriate. Gretchen Chapman has used a MAUT model to assess prostate cancer patients' preferences for health states. She describes metastatic prostate cancer by the attributes pain, mood, sexual function, bladder and bowel function, and fatigue and energy, each at three levels of functioning. The attributes had been predefined, and the patients were asked to rate the relative importance of these by dividing 100 points among them. Next, the patients indicated their current level of health for each attribute. MAUT scores were computed by multiplying, for each attribute, the level by the attribute importance weight and summing across the attributes.

Analytical Hierarchy Process

The analytical hierarchy process (AHP) decomposes options into a hierarchy of criteria that include a person's ultimate goal for the decision. First, participants identify their ultimate goal (e.g., maximum possible health and well-being) and the subgoals (criteria) that contribute to it (e.g., avoiding side effects, decreasing the risk of cancer). The participants compare options in a pairwise fashion in terms of these criteria: They give them a rating to indicate which is better or whether they are similar. These pairwise ratings can be combined to give each option a score in terms of each criterion and to work out how the attributes describing an option contribute to achieving the criteria. The participants then prioritize these criteria, giving them a weight to indicate how much they contribute to achieving the ultimate goal. These can be combined to give each option a score in terms of the ultimate goal.

Health State Classification Systems

Both MAUT and statistically inferred regression methods have found well-known applications in the health state classification systems. The two most often used systems are the Health Utilities Index (HUI) and the EQ-5D. Health state classification systems, or health indexes, are customarily composed of two components: a descriptive system and a formula for assigning a utility to any unique set of responses to the descriptive system. The descriptive system consists of a set of attributes, and a health state is described by indicating the appropriate level of functioning on each attribute. For instance, in the EQ-5D, the attributes, or domains, are mobility, self-care, usual activities, pain/discomfort, and anxiety/depression. Each domain is divided into three levels of severity, corresponding to no problem, some problem, and extreme problem. By combining each of the three levels from each of the five domains, a total of 3^5—that is, 243—EQ-5D health states are defined. The formula is generally based on utilities that have been obtained in part from direct measurement and in part from application of MAUT (in the HUI) or statistical inference (in the EQ-5D) to fill in values not measured directly. In both instances, only a limited number of valuations have been obtained from the surveyed population, usually the general public. Of more recent date is a scoring formula based on the SF-36 descriptive quality-of-life instrument. Researchers in the United Kingdom have created from this instrument a six-dimensional health classification system called the SF-6D.

Valuing Aspects of Health States and Treatments

Whereas the ultimate aim of techniques such as MAUT is to assess preferences for health states, or treatments, via decomposition, other techniques aim to measure how treatment or health state attributes in themselves are valued. Judgment analysis, conjoint analysis, discrete choice experiments, and

the repertory grid method each examine how aspects of a treatment or health state influence preferences. In these methods, a holistic valuation technique is used to derive the underlying value of the dimensions described in scenarios. In these cases, a rater holistically values a set of scenarios in which the dimensions appear together in various combinations. The full set of these holistic scores is then analyzed with multiple regression techniques to derive the underlying value each rater was assumed to have assigned to each dimension while making a holistic judgment.

Conjoint Analysis

Conjoint analysis has been widely used to examine consumer preferences, particularly in marketing, and its use in examining patient preferences is increasing with the availability of both generic and specialist software. The principle of conjoint analysis is that evaluations of options are compared to reveal the importance of differences between them. Similar to the statistically based decomposition techniques described above, participants judge hypothetical cases (health states or treatments) that are described in terms of combinations of attributes at particular levels. Statistical analysis reveals the relative importance weights of attributes and identifies sets of attribute-level utilities. Discrete choice experiments are variations on forced-choice conjoint analysis with their roots in economics. Analysis of the data is based on random-utility theory. Judgment analysis is technically similar to conjoint analysis but has its roots in a Brunswikian tradition of psychology, seeking to describe participants' natural judgment processes as they happen, rather than what they would prefer if they had a range of options.

Repertory Grid Technique

The use of repertory grid techniques has been proposed as a bottom-up approach to analyzing what is of more or less importance to patients choosing between treatments. While conjoint analysis and other statistical inference techniques have their roots in psychophysics, perception, and cognition, repertory grid techniques emerged from Kelly's construct theory in social psychology. It has been used to assess patients' quality-of-life measures in relation to their previous and desired states of health. In the statistical inference techniques discussed above, option attributes are defined or identified by the researcher prior to analyzing their relative importance. In the case of the analytical hierarchy process, this may happen after discussion with respondents. In repertory grid analysis, the defining attributes, and their hierarchical combinations, emerge from participants' contrasts between options.

Repertory grid analysis involves four steps. First, in a series of judgments, a participant indicates which of three options (such as treatments) differs from the other two and in what way. This is repeated for all possible triplets of options. Second, each option is rated to indicate to what degree it has this characteristic. Third, characteristics are rated to indicate how important they are. Fourth, a grid of options by characteristics (termed constructs) is analyzed, using simple frequency counts (the number of times a particular construct appears in the option set or the number of overlapping constructs that options have is counted) or using some sort of computer-based cluster analysis. Principal components analysis identifies the correlations between patterns of constructs for each option to reveal which are similar to each other and which constructs tend to co-occur and form a principal component. Generalized procrustes analysis (GPA) is similar to principal components analysis, but it can summarize results across participants even if they have not produced an identical set of constructs.

Anne M. Stiggelbout

See also Conjoint Analysis; Discrete Choice; EuroQoL (EQ-5D); Health Utilities Index Mark 2 and 3 (HUI2, HUI3); Holistic Measurement; Multi-Attribute Utility Theory; SF-6D; Social Judgment Theory

Further Readings

Chapman, G. B., Elstein, A. S., Kuzel, T. M., Nadler, R. B., Sharifi, R., & Bennett, C. L. (1999). A multi-attribute model of prostate cancer patients' preferences for health states. *Quality of Life Research, 8,* 171–180.

Dolan, J. G. (1995). Are patients capable of using the analytic hierarchy process and willing to use it to help make clinical decisions? *Medical Decision Making, 15,* 76–80.

Harries, C., & Stiggelbout, A. M. (2005). Approaches to measuring patients' decision-making. In A. Bowling & S. Ebrahim (Eds.), *Handbook of health research methods: Investigation, measurement and analysis* (pp. 362–393). Maidenhead, UK: Open University/ McGraw-Hill.

Kelly, G. A. (1955). *The psychology of personal constructs*. New York: Norton.

Morera, O. F., & Budescu, D. V. (1998). A psychometric analysis of the "divide and conquer" principle in multicriteria decision making. *Organizational Behavior and Human Decision Processes, 75,* 187–206.

Rowe, G., Lambert, N., Bowling, A., Ebrahim, S., Wakeling, I., & Thomson, R. (2005). Assessing patients' preferences for treatments for angina using a modified repertory grid method. *Social Science Medicine, 60,* 2585–2595.

Ryan, M., & Farrar, S. (2000). Using conjoint analysis to elicit preferences for health care. *British Medical Journal, 320,* 1530–1533.

Von Winterfeldt, D., & Edwards, W. (1986) *Decision analysis and behavioral research*. Cambridge, UK: Cambridge University Press.

Deliberation and Choice Processes

Deliberation is consideration of the reasons for and against an action, issue, or measure. Deliberation may be carried out with attention and without attention. The notion of deliberation without attention brings to realization the notion of dual (or multiple types of) processing of information underlying and affecting acts of deliberating.

Much attention is directed to the notion of how individuals can better focus mentally on their decision problems and structure (formulate) their decision problems to better optimize the decisions they make. These issues become even more of a concern for individuals with declining brain function. A population whose members have a susceptibility to developing neurodegenerative diseases focuses public attention on the loss of the affected individuals' contribution to society and their increased dependence on societal resources to care for them as they continue to age. Today, we have rough estimates of the levels of success in slowing cognitive deterioration from neurodegenerative disease through various modalities. A Dutch team of investigators used brighter daytime lighting to improve patients' sleep and mood and cut aggressive behavior. These researchers found that brighter daytime lighting can slow cognitive deterioration by 5%. This figure is judged at this time to compare well with the rate of slowing of cognitive deterioration in humans through the use of current prescription medicines.

Deliberations With Attention

Research on deliberations with attention may be carried out in nonmedical and medical contexts. In the research arena, investigators have studied nonmedical deliberations involving purchase decisions, which they describe as "simple" (a choice of which towels to buy among a set of towels available at a time) or "complex" (a choice of which car to buy among a set of cars available). One characteristic shared by such simple and complex product purchase decisions is that deliberations with attention regarding such purchases (barring extenuating circumstances) do not have to necessarily be made at that time but can be delayed until a future time, as long as the items are available and the price is right.

Medical Decision Making

While investigators may study simpler and more complex choices in medical deliberations, medical deliberations have a unique quality: the nondelay factor. Many decisions in medicine are a matter of life and death—act now or face the consequences later. Deliberations may not be delayed without adverse consequences for the patient.

The issue of nondelay of medical decisions arises because of the chance that a medical condition will advance if it is not acted on quickly or soon enough. A delay in medical intervention in an individual may cause consequences for that individual in the future. The delayed decision may no longer be about the initial medical condition or disease process but may evolve into a different decision wherein the medical condition or disease considered for intervention is more advanced than it was at that earlier time when first identified.

Consider the following example. The treatment decision in an oncologic disease, such as early-stage

Hodgkin's disease, when it is first diagnosed in a patient is much different from a delayed decision. In the latter case, Hodgkin's disease may have progressed because the intervention was not applied in the early stage. Here, because the competent adult (for whatever reason) chooses not to undergo treatment when the disease is first diagnosed (and is most responsive to therapy), the decision is delayed. Delaying this particular decision means that the disease will continue to grow and continue to progress in its development; the disease may evolve from being curable to being potentially curable and then to being incurable. At the incurable stage, the only treatment option available will be palliation.

Medical Care and Medical Research

Deliberation is at the very heart of some conceptions of decision making in two very different areas: medical care and medical research. The fact that medical care needs to be distinguished from medical research is a view traceable to *The Belmont Report* (1979), created by the National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research.

Medical care and medical research are distinguished in the following ways. In medical care, (a) the care is being provided with one purpose only—that is, to best screen, diagnose, and treat the individual patient, and (b) the patient bears all risks of the screening, diagnostic, and treatment interventions that are undertaken with the patient's permission, but the interventions themselves for the most part have undergone study with the approval of regulatory bodies within government for their use in the population. This is particularly true of medical care with the use of medical products such as prescription medicines and medical devices. These medical products have gone through preapproval research studies, which have developed an evidentiary base for the medical product, and the medical product is approved based on the scientific evidence developed during its research and development phases. In medical care, it is recognized that a competent adult individual—except in the context of a medical emergency—chooses to come to the physician for medical care.

In medical research, the principal investigator (or designee) pursues the individual as part of an act of recruitment, where the individual in many cases is not aware at all of the study's existence prior to being pursued as a study volunteer. Here, the individual is made aware of the study's existence by the pursuit and recruitment processes that have been put into place for a particular research study. This individual is asked to consider study participation and asked to engage in an informed-consent session, where the principal investigator or designee presents the research study, its goals, its methods, its risks, and how liability for injury will be handled within the research study, among other points. Part of the information provided to the individual being recruited into a research study is a discussion of the nature of "research" itself as an activity with one focus only: to attempt to develop new scientific knowledge that might benefit future generations. This new scientific knowledge will then become the evidentiary base for the medical product or medical intervention that may result in approval for its use in the population.

Legal Concepts

Disclosure

The imparting of information for deliberation is termed *disclosure*. The notion of disclosure in the court-defined concepts of consent and informed consent refers to disclosure of information by the physician to the patient (or by the principal investigator to the study volunteer) for the purposes of the patient's deliberations about whether to accept a physician-recommended intervention for the patient's care (or the deliberations of an individual being recruited into a medical research study considering whether or not he or she will enroll in a research trial as a study volunteer). Yet the courts—for example, the landmark 1972 U.S. federal decision in *Canterbury v. Spence*—are also very clear that a competent adult patient in medical care can base his or her decision on whatever grounds the patient sees fit. Similarly, an individual can not enroll in a research study and, even after enrolling, can terminate his or her enrollment in the research study within the bounds of safety for any reason that the individual sees fit.

Autonomy, Trust, and Accountability

Autonomy and self-decision have been foundational concepts in court and medical decision making

in the United States, Canada, and Australia. In England, Onora O'Neill has argued that there are foundational truths besides autonomy and self-decision on which legal structure can be based for the protection of patients and study volunteers. Here, O'Neill argues for "trust" in decision making and "accountability" of those responsible for medical care and medical research as the focus of attention in consent and informed consent. Accountability in this sense involves increased institutional efforts to check on the responsibilities of those overseeing decision making in medical care or medical research to ensure that the best decisions are being made and carried out on the patient's or study volunteer's behalf in medical care and in medical research on humans, respectively.

O'Neill argues that trust and accountability in issues of informed consent can provide a level of protection that is more durable in decision making in medical care and in medical research than considerations solely of individual autonomy. Neil C. Manson and O'Neill discuss the notion of informed consent in terms of waivers against receiving certain types of information. Ethicists continue to examine the conceptual developments related to consent, informed consent, choice, and decision making as conscious choices involving deliberation with attention.

Deliberation Without Attention

Ap Dijksterhuis and colleagues have described what they call the *deliberation-without-attention effect*. The authors argue that it is not always advantageous to engage in thorough conscious deliberation before choosing in the arena of product purchases. As noted earlier, the authors studied simple choices (choices between or among different towels or different sets of oven mitts) and complex choices (choices between different houses or different cars). The authors found that simple choices produce better results after conscious thought but that choices in complex matters should be left to unconscious thought (deliberation without attention).

Neuroeconomics

Alan G. Sanfey and Luke J. Chang define "neuroeconomics" as the science that seeks to gain a greater understanding of decision making by combining theoretical and methodological principles from the fields of psychology, economics, and neuroscience. Key among the early findings of neuroeconomics is evidence that the brain itself may be capable of employing dual-level (or even multiple level) processing of information when making decisions. Sanfey and Chang argue that while behavioral studies provide compelling support for the distinction between automatic and controlled processing in judgment and decision making, less is known about to what extent these components have a corresponding neural substrate. Yet there are other effects on judgment and decision making that need further clarification with neuroeconomics.

Deliberation Deficits

Disinhibition is a process whereby an individual with a measurable capacity to edit his or her immediate impulsive response to a stimulus or situation is rendered to have a deficit in this capacity. Such incapacities are found (a) after brain injuries to the orbitofrontal and basotemporal cortices of the right hemisphere of the brain (caused by closed-head traumatic brain injuries, brain tumors, stroke lesions, and focal epilepsy), which selectively inhibit or release motor, instinctive, affective, and intellectual behaviors elaborated in the dorsal cortex; (b) after the application of agents such as alcohol; and (c) after the use of prescription medicines such as the benzodiazepines alprazolam and flunitrazepam. Benzodiazepines have an effect on gamma-aminobutyric acid, the chief inhibitory neurotransmitter in the central nervous system and the retinas of humans.

Future Research

Future research in the area of deliberations and choice will continue to clarify three areas: deliberation with attention, deliberation without attention, and the impact of brain lesions, agents, and prescription medicines on choice and deliberation. Research is also needed on how best to define and measure nonrisky and risky options over which deliberations are carried out in research trials on medical decision making.

Dennis J. Mazur

See also Decisions Faced by Institutional Review Boards; Informed Consent

Further Readings

Bhugra, D. (2008). Decision making by patients: Who gains? *International Journal of Social Psychiatry, 54,* 5–6.

Griffin, R. J., Yang, Z., ter Huurne, E., Boerner, F., Ortiz, S., & Dunwoody, S. (2008). After the flood: Anger, attribution, and the seeking of information. *Science Communication, 29,* 285–315.

Lane, S. D., Cherek, D. R., & Nouvion, S. O. (2008). Modulation of human risky decision making by flunitrazepam. *Psychopharmacology (Berlin), 196,* 177–188.

Lau, H. C., & Passingham, R. E. (2007). Unconscious activation of the cognitive control system in the human prefrontal cortex. *Journal of Neuroscience, 27,* 5805–5811.

Manson, N. C., & O'Neill, O. (2007). *Rethinking informed consent*. New York: Cambridge University Press.

National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research. (1979). *The Belmont report: Ethical principles and guidelines for the protection of human subjects of research*. Washington, DC: U.S. Department of Health, Education, and Welfare.

O'Neill, O. (2004). Accountability, trust and informed consent in medical practice and research. *Clinical Medicine, 4,* 269–276.

O'Neill, O. (2004). Informed consent and public health. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences, 359,* 1133–1136.

Riemersma-van der Lek, R. F., Swaab, D. F., Twisk, J., Hol, E. M. Hoogendijk, W. J. G., & Van Someren, E. J. W. (2008). Effect of bright light and melatonin on cognitive and noncognitive function in elderly residents of group care facilities: A randomized controlled trial. *Journal of the American Medical Association, 299,* 2642–2655.

Sanfey, A. G., & Chang, L. J. (2008). Multiple systems in decision making. *Annals of the New York Academy of Sciences, 1128,* 53–62.

Deterministic Analysis

Deterministic analysis and decision analysis are not interchangeable. Instead, deterministic analysis is one of the analytical approaches under decision analysis. Under the framework of decision analysis, deterministic analysis conducts mathematical calculations to compare the outcomes of interest.

In medical decision making, the outcome of medical interventions is usually measured by clinical efficacy, effectiveness, or cost-effectiveness. Deterministic analysis compares outcomes of alternative interventions by developing a mathematical model to calculate the value of the outcomes associated with each intervention. The model is often structured in the form of a decision analytical model and contains a number of parameters that affect the outcome of interventions. Deterministic analysis uses the best available estimate of each parameter as the model input and the report point estimate, such as means or median, of the outcome of interest as the model output. For example, deterministic analysis of a cost-effectiveness analysis comparing two interventions may include probabilities of the occurrence of certain clinical events, utilization patterns of healthcare resources, and unit cost associated with each type of healthcare resource as the model inputs and may report the results in terms of a point estimate of the incremental cost-effectiveness ratio (ICER), calculated as the difference in the mean cost between the two competing interventions divided by the difference in the mean effectiveness between these two interventions. Deterministic analysis is not only a terminology used in the field of medical decision making, it is also mentioned in the literature of operational research, civil engineering, and risk assessment, among others.

Sensitivity Analyses

Findings from deterministic analyses serve as the base case scenario of the model output, and researchers apply sensitivity analyses to evaluate whether the conclusions derived from the model are sensitive to the model parameters. Sensitivity analyses vary the model parameters within reasonable ranges to examine the effect of these parameters on the conclusion of the analyses. The number of parameters assessed in sensitivity analyses often ranges from one (known as the one-way sensitivity analyses) to three (three-way sensitivity analyses) because it becomes extremely difficult to interpret the findings of sensitivity analyses if the number of

parameters exceeds three. For ease of illustration, one-way sensitivity analyses are most frequently used to address uncertainties in deterministic modeling. In these analyses, researchers will vary parameters of interest one at a time to determine which parameter(s) has the largest effect on the study findings.

When reporting findings from deterministic analyses, it is common to add a *tornado diagram* to summarize the results of one-way sensitivity analyses graphically. Tornado diagrams are charts that use horizontal bars to describe the magnitude of effect associated with each parameter. Decision makers can visually identify the most influential parameters based on the width of each bar in the diagram. Another analysis commonly added to sensitivity analyses, but not limited to one-way sensitivity analyses, is the *threshold analysis*, in which the deterministic model calculates the parameter value or values indicating that decision makers are indifferent between two interventions (i.e., the break-even point). The idea of threshold analysis is to inform decision makers of the minimum or maximum value (i.e., the threshold value) of a certain model parameter for an intervention to be considered effective or cost-effective. Readers who are looking for straightforward examples and clear graphical illustrations of various forms of sensitivity analyses should read the chapter "Sensitivity Analysis" in Petitti (2000).

Advantage

The advantage of deterministic analyses is that the output of the model is summarized in an exact number (e.g., life expectancy, quality-adjusted life years [QALY], and ICER), which makes it easier for decision makers to select the best intervention. For example, in a deterministic analysis comparing the life expectancy of various interventions, decision makers can simply identify the best intervention by picking the intervention that yields the highest mean life expectancy calculated from the model. Similarly, in a deterministic cost-effectiveness analysis comparing a new intervention with a standard-of-care intervention, decision makers can determine whether the new intervention is cost-effective by assessing whether the ICER calculated from the model is lower than the level of willingness to pay society sets forth for new medical interventions (e.g., $50,000 or $100,000 per QALY). However, as the model becomes more complex, the number of parameters involved increases accordingly, and it becomes more difficult to understand the results of sensitivity analyses due to the excessive number of parameters (for one-way sensitivity analyses) or combinations of parameters (for two- or three-way sensitivity analyses) to be explored.

Relationship With Stochastic Analyses

Although deterministic analyses have the advantage of being exact, the information presented in these analyses is not sufficient to perform hypothesis testing. Therefore, in studies comparing two interventions, deterministic analyses are able to calculate the mean difference in effectiveness between these two interventions but cannot inform decision makers whether the calculated difference can be considered statistically significant. For the purpose of hypothesis testing and to obtain information on the uncertainties associated with model parameters or estimates, it is necessary to conduct another type of analysis known as *stochastic analysis* (or *probabilistic analysis*). The distinction between deterministic and stochastic analyses can be clearly understood in the context of assessing the effectiveness of health interventions, in which deterministic analyses are viewed as analyses that use information on *the average number of events per population*, whereas stochastic analyses use randomization to simulate *the probability distributions of events that may occur*.

There are a number of important differences between deterministic and stochastic analyses. First, deterministic analyses report results as exact numbers, while stochastic analyses present findings either in 95% confidence intervals or as the probability that one treatment is more effective (or more cost-effective) than the other(s). The former presentation is based on analyses taking a classical statistical approach (also known as the frequentist approach), and the latter uses the Bayesian approach. Second, deterministic analyses assume certainty about parameter values that are used as model inputs, whereas stochastic analyses explicitly acknowledge uncertainties in parameter values and describe them in probability distributions. For example, when incorporating hospitalization cost as one of the components in the estimation of total

medical costs, deterministic analyses will include the average cost per hospitalization as the model input, but stochastic analyses will use either lognormal or gamma distribution to characterize this parameter. Last, the lack of knowledge about model parameters was addressed with sensitivity analyses in deterministic analyses and probabilistic sensitivity analyses in stochastic analyses. As discussed previously, sensitivity analyses vary the model parameters within a reasonable range to determine the impact of each parameter (or a combination of two or more parameters) on the study findings. In probabilistic sensitivity analysis, model parameters are described as random variables, each with its own designated probability distribution, and researchers can perform Monte Carlo simulations to estimate the mean and standard deviation of the expected outcome(s) or calculate the probability that one strategy performs better than the other(s). Doubilet and colleagues provided a clear illustration of probabilistic sensitivity analysis. In their article comparing the expected utility among three treatment strategies, (1) biopsy but no treatment, (2) treat but no biopsy, and (3) no biopsy and no treatment, deterministic analyses showed that the expected utilities associated with the above three strategies were .558, .566, and .494, respectively. That is, the strategy "treat but no biopsy" had the highest expected utility. However, such an analysis did not inform the decision maker whether this strategy was significantly better than the other two strategies. On the contrary, the results from the probabilistic sensitivity analyses indicated that the likelihood that "treat but no biopsy" was the best strategy was 80%, as compared with 18% and 2% for the "biopsy but no treatment" and "no biopsy and no treatment" strategies, respectively.

Deterministic and stochastic analyses should not be viewed as rival analytical approaches. Indeed, a comprehensive study is expected to present results from both deterministic and stochastic analyses. Perhaps the best way to describe the relationship between these two types of analysis was expressed in a review article by Corner and Corner in 1995. The authors envisioned a decision problem from a systems engineering perspective and characterized the decision-making process in four steps. In Step 1, a basic structure was developed to model the decision problem and identify the relevant parameters in the model. In Step 2, deterministic analysis was performed, along with sensitivity analysis, to remove those variables that would not affect the final results. Step 3 involved a complete analysis of uncertainty using stochastic analysis and concluded with a recommendation of the best (or most cost-effective) strategy. Step 4 related to model validation and the value of information analysis. Together these four steps complete a decision analysis cycle. The decision-making process can become iterative as information gained from Step 4 may lead to modification of the model structure, thus starting the decision cycle from Step 1 again.

Recent methodological development has made substantial improvements in the statistical and computational methods used in stochastic analysis. This does not mean that deterministic analysis has lost its role in medical decision making. Regardless of how sophisticated the analytical techniques have become, the exact value calculated from deterministic analyses is often what matters most to decision makers, or at least what is most remembered by them.

Ya-Chen Tina Shih

See also Bayesian Analysis; Confidence Intervals; Cost-Effectiveness Analysis; Cost-Utility Analysis; Expected Value of Perfect Information; Frequentist Approach; Hypothesis Testing; Life Expectancy; Managing Variability and Uncertainty; Marginal or Incremental Analysis, Cost-Effectiveness Ratio; Probability; Quality-Adjusted Life Years (QALYs); Statistical Testing: Overview; Threshold Technique; Tornado Diagram; Uncertainty in Medical Decisions; Variance and Covariance

Further Readings

Briggs, A. H. (1999). A Bayesian approach to stochastic cost-effectiveness analysis. *Health Economics, 8*(3), 257–261.

Briggs, A. H. (2000). Handling uncertainty in cost-effectiveness models. *PharmacoEconomics, 17*(5), 479–500.

Claxton, K., Sculpher, M., McCabe, C., Briggs, A., Akehurst, R., Buxton, M., et al. (2005). Probabilistic sensitivity analysis for NICE technology assessment: Not an optional extra. *Health Economics, 14*(4), 339–347.

Corner, J. L., & Corner, P. D. (1995). Characteristics of decisions in decision analysis practice. *Journal of the Operational Research Society, 46*(3), 304–314.

Doubilet, P., Begg, C. B., Weinstein, M. C., Brawn, P., & McNeil, B. J. (1985). Probabilistic sensitivity analysis using Monte Carlo simulation: A practical approach. *Medical Decision Making, 5*(2), 157–177.

Mandelblatt, J. S., Fryback, D. G., Weinstein, M. C., Russell, L. B., Gold, M. R., & Hadorn, D. C. (1996). Assessing the effectiveness of health interventions. In M. R. Gold, J. E. Siegel, L. B. Russell, & M. C. Weinstein (Eds.), *Cost-effectiveness in health and medicine* (Chap. 5). New York: Oxford University Press.

Petitti, D. B. (2000). *Meta-analysis, decision analysis, and cost-effectiveness analysis: Methods for quantitative synthesis in medicine* (2nd ed.). New York: Oxford University Press.

Shih, Y. C. T., & Halpern, M. T. (2008). Economic evaluation of medical interventions for cancer patients: How, why, and what does it mean? *CA: A Cancer Journal for Clinicians, 58,* 231–244.

Developmental Theories

Developmental theories concern changes that occur over the lifespan as a result of maturation and experience. The nature of decision making shifts as children become adolescents and, as more recent research shows, as adolescents become adults and adults age. Two major theories of decision making are discussed that are also theories of development: the prototype/willingness model and fuzzy-trace theory. When discussing decision making in a medical context, it is important to keep in mind the key concepts of risk perception and informed consent (including issues of autonomy). How these theories address each of these issues and their implications for development and rationality are discussed.

In discussing what rationality in decision making is, it important to note two approaches offered as criteria: coherence and correspondence. The coherence criterion for rational decision making is that a decision is rational if the process used is internally consistent. For example, decision makers use a logical rule to combine their assessments of the costs and benefits of each option. Furthermore, the choice made must reflect the decision makers' goals. This coherence criterion is what is traditionally referred to when a process is described as rational. For the coherence criterion, the outcome of the decision is not involved in denoting a decision as rational. The correspondence criterion argues that outcomes do matter. To the extent that the decisions made correspond with good outcomes in reality (e.g., they cause no harm to the decision maker or to others), the decision can be considered rational. Researchers who focus on the health of children and youth often emphasize positive outcomes. However, coherent reasoning is also relevant for issues such as whether young people are capable of giving informed consent for medical treatments.

The two theories discussed here are dual-process theories of decision making. These theories argue that there are two ways in which a decision maker can arrive at a decision. One process is rational (in the traditional sense) and analytic. This process involves the decision maker combining relevant factors using a logically defensible decision rule; behavior resulting from this process is a planned and intentional action. The other process is described as intuitive. This process is quick and does not involve deliberation. Although both theories are similar in that they propose a dual-process distinction, they differ in what is proposed for developing and what is considered rational. Crucially, intuition in prototype/willingness theory is developmentally primitive, whereas intuition in fuzzy-trace theory characterizes advanced thinking.

Prototype/Willingness Model

A standard dual-process theory, the prototype/willingness model has been applied to many health decisions, such as the decision to smoke or drink, and to health-promoting behaviors, such as cancer screening and family planning. The prototype/willingness model argues that there are two paths to a decision, a reasoned path and a reactive path. For the reasoned path, intentions are the direct antecedent to behavior. In turn, intentions are a function of subjective norms and attitudes. Decisions using the reasoned path are deliberative and planned and characterize more mature decision makers. The reactive path was proposed to capture behavior that is not deliberative and is captured by the construct of willingness. Research has shown that willingness is able to explain unique variance when included in a model with behavioral intentions. For

the reactive path, individuals are said to form images of the prototypical person who regularly performs the behavior. What dictates behavior from this process is the reaction that the individual has to this prototype. For instance, producing a prototype of a smoker, an individual can have a positive reaction to the prototype, increasing the probability that the individual will smoke, or a negative reaction to the prototype, decreasing the probability that the individual will smoke. (The theory also holds that a negative image can sometimes be viewed as a cost of engaging in the behavior.) Furthermore, individuals recognize that the more they do the behavior, the more they will come to be perceived as similar to the prototype.

For the prototype/willingness model, development progresses from greater use of the reactive path as children get older to greater reliance on the reasoned path as adults. Therefore, the reasoned path is considered the rational process. Because adolescents are said to be preoccupied with social images and identities, they are more likely to rely on the reactive path than adults. Studies have shown that a positive relationship between intentions and behavior increases with age. Risk perception for the reactive path is defined by the reaction the individual has to the prototype, yet for the reasoned path, it is dictated by the knowledge the individual has of the risk.

Fuzzy-Trace Theory

A more recent dual-process theory, fuzzy-trace theory is based on studies of memory, reasoning, social judgment, and decision making. The theory has been applied to children, adolescents, younger adults, and older adults as well as to groups varying in expertise, such as medical students and physicians. The phrase *fuzzy trace* refers to a distinction between gist memory representations that are fuzzy (i.e., they are vague and impressionistic) and verbatim memory representations that are vivid. Reasoning gravitates to using gist (or fuzzy) representations, which minimizes errors. Moreover, this adaptive tendency to use gist representations—the fuzzy-processing preference—increases with development as children and youth gain experience. Studies of children (comparing older with younger children) and of adults (comparing experts with novices in a domain of knowledge) have demonstrated that reliance on gist representations increases with development. People make decisions using simple gist representations of information, often processing it unconsciously, and engage in parallel rather than serial processing of that information (leaping ahead based on vague gist impressions of the relations and patterns in information without fully encoding details). This kind of thinking is what is meant by "gist-based intuitive reasoning." What develops with age and experience, therefore, is a greater reliance on gist-based intuition in decision processes. Fuzzy-trace theory has been used to describe developmental trends in adolescent risky decision making, HIV prevention, cardiovascular disease, and cancer prevention.

Specifically, fuzzy-trace theory relies on four basic principles in explaining decision making: (1) parallel encoding, (2) the fuzzy-to-verbatim continua, (3) the fuzzy-processing preference, and (4) task calibration. Parallel encoding states that people extract patterns from the environment and encode them along with exact surface form information. These traces (verbatim and gist) are independent, as previously discussed. The second principle, the fuzzy-to-verbatim continua, states that people encode multiple representations at varying levels of precision. At one end are factual, detailed verbatim representations, and at the other end are simplified, abstracted gist representations. These representations are sensitive to environmental cues, meaning that either could be used in the decision process, depending on which representation is cued in context. Verbatim representations support a quantitative, analytic process, while gist representations support an intuitive/holistic process. Since problems are represented at multiple levels of specificity, the same problem can be approached analytically (verbatim) or intuitively (gist) depending on which representation is retrieved. The third principle, task calibration, states that the lowest level of gist required is used to perform the task. For instance, when deciding between Option A, gaining $5, or Option B, gaining $7, one need only remember the ordinal distinction between the two, B > A, to choose B. Finally, the fuzzy-processing preference states that individuals prefer to operate on the simplest representation (gist) needed to accomplish their goals. For development, studies have shown that young children are more likely to make decisions based

on quantitative differences and that what develops with experience is a greater reliance on gist representations, a finding predicted by fuzzy-trace theory. Therefore, consistent with fuzzy-trace theory, gist-based intuitive reasoning has been shown to be the more advanced (and consequently more rational) mode of processing.

Risk perception can vary along the fuzzy-to-verbatim continua in that it can be precise, for example, remembering the exact risk that was conveyed if the surgery were done, or it can be fuzzy, for example, remembering that there is a risk with surgery but not the exact number. Fuzzy-trace theory explains and predicts the major findings in risk perception and risk taking—for example, that risk perceptions vary greatly depending on how they are elicited. The theory also predicts reversals in the relation between risk perception and risk taking depending on whether people use gist-based intuition or verbatim-based analysis. Paradoxically, adolescents often take risks that compromise health because they logically analyze the details of decisions. Adults avoid unhealthy risk taking by considering the gist, or bottom line, of the decision. Fuzzy-trace theory also explains most of the biases and fallacies exhibited in judgment and decision making (ratio bias, framing effects, hindsight bias, base-rate neglect, conjunction fallacy, disjunction fallacy, and others). Many of these biases and fallacies have been demonstrated in medical decision making by patients and healthcare professionals. Fuzzy-trace theory also predicts (and this prediction has been borne out by data) that many biases increase from childhood to adulthood because they are caused by gist-based intuition.

Informed Consent

Recently, there has been an emphasis on increasing the role the patient has in his or her medical decisions. The patient-practitioner relationship has been steadily growing from paternalism to egalitarianism. Evidence has shown that involving patients in their own medical decisions has a positive effect on their well-being. One of the central issues of this move centers on the concept of informed consent. Informed consent involves a decision, or authorization, given without coercion and involves the decision maker having a fundamental understanding of the risks and benefits. Informed consent is given with volition and is usually assumed to involve an underlying rational process. Given that it is rational, it is assumed that to give fully informed consent, the decision maker must be intellectually competent and mature. In discussing the matter of young children, the issue is not one of consent, in that it is clear that children are not considered on par in maturity and cognitive capacity with adults. For young children, decisions are left up to the parent or guardian. However, the case of whether or not an adolescent is capable of providing informed consent is still an ongoing debate. Evidence supporting both sides of the issue has been found. For instance, older adolescents were found to perform on par with adults in a task involving hypothetical medical scenarios. These adolescents were able to select options based on logical reasoning and give valid evidence for their choices, and they had a clear understanding of the costs and benefits of the options. However, other studies have shown that real differences between adults and adolescents do exist. For example, adolescents' goals are more likely than adults' to maximize immediate pleasure, adolescents take more risks in the presence of peers than adults, and the brain is still not fully mature in adolescence. Therefore, the issue of autonomy in adolescence and of whether adolescents can make a rational decision is still unresolved. How each theory handles consent is important with respect to medical decision making.

Prototype/willingness does not specifically address the concept of consent. For the prototype/willingness model, however, using the reasoned path is considered the preferred process. Therefore, deliberating about details and precise knowledge of the options involved in the process matter greatly. For fuzzy-trace theory, making an informed decision requires a grasp of the bottom-line meaning of the situation (e.g., there is a fatal risk involved in the surgery), not simply regurgitating the minutia. For example, imagine that two patients are informed that the risk of death from surgery is 2% and each is later asked to recall what the risk they were informed is. One patient says 0% and the other 10%. Although the patient reporting 0% is objectively more correct (2% off is closer than 8% off), the patient reporting 10% is more informed because he or she understands that the surgery does have some risk. Research has shown that patients often

cannot recall the details of surgical risks and that consent is driven instead by their understanding of the gist of the options. People low in numeracy, the ability to understand and use numbers, have difficulty getting the gist of health information, which impairs informed medical decision making. In sum, developmental differences related to age, experience, and knowledge determine informed consent and the quality of medical decisions.

Steven Estrada, Valerie F. Reyna, and Britain Mills

See also Dual-Process Theory; Fuzzy-Trace Theory; Intuition Versus Analysis; Risk Perception

Further Readings

Fischhoff, B. (2008). Assessing adolescent decision-making competence. *Developmental Review, 28,* 12–28.

Gerrard, M., Gibbons, F. X., Houlihan, A. E., Stock, M. L., & Pomery, E. A. (2008). A dual-process approach to health risk decision-making: The prototype–willingness model. *Developmental Review, 28,* 29–61.

Jacobs, J., & Klaczynski, P. A. (2005). *The development of judgment and decision making in children and adolescents.* Mahwah, NJ: Lawrence Erlbaum.

Kuther, T. L. (2003). Medical decision-making and minors: Issues of consent and assent. *Adolescence, 38,* 343–358.

Reyna, V. F. (2004). How people make decisions that involve risk. A dual-processes approach. C*urrent Directions in Psychological Science, 13,* 60–66.

Reyna, V. F., & Adam, M. B. (2003). Fuzzy-trace theory, risk communication, and product labeling in sexually transmitted diseases. *Risk Analysis, 23,* 325–342.

Reyna, V. F., & Farley, F. (2006). Risk and rationality in adolescent decision making: Implications for theory, practice, and public policy. *Psychological Science in the Public Interest, 7,* 1–44.

Reyna, V. F., & Lloyd, F. (2006). Physician decision making and cardiac risk: Effects of knowledge, risk perception, risk tolerance, and fuzzy processing. *Journal of Experimental Psychology: Applied, 12,* 179–195.

Diagnostic Process, Making a Diagnosis

The diagnostic process is central to clinical medicine. Patients come to a physician with complaints, and the physician attempts to identify the illnesses responsible for the complaints. The physician accomplishes this task by eliciting from patients their collection of signs (manifestations of the disease perceived by the physician, brought forth during the physical examination) and symptoms (manifestation of the disease perceived by the patient and brought forth during the history taking). Generally, physicians make treatment errors as the result of diagnostic errors. If the disease responsible for the complaints is correctly diagnosed, the correct treatment has a high probability of being prescribed. This makes good sense. Treatments can be looked up in reference materials; making the correct diagnosis is more complex.

Methods of Diagnosis

How physicians make a medical diagnosis has received considerable study and attention, although our understanding remains incomplete. Traditionally, physicians were thought to first systematically collect a complete clinical data set on the patient. This included the chief complaint, the history of present illness, the patient's complete past medical history, the patient's social history, a detailed family history, a comprehensive review of systems, and, finally, the complete physical exam. Only as a second and separate step were physicians thought to analyze the data and diagnose the responsible disease.

Despite the belief of some physicians that this method is central to diagnostic success, when psychologists study the process of diagnosis, they find that expert physicians do not blindly collect clinical information. In fact, expert physicians are often observed to collect less information than novice physicians when making diagnoses but are much more likely to make the correct diagnosis. While the novice physician collects a great deal of information, the novice can miss collecting the data needed to make the diagnosis. Expert physicians might be expert in knowing which data to collect as well as expert in knowing which data are irrelevant to the diagnostic task.

It appears that physicians use intuition, deliberate reasoning, or a combination of these two processes when engaged in making medical diagnoses. Many psychologists describe two different and complementary mental systems used by humans:

an automatic, experiential, recognition-based system and a rational, conscious, analytic system. The recognition-based system generates impressions of the attributes of objects of perception and thought. These impressions are not necessarily voluntary and often cannot be explained by the person. In contrast, the conscious, analytic system involves deliberate reasoning, which is much slower and effortful but more controlled. Although these two cognitive systems operate independently, there is reason to suspect that skilled diagnosticians learn to use their analytical brains to double-check their intuitive brains. In the discussion below, three distinct diagnostic methods are presented. These are presented as prototypes; physicians might use one of these three methods and a mixture of these methods when involved in medical diagnosis.

Pattern Recognition Method

In many studies, physicians appear to use pattern recognition when making diagnoses; that is, they make diagnoses based on sensory input and without deliberate or conscious analysis. This process is typically fast and accurate but difficult for the physician to explain. For example, when an experienced physician sees a psoriatic plaque on a patient's elbow, the physician will instantly diagnose the disease as psoriasis. Ask why it is psoriasis, one might observe a pause, and only after a few seconds will the physician come forth with the observation that the skin plaque is salmon-colored and covered with distinct, silvery scales. The explanation of the diagnosis takes considerably longer and more effort than does making the diagnosis itself!

Recognized patterns can be visual (the plaque of psoriasis), olfactory (the smell of an anaerobic infection), tactile (the hard, gritty feel of a cancer), or auditory (the confined speech of a patient with a peritonsillar abscess). Patterns can be learned through instruction or clinical experience, but novices frequently need a guiding mentor to point out clinically important patterns.

Pattern recognition has another interesting characteristic—the accuracy of a diagnosis is inversely correlated with the time it takes the physician to make the diagnosis. Thus, diagnoses that are made almost instantaneously are more likely to be correct than those made only after a more drawn-out review. This finding suggests that pattern recognition involves a cognitive process that is not based on deliberate reasoning.

Pattern recognition is a powerful and impressive tool when it works, but it also has weaknesses. The patient's signs and symptoms might resemble the patterns of two or more diseases. Again turning to the discussion of psoriasis, a physician might come across an isolated scalp lesion that is scaling, but it is not clear whether it is psoriasis or seborrheic dermatitis. In this situation, the physician needs to use more than pattern recognition. Physicians can also perceive specific patterns when they are not present, because visual pattern recognition appears to be influenced by nonvisual data. For example, researchers showed that they could manipulate the findings expert radiologists report on radiographs by manipulating the brief clinical histories that accompany each radiograph.

Prediction Rules Method

Another strategy used by physicians for arriving at a medical diagnosis is the prediction rule. When using a prediction rule, the clinician moves through a series of predetermined steps of an algorithm based on the presence or absence of clinical findings at branch points. Prediction rules can also be presented as mathematical functions that generate scores based on clinical findings. In contrast to pattern recognition, the use of prediction rules is slow, deliberate, effortful, and controlled.

Often, prediction rules are in the form of algorithms that are branching flow diagrams. Following a flow diagram does not require great domain knowledge. Therefore, this is a powerful method for physicians and other clinicians when they encounter a problem that they infrequently see and are unfamiliar with. As long as the algorithm is followed, the physician has a good chance of ending with the correct diagnosis.

Prediction rules can be stored mentally, on paper (now a prominent feature in review articles, textbooks, and practice protocols), or as Web pages. Some prediction rules used by physicians come from more expert physician colleagues and are transmitted via curbside consults. The systematic use of a prediction rule can improve physicians' diagnostic accuracy. For example, a short algorithm for diagnosing acute myocardial infarction in patients with

chest pain was shown to perform at least as well as third-year internal medicine residents.

Despite this power, prediction rules also have a weakness: If a sign or symptom is not in the algorithm, it can not be used in the diagnostic process. For example, egophony is a specific finding for pneumonia, but its sensitivity is low. In the effort to keep a prediction rule manageably simple so that it can be easily presented and followed, egophony would not be included in the rule. Therefore, the examiner using the prediction rule would not be prompted to look for this finding. The lean and mean prediction rule might be efficient and helpful but not particularly nuanced.

A second problem with using prediction rules to make diagnoses is that they force a physician to lumber through a series of decision steps that an expert would bypass because of a more expert way of approaching a diagnostic problem. A final problem is that while prediction rules are readily available to physicians, many have never been validated.

Hypothetico-Deductive Method

The hypothetico-deductive approach to medical diagnosis can be placed between the sudden insight of pattern recognition and the slow, deliberate movement through a prediction rule. This is the method of medical diagnosis probably most frequently used by physicians. The method involves rapidly generating a differential diagnosis based on limited information about the chief complaint and then deliberately collecting additional clinical information to assess the likelihood of the different diseases in the differential.

Imagine that a physician is seeing a middle-aged patient in the office and has collected some initial information about chest pain. Instead of recognizing a single diagnosis, he or she might have several competing hypotheses (the differential diagnosis) and will usually set off to collect specific pieces of data (e.g., location, duration, provokers and relievers of the pain) that increase the probability of one diagnosis and decrease the probabilities of others. The additional data can come from further questioning, maneuvers on the physical exam, the laboratory, or the radiology suite or by using time to observe the change in signs and symptoms.

Central to the hypothetico-deductive method is the differential diagnosis. This list of competing diagnoses is typically short, usually three to five diseases, and is formulated by the physician early in the clinical encounter, usually within the first few minutes. Physicians spend much of their time during the patient visit collecting data to evaluate these different diagnostic possibilities. Data not pertinent to the differential are not collected by the expert clinician because it is not relevant to the task at hand.

Physicians use a number of approaches to judge the probabilities for the different diseases in a differential. One approach is to look up this information in a medical reference, although this information is often unavailable or difficult to find. A more commonly used approach is that of employing heuristics—simple, efficient rules, which are either hard-coded or learned. These work well under many circumstances but in certain situations are linked to systematic cognitive biases. These biases have gained a great deal of notoriety.

One frequently used heuristic is based on availability—diseases that come more easily to the physician's mind after learning the initial symptoms are taken to be more probable. This strategy is taught as part of the informal curriculum in most medical schools, transmitted through aphorisms such as "If you hear hoof beats, think horses, not zebras." However, attributes other than greater likelihood might make a disease come easily to mind. For example, when a physician has recently attended a talk about an uncommon disease, it might be more mentally available to him or her when he or she next goes to see patients. If a physician errs by missing an important diagnosis, the disease will often easily come to mind when encountering a future patient with similar symptoms. If the physician has recently been diagnosed with a disease, it may more readily come to mind when the physician is evaluating patients.

A second commonly used heuristic is basing the probability of a disease on the representativeness of the symptoms. Using this heuristic, a physician will estimate the probability that a person with symptom complex *A*, *B*, and *C* has disease *X* by judging the degree to which the complex of symptoms *A*, *B*, and *C* is representative or typical of disease *X*. This heuristic is often taught to medical students using the aphorism "If it walks like a duck and quacks like a duck, it probably is a duck." However, this heuristic ignores the underlying

probably of a disease in a population. Because it ignores base rates, this heuristic can lead to errors in probability estimation. For example, a middle-aged woman who recently developed truncal obesity, excess sweating, telangiectasia, and hypertension might fit our image of Cushing's disease; but this is an uncommon disease. Despite the close match to our profile of Cushing's disease, simple obesity accompanied with hypertension is much more common and therefore a much more likely diagnosis.

The hypothetico-deductive method not only requires that physicians determine the correct prior probability for a disease, but physicians must also correctly revise this probability given the additional information uncovered during the clinical evaluation. Bayes's theorem is a normative standard by which intuitive probability revision can be assessed. Researchers have raised considerable doubt about physicians' abilities to intuitively revise this probability after gathering new information. Using Bayes's theorem as a comparison, physicians have been shown to badly err when asked about the effect of new information on the likelihood of a disease. More recently, other researchers have suggested that physicians are quite skilled at probability revision. They suggest, however, that the format in which physicians are provided with the information about probabilities is a major determination of whether they revise probabilities in a way consistent with Bayes's theorem. These researchers suggest that physicians do poorly with likelihood information in the format of probabilities but perform well when the information is in the format of natural frequencies.

George Bergus

See also Cognitive Psychology and Processes; Decision Psychology; Errors in Clinical Reasoning; Hypothesis Testing; Judgment; Learning and Memory in Medical Training; Pattern Recognition; Problem Solving; Teaching Diagnostic Clinical Reasoning

Further Readings

Ark, T. K., Brooks, L. R., & Eva, K. W. (2006). Giving learners the best of both worlds: Do clinical teachers need to guard against teaching pattern recognition to novices? *Academic Medicine, 81,* 405–409.

Brooks, L. R., LeBlanc, V. R., & Norman, G. R. (2000). On the difficulty of noticing obvious features in patient appearance. *Psychological Science, 11*(2), 112–117.

Elstein, A. S., & Schwartz, A. (2002). Clinical problem solving and diagnostic decision making: Selective review of the cognitive literature. *British Medical Journal,* 324(7339), 729–732.

Eva, K. W., Hatala, R. M., Leblanc, V. R., & Brooks, L. R. (2007). Teaching from the clinical reasoning literature: Combined reasoning strategies help novice diagnosticians overcome misleading information. *Medical Education, 41*(12), 1152–1158.

Gigerenzer, G. (1999). *Simple heuristics that make us smart.* Oxford, UK: Oxford University Press.

Gilovich, T., Griffin, D., & Kahneman, D. (Eds.). (2002). *Heuristics and biases: The psychology of intuitive judgment.* New York: Cambridge University Press.

Hogath, R. (2001). *Educating intuition.* Chicago: University of Chicago Press.

Kahneman, D. (2002). *Maps of bounded rationality* (The Sveriges Riksbank Prize Lecture in Economic Sciences in Memory of Alfred Nobel 2002). Retrieved February 10, 2009, from http://nobelprize.org/nobel_prizes/economics/laureates/2002/kahneman-lecture.html

Norman, G. (2005). Research in clinical reasoning: Past history and current trends. *Medical Education, 39,* 418–427.

Rikers, R. M., Schmidt, H. G., Boshuizen, H. P., Linssen, G. C., Wesseling, G., & Paas, F. G. (2002). The robustness of medical expertise: Clinical case processing by medical experts and subexperts. *The American Journal of Psychology, 115*(4), 609–629.

Diagnostic Tests

Numerous diagnostic tests exist that can provide information to guide medical decision making. In a broad sense, diagnostic tests include symptoms and signs (e.g., chest pain, fatigue, varicose veins, ankle edema); measurements on physical examination (e.g., height, weight, blood pressure); special measurements (e.g., ankle-brachial pressure index, electrocardiogram [ECG], electroencephalogram [EEG]); blood tests (e.g., cholesterol, lipid profile, glucose); cytology and histology (e.g., Papanicolaou smears, biopsy); and imaging tests (e.g., endoscopy, ultrasound, computerized tomography [CT], magnetic resonance imaging [MRI],

single photon emission computed tomography [SPECT], positron-emission tomography [PET]).

Tests results can be dichotomous—that is, the result is either positive or negative, or the test may have multiple possible results on a categorical, ordinal, or continuous scale. Interpreting information obtained from diagnostic tests correctly is key in optimizing medical decision making.

Tests With Two Results, Positive Versus Negative

A test result is said to be "positive" if it shows a particular finding to be present and "negative" if the finding is absent. Note that a positive test result suggests that a patient has the disease in question—which is usually not a positive thing for the patient—and vice versa.

Most diagnostic information is not perfect but rather subject to some degree of error. A positive test result may be

- *true positive (TP)*—the test result indicates disease, and the patient has the disease, or
- *false positive (FP)*—the test result indicates disease, but the patient does not have the disease.

A negative test result may be

- *true negative (TN)*—the test result indicates no disease, and the patient has no disease, or
- *false negative (FN)*—the test result indicates no disease, but the patient has the disease.

Whether a patient has the disease or not is determined by the "truth" as established by a reference (gold) standard test, which is generally an invasive and/or expensive test and one that many patients would like to avoid. The occurrence of false-positive (FP) and false-negative (FN) test results implies that medical professionals need to be careful in interpreting diagnostic test information to minimize the impact of such errors.

Diagnostic performance (also referred to as accuracy or validity) of a test is its correspondence with the underlying truth and is expressed using the test's characteristics, sensitivity, and specificity. Alternatively, the diagnostic test performance may be characterized with true- and false-positive ratios, which is particularly convenient when a test has more than two possible results. Sensitivity and specificity describe how often the test is correct in the diseased and nondiseased groups, respectively. True- and false-positive ratios describe how often the test yields a positive result in the diseased and nondiseased groups, respectively.

Sensitivity, or *true-positive ratio* (TPR), is the probability of a positive test result given that the disease is present, denoted by $p(T+|D+)$. *Specificity*, or *true-negative ratio* (TNR), is the probability of a negative test result given that the disease is absent, denoted by $p(T-|D-)$. The *false-negative ratio* (FNR) is the complement of sensitivity, that is, 1.0 − TPR, and is the proportion of patients with disease who have a negative test result, denoted by $p(T-|D+)$. The *false-positive ratio* (FPR) is the complement of specificity, that is, 1.0 − TNR, and is the proportion of patients without disease who have a positive test result, denoted by $p(T+|D-)$.

Algebraically, these can be summarized as follows:

- Sensitivity = $p(T+|D+)$ = TPR = TP/(TP+FN).
- Specificity = $p(T-|D-)$ = TNR = TN/(TN+FP).
- 1 − Sensitivity = $p(T-|D+)$ = FNR = FN/(TP+FN).
- 1 − Specificity = $p(T+|D-)$ = FPR = FP/(TN+FP).

There is an analogy between diagnostic tests and research studies. The FPR is the rate of Type I errors (α value) that are errors of commission: We are saying there is a finding that is in fact not there. The FNR is the rate of Type II errors ($1 - \beta$ value) that are errors of omission: We omit to identify the finding.

Although sensitivity and specificity are important characteristics of a test, they are not the conditional probabilities required to decide how to treat a patient. Sensitivity and specificity are the probabilities of test results conditional on the presence versus absence of disease. In practice, medical professionals do not know whether or not someone has the disease, but rather, they find a test result is positive or negative, and from this information, they infer the probability of disease. Thus, medical professionals usually need to know the probabilities of disease given positive or negative test results, which are very different. Posttest revised (or posterior) probabilities are defined as follows:

- The *post-positive-test probability of disease*, or *positive predictive value (PPV)*, is the conditional probability of disease given a positive test result, $p(D+|T+)$—that is, the probability that a patient with a positive test result has the disease.
- The *post-negative-test probability of disease* is the conditional probability of having the disease given a negative test result, $p(D+|T-)$—that is, the probability that in spite of a negative test result, the patient does have the disease.
- The *post-positive-test probability of absence of disease* is the conditional probability of absence of the disease given a positive test result, $p(D-|T+)$—that is, the probability that in spite of a positive test result, the patient does not have the disease.
- The *post-negative-test probability of absence of disease*, or *negative predictive value (NPV)*, is the conditional probability of not having the disease given a negative test result, $p(D-|T-)$—that is, the probability that a patient with a negative test result does not have the disease.

If the number of TP, FN, FP, and TNs in the population is known, then these probabilities can be calculated as follows:

- Post-positive-test probability of disease = $p(D+|T+) = TP/(TP + FP) = PPV$.
- Post-negative-test probability of disease = $p(D+|T-) = FN/(TN + FN)$.
- Post-positive-test probability of absence of disease = $p(D-|T+) = FP/(TP + FP)$.
- Post-negative-test probability of absence of disease = $p(D-|T-) = TN/(TN + FN) = NPV$.

Estimates of probabilities of disease conditional on test results are not readily available, and if they are available, they are highly influenced by the pretest (prior) probability of the disease in the patient population studied. Sensitivity and specificity values are, however, generally available and under certain conditions can be transferred from one population to another in spite of a different prior probability because they are conditional on disease status. Converting the probabilities of test results given the disease to probabilities of disease given the test results is done with Bayes's theorem.

Tests With Multiple Results

Many tests have multiple possible test results, which may be on a categorical, ordinal, or continuous scale. In the setting of multiple test results, diagnostic test performance is best characterized with true- and false-positive ratios of each of the test results and the corresponding likelihood ratio. The likelihood ratio (LR) for test result R is the ratio of the conditional probability of R given the disease under consideration to the probability of R given absence of the disease under consideration. The LR summarizes all the information medical professionals need to know about the test result R. A high LR indicates that the test result argues in support of the diagnosis. A low LR indicates that the test result argues against the diagnosis.

In the setting of multiple test results, medical professionals frequently need to choose a cut-off value that defines a positive test result that requires treatment or further workup versus a negative result that does not require further action. Shifting the chosen cut-off value will yield pairs of FPR and TPR rates that together give the receiver operating characteristic (ROC) curve of the test.

M. G. Myriam Hunink

See also Bayes's Theorem; Conditional Probability; Likelihood Ratio; Receiver Operating Characteristic (ROC) Curve

Further Readings

Hunink, M. G. M., Glasziou, P. P., Siegel, J. E., Weeks, J. C., Pliskin, J. S., Elstein, A. S., et al. (2001). *Decision making in health and medicine: Integrating evidence and values*. Cambridge, UK: Cambridge University Press.

Differential Diagnosis

The term *differential diagnosis* is generally thought of as both a noun and verb by clinicians. The noun form of differential diagnosis is the list of all possible conditions that could explain the collection of signs, symptoms, and test results observed in a particular patient at a particular point in time. This list of conditions is organized from most

likely (high on the list) to least likely (low on the list). The verb form of differential diagnosis is the medical decision-making process whereby this list is continually updated by eliminating conditions that are considered to be ruled out and adding conditions that may not have been previously considered based on the acquisition of new information. Conditions that remain on the list are also moved up and down in priority based on a continual reanalysis of their likelihood. The goal of diagnostic investigation and problem solving is the elimination of all conditions from the differential diagnosis until a single unifying diagnosis remains.

Casting a Broad Net

The number of conditions contained in a differential diagnosis is referred to as its breadth. Generally, the smaller the number of signs, symptoms, and test results available for consideration, the broader the differential diagnosis. A clinician always has the least amount of information available at the time of the patient's initial presentation, and so it behooves him or her to "cast a broad net" by adding many conditions to the list even if they are only remote possibilities. An initial broad differential can be winnowed down later using additional information obtained during the course of further diagnostic investigation. The choice of which conditions to add occurs by pattern recognition, whereby clinicians recognize patterns of signs and symptoms present in disease states that they have seen before. The process of recognizing these patterns, identifying the condition, and adding it to the differential diagnosis is referred to as hypothesis generation.

Accurate pattern recognition and hypothesis generation are the foundation of accurate differential diagnosis, because if a condition does not make it onto the list of differential diagnoses, it can never be confirmed or refuted via further investigation. Research has shown that the source of the improved accuracy of expert diagnosticians is not better acquisition of the signs and symptoms that provide the data set for hypothesis generation, nor is it the number of hypotheses generated from a given data set, but instead, it is the generation of more accurate hypotheses compared with novice diagnosticians. The source of this improved accuracy has been the topic of much debate. However, a rule of thumb that is used frequently by clinicians to describe hypothesis generation is "common things are common." This seemingly obvious adage means that a given sign or symptom is more likely to be an uncommon manifestation of a common disease than a common manifestation of an uncommon disease. In other words, one should focus on generating hypotheses that are epidemiologically most likely even if they do not seem to fit the pattern perfectly. Recall or availability bias is a type of cognitive error in this process wherein the clinician has a distorted sense of the prevalence of a particular condition based on his or her own personal experience rather than that reported in the scientific literature.

While the amount of data available to the diagnostician is the principal determinant of the breadth of an initial differential diagnosis, the particular characteristics and the quality of the data being considered can also have a profound effect. Certain findings are considered pathognomonic for particular diseases, meaning that the finding is so specific and sensitive that a patient should be considered to have the condition until proven otherwise. An example would be the presence of Kaiser-Fleischer rings in the eyes of patients with Wilson's disease. This single observation on the physical exam would eliminate nearly all other conditions from consideration. Similarly, the quality of the data also has dramatic effects on the breadth of the differential. Demented, mentally ill, or malingering patients may supply a wealth of historical details; however, the reliability of this information would remain suspect, and it might add little value despite its abundance. In these situations, the differential would remain broad despite obtaining a relatively large amount of data.

Narrowing the Differential Diagnosis

Once a broad differential has been established based on initial data gathering and hypothesis generation, the list is narrowed by either confirming (ruling in) a single diagnosis or eliminating (ruling out) conditions one by one until a single diagnosis remains. Usually, both approaches are used simultaneously. The order in which conditions in the differential are investigated depends on (a) the urgent or emergent nature of diagnoses on the list,

(b) the logistical expediency of obtaining a definitive answer for a particular diagnosis, and (c) the particular cognitive preferences of the diagnostician. Diagnoses that threaten loss of life or function are always investigated first even if they are low on the differential. Dissecting thoracic aortic aneurysm is a relatively rare cause of chest pain and is often near the bottom of the differential. However, it is investigated rapidly, as the consequences of a delayed diagnosis would be devastating. Once all the life-threatening diagnoses have been eliminated from the differential, diagnostic investigation can proceed at a more leisurely pace. If a condition can be excluded simply and easily, it is often pursued next. These are the so-called low-hanging fruit of the diagnostic process, and an example would be excluding a diagnosis of anemia with a simple complete blood count. In general, ruling in is a quicker way to narrow the differential than ruling out because one need only be correct once in the former approach and one needs to be correct $N - 1$ times (N being the number of conditions in the differential) in the latter.

Once a diagnosis has been ruled in, the remainder of the diagnoses are assumed to be ruled out based on the principle of parsimony, or Ockham's razor. The principle is attributed to the 14th-century logician William of Ockham and states that "the explanation of any phenomenon should make as few assumptions as possible, eliminating those that make no difference in the observable predictions of the explanatory hypothesis." Practically, this means that all of the observable signs, symptoms, and test results should be explained by a single diagnosis. If a single condition has crossed the threshold of evidence to be accepted as the unifying diagnosis, then all other diagnoses must be rejected. Even if a diagnosis has been confirmed, the particular cognitive preferences of a diagnostician will still factor into the ongoing investigation. Some diagnosticians may continue to rule out conditions as they prefer to "leave no stone unturned."

Cognitive Bias

All diagnosticians are subject to bias in the medical decision making involved in narrowing the differential diagnosis. Two common types of cognitive bias are confirmation bias and anchoring bias. Confirmation bias arises when a clinician only performs further testing in an effort to confirm a diagnosis that he or she already believes to be true and does not test other hypotheses that might refute the favored diagnosis. Anchoring bias is similar but distinct in that it results from a failure to add new diagnoses to the differential or adjust the position of old diagnoses based on new information. The clinician becomes anchored to the original differential and is blinded to new possibilities.

Negative Diagnostic Workups

"No evidence of disease is not evidence of no disease" is a phrase often used to describe the fact that a clinician's inability to detect a condition at a particular point in time does not mean that it is not present currently or was not present in the recent past. This is especially true for conditions that have waxing-and-waning courses, such as occult gastrointestinal bleeding. Commonly, 80% of upper gastrointestinal bleeding has stopped by the time of presentation to medical attention. Nonbleeding ulcers or varices are often found on esophagogastroduodenoscopy and presumed to be the source, but in a significant number of cases, the source of the bleeding cannot be found because active bleeding is no longer visible at the time of the diagnostic investigation. Failure to find a source of bleeding despite thorough investigation does not mean that gastrointestinal bleeding has been ruled out as a cause of the patient's presenting signs and symptoms, and consequently, it cannot be eliminated from the differential diagnosis.

When a workup is entirely negative and the differential diagnosis still contains more than a single diagnosis, watchful waiting is sometimes employed as a passive diagnostic strategy if the patient's condition is stable. The hope is that the condition causing the presenting symptoms will reactivate and new observations can be made at that time, which will allow the differential diagnosis to be narrowed.

When a workup is negative but the differential is relatively small and/or the patient's condition is deteriorating, a strategy of diagnostic and therapeutic intervention can be employed to confirm a diagnosis. If the therapy is narrowly directed at a particular diagnosis and the patient responds to treatment, the individual diagnosis in question is considered to be ruled in, and further diagnostic

workup is unnecessary. When this strategy is employed, it is important that a diagnostic response to treatment be defined clearly and prospectively and that only a single, narrowly directed therapy be used at any one time. If multiple therapies are employed simultaneously, a causal relationship between treatment and disease cannot reliably be inferred, and therefore, a diagnosis cannot be reliably confirmed based on response to treatment. This type of obfuscation of the differential diagnosis often occurs when broad-spectrum antibiotics are used to treat an infection of unclear etiology. The patient may have improvement in fever, white blood cell count, and bacteremia, but the signs and symptoms that would have helped localize the infection have not been allowed to develop.

The process of differential diagnosis is critical to medical decision making, because without an accurate diagnosis, decisions about treatment become extremely difficult. The medical decision making involved in differential diagnosis is complex and subject to the underlying cognitive biases of clinicians. Diagnostic testing is not without the potential to harm patients. Consequently, risk/benefit decisions must be made to determine whether the additional diagnostic information provided by a test or procedure is warranted. Skilled differential diagnosticians balance these risks with their degree of confidence that the correct single unifying diagnosis has been selected from the list of possibilities generated during the process of differential diagnosis.

Robert Patrick

See also Diagnostic Process, Making a Diagnosis; Errors in Clinical Reasoning; Heuristics

Further Readings

Adler, S. N., Adler-Klein, D., & Gasbarra, D. B. (2008). *A pocket manual of differential diagnosis* (4th ed.). Philadelphia: Lippincott Williams & Wilkins.

Barrows, H. S., & Feltovich, P. J. (1987). The clinical reasoning process. *Medical Education, 21*(2), 86–91.

Bordage, G. (1999). Why did I miss the diagnosis? Some cognitive explanations and educational implications. *Academic Medicine, 74*(Suppl. 10), S138–S143.

Elstein, A., Shulman, L., & Sprafka, S. (1978). *Medical problem solving: An analysis of clinical reasoning.* Cambridge, MA: Harvard University Press.

Eva, K. W. (2005). What every teacher needs to know about clinical reasoning. *Medical Education, 39*(1), 98–106.

Groves, M., O'Rourke, P., & Alexander, H. (2003). The clinical reasoning characteristics of diagnostic experts. *Medical Teacher, 25*(3), 308–313.

Norman, G., Young, M., & Brooks, L. (2007). Non-analytical models of clinical reasoning: The role of experience. *Medical Education, 41*(12), 1140–1145.

Disability-Adjusted Life Years (DALYs)

The disability-adjusted life year (DALY) measure combines nonfatal outcomes and mortality in a single summary measure of population health. One DALY represents 1 lost year of healthy life. The basic philosophy associated with the estimation of DALYs is (a) use the best available data, (b) make corrections for major known biases in available measurements to improve cross-population comparability, and (c) use internal consistency as a tool to improve the validity of epidemiological assessments. For the latter purpose, a software application, DISMOD II, is available from the World Health Organization (WHO) Web site.

Uses

DALYs were first employed in the 1993 *World Development Report* to quantify the burden of ill health in different regions of the world. The Global Burden of Disease (GBD) study, edited by Murray and Lopez and published in 1996, used a revised DALY measure. The DALY was developed to facilitate the inclusion of nonfatal health outcomes in debates on international health policy, which had often focused on child mortality, and to quantify the burden of disease using a measure that could also be used for cost-effectiveness analysis. DALYs have been widely used in global- and national-burden-of-disease studies and to assess disease control priorities. They have also been used to make the case for primary prevention programs for disorders such as stroke prevention in Australia and in assessing funding allocations in medical research programs in Australia, Canada, and the United States in relation to the burden associated with different diseases.

DALYs are also frequently used in economic evaluations of public health interventions, particularly in low- and middle-income countries. The DALY is the health effect measure that is recommended by the WHO's Choosing Interventions that are Cost Effective (WHO-CHOICE) program for generalized cost-effectiveness analysis (GCEA) and is also used in the World Bank's Disease Control Priorities in Developing Countries program. In GCEA, the costs and effectiveness of all possible interventions are compared with the null set for a group of related interventions to select the mix that maximizes health for the given resource constraints. DALYs and quality-adjusted life years (QALYs) are both health-adjusted life year (HALY) measures that use time as a common metric. QALYs were developed for the economic evaluation of clinical interventions and remain the dominant outcome measure used in cost-utility analyses that compare the costs and health effects of specific interventions using a preference-based measure of health. It is standard for cost-utility analyses using QALYs to subtract averted direct costs of care (cost offsets) from intervention costs to calculate the net cost of interventions used to calculate cost-effectiveness ratios, which can be negative. In contrast, most analyses that use DALYs do not calculate cost offsets, primarily because reliable information on such costs is extremely scarce in low- and middle-income countries.

Components

DALYs are composed of two components, years of life lost (YLL) due to premature death and years lived with disability (YLD) associated with nonfatal injuries and disease. YLL represents the stream of lost healthy life due to premature death at a particular age. It is calculated as the product of the number of deaths due to a specific cause and the years lost per death. YLD is calculated as the product of incidence of a specific cause and its average duration, multiplied by a disability or severity weight for that condition. Disability weights are assigned on a scale from 0 (representing *perfect health*) to 1 (representing *death*), in addition to an optional age-weighting parameter. The scale of DALY weights is inverted from that used to calculate QALYs. Consequently, when DALYs are used as the denominator in cost-effectiveness ratios, one refers to the cost per DALY averted as opposed to cost per QALY gained. Equivalently, the DALY is a health gap measure, whereas the QALY is a health gain measure. When different interventions are evaluated by some studies using DALYs and by others using QALYs, ranking interventions according to cost-effectiveness ratios may be possible even though there is no systematic formula for converting between the two measures, as long as the same approach is used in each study to calculate costs.

Weights

The GBD study derived DALY weights for 22 indicator conditions through a person trade-off (PTO) process, in which panels of health experts from various countries were asked to assess the expected relative burden of conditions in two trade-off exercises. In one exercise (PTO1), participants were asked to trade off extending the lives of different numbers of "healthy" people and people with a condition such as blindness. In the second exercise (PTO2), participants were asked to choose between prolonging life for 1 year for people with perfect health and restoring to perfect health a different number of people with the same condition used in PTO1. If the results of the PTO1 and PTO2 exercises differed, participants were required to individually reconcile their estimates in order to reach internal consistency using PTO1-PTO2 equivalence tables. Afterward, participants shared their PTO1 and PTO2 assessments through a deliberative group process in which participants were confronted with the implications of their choices and allowed to discuss the basis for their viewpoints, to reflect on the implications of their preferences, and to revise their assessments. Subsequently, DALY weights were derived for several hundred other conditions by comparison with the indicator conditions. The PTO exercises have been repeated in many countries and have generally yielded comparable weights, which supports the use of the same weights in different populations. Potential facilitator biases in the PTO valuation process can be reduced through the training of facilitators, and potential participant biases are minimized by the deliberative process and by replication across multiple groups of participants.

The standard DALY used in the GBD study is calculated using a 3% discount rate to calculate

present values and an age-weighting parameter (which is optional). Discounting of future benefits is standard practice in economic analysis, but the use of age weighting is more controversial. The age-weighting parameter gives greater weight to young-adult years, peaking at around age 20 years, than to years lived in childhood or older adulthood. It is also possible to calculate DALYs with discounting but without age weighting or with neither discounting nor age weighting (see Figure 1), as has been done, for example, in the Australian burden-of-disease study.

One distinctive feature of DALYs as estimated in the GBD study is the use of Standard Expected Years of Life Lost (SEYLL). To define the standard, the highest national life expectancy observed was used, 82.5 years for Japanese females and 80.0 years for Japanese males. The use of a standard life expectancy, regardless of local life expectancy, is to express the social value of people being equal regardless of country or location. For the calculation of DALYs in cost-effectiveness analyses, as opposed to burden-of-disease studies, national life expectancies are typically used.

"Disability," as used in DALYs, encompasses all nonfatal outcomes and aggregates various aspects of an individual's health such as mobility, anxiety, and pain. The calculation of YLD does not entail an empirical assessment of functional or activity limitations experienced by individuals with impairments, which is how disability is conventionally defined and measured. The DALY weights reflect the preferences regarding different disease/health states or impairments in relation to the societal "ideal" of good health. The health state valuations used to estimate the burden of disease in terms of DALYs lost do not represent the lived experience of any disability or health state or imply societal value of the person in a disability or health state. A relatively high DALY weight for a condition means that 1 year lived in that condition is less preferable than 1 year lived in health states with lower disability weights. For example, the disability weight of .43 for blindness implies that 1 year spent with blindness is preferable to 1 year with paraplegia (weight .57) and 1 year with paraplegia is preferable to 1 year with unremitting unipolar major depression (weight .76). Equivalently, these weights imply that 1 year of living in good health followed by death (1 year × [1.0 − 0.0 disability weight] = 1.0 healthy life year) is less preferable than 3 years of living with paraplegia followed by death (3 years × [1.0 − .57 disability weight] = 1.3 healthy years). Based on these weights, other things being equal, it is preferable to prevent or cure a case of paraplegia (weight .57) rather than a case of low back pain (weight .06) if the prevention or cure for each case would cost the same and there were not enough resources to do both.

In the GBD study, disability weights for selected conditions and sequelae were adjusted according to whether a person was assumed to have received medical treatment and whether the treatment was believed to decrease the severity of the condition. For example, the disability weight was .583 for patients with untreated bipolar disorder and .383 for bipolar patients whose condition improved due to the treatment but was not in remission. For most disabling conditions (e.g., spina bifida, limb loss, spinal cord injuries), disability weights reflected the assumption that no improvement in functioning occurred as the result of rehabilitation. Disability weights could also be modified to incorporate data on the effectiveness of rehabilitation therapies.

A major attraction for the use of DALYs in comparison with QALYs is that they provide a means of comparing the health impact of a wide range of medical conditions through the use of a standardized set of disability weights. However, additional sources of disability weight estimates are appearing. The Dutch Disability Weights study has provided additional estimates for disorders or sequelae that were not fully included in the GBD study, and these have been used in national burden-of-disease studies conducted in the Netherlands, Australia, and the United States. In particular, the Dutch Disability Weights study estimated disability weights stratified on the basis of disease stages and complications. For example, that study estimated a weight of .07 for Type 2 diabetes, with weights of increasing severity for complications, such as a weight of .17 for moderate vision loss and .43 for severe vision loss. To take one more example, the GBD study assigned a weight of .73 for adults with dementia, whereas the Australian and Dutch studies calculated weights of .27 for mild dementia (with impairments in daily activities of living),

Formulas for DALY calculations without discounting or age weighting

$DALY_i = YLL_i + YLD_i$

YLL_i = Number of deaths due to cause i * Years lost per death

YLD_i = Number of incident cases of cause i * Average $duration_i$ * DW_i

$DALY_i$ = Disability-adjusted life years due to cause i

YLL_i = Years of life lost due to cause i

YLD_i = Years lived with disability due to cause i

DW_i = Disability weight for cause i

Example using individual-level data (for population data, incidence would be used):

Motor vehicle collision results in two fatalities and two injuries

A 55-year-old woman dies, resulting in 29.37 standard expected years of life lost (SEYLLs)

A 60-year-old man dies, resulting in 21.81 SEYLLs

Total YLL = 51.18 (without discounting or age weighting)

A 35-year-old woman gets a fractured skull, for which she is treated, but the effects are lifelong. The duration is equal to 48.38 SEYLLs, with a disability weight of .35. YLD for this injury is 48.38 * .35 = 16.933

A 40-year-old man treated for fractured sternum. The average duration is .115 years, with a disability weight of .199. YLD for this condition is .115 * .199 = .022885

Total YLD = 16.95589 (without discounting or age weighting)

Total DALY loss = 68.13589 (without discounting or age weighting)

Figure 1 How disability-adjusted life years (DALYs) are calculated

Source: The SEYLL and DW estimates were taken from Murray and Lopez (1996), in which discounting and age weighting were used in the estimation of DALYs.

.63 for moderate dementia (unable to live independently), and .94 for severe dementia (requiring permanent supervision). Future empirical studies may provide still more detail and better reflect the heterogeneity among health conditions. Currently, efforts are being undertaken to update disability weights for DALYs both globally and in the United States, to address the relevance of the social values that have been incorporated in the calculation of DALYs, and to assess changes in weights due to new developments in treatments for various diseases and conditions.

Cost-Effectiveness Ratios

The use of fixed thresholds for cost-effectiveness ratios to conclude that a particular intervention is or is not cost-effective is widespread but still controversial. Because of the interaction between cost-effectiveness, disease burden, and available resources, a single threshold for maximum cost per health gain cannot be specified. Nonetheless, a consensus has emerged that an intervention with a cost-effectiveness ratio less than three times the per capita gross domestic product (GDP) in a given country can be considered cost-effective, and one with a cost-effectiveness ratio less than one time the GDP per capita is "very cost-effective." This does not mean that clinical interventions with higher cost-effectiveness ratios do not provide good value but that more health gains could be achieved by prioritizing funding to interventions with lower cost-effectiveness ratios, which is the rationale for the Disease Control Priorities in Developing Countries program. However, even cost-effective interventions may not be feasible to implement if the costs are monetary and come from a public budget and the

benefits are nonmonetary and diffused over the population.

Scott D. Grosse and Armineh Zohrabian

Authors' Note: Authors have contributed equally and are listed alphabetically. The findings and conclusions in this article are those of the author and do not necessarily represent the official position of the Centers for Disease Control and Prevention.

See also Cost-Effectiveness Analysis; Cost-Utility Analysis; Person Trade-Off; Quality-Adjusted Life Years (QALYs)

Further Readings

Gold, M. R., Stevenson, D., & Fryback, D. G. (2002). HALYs and QALYs and DALYs, oh my: Similarities and differences in summary measures of population health. *Annual Review of Public Health, 23,* 115–134.

Grosse, S. D., Lollar, D. J., Campbell, V. A., & Chamie, M. (in press). Disability and DALYs: Not the same. *Public Health Reports*.

Jamison, D. T., Breman, J. G., Measham, A. R., Alleyne, G., Claeson, M., Evans, D. B., et al. (Eds.). (2006). *Disease control priorities in developing countries* (2nd ed.). New York: Oxford University Press (published for the World Bank).

Mathers, C. D., Vos, T., Lopez, A. D., Salomon, J., & Ezzati, M. (Eds.). *National burden of disease studies: A practical guide* (2nd ed.). Geneva: World Health Organization. Retrieved February 10, 2009, from http://www.who.int/healthinfo/nationalburdenofdiseasemanual.pdf

Mathers, C., Vos, T., & Stevenson, C. (1999). *The burden of disease and injury in Australia*. Canberra, Australian Capital Territory, Australia: Australian Institute of Health and Welfare. Retrieved February 10, 2009, from http://www.aihw.gov.au/publications/phe/bdia/bdia.pdf

McKenna, M. T., Michaud, C. M., Murray, C. J. L., & Marks, J. S. (2005). Assessing the burden of disease in the United States using disability-adjusted life years. *American Journal of Preventive Medicine, 28,* 415–423.

Murray, C. J. L., & Lopez, A. D. (Eds.). (1996). *The global burden of disease: A comprehensive assessment of mortality and disability from diseases, injuries, and risk factors in 1990 and projected to 2020*. Cambridge, MA: Harvard University Press.

Stouthard, M. E. A., Essink-Bot, M. L., Bonsel, G. J., & Dutch Disability Weights Group. (2000). Disability weights for diseases: A modified protocol and results for a Western European region. *European Journal of Public Health, 10,* 24–30.

World Health Organization. (2003). *Making choices in health: WHO guide to cost-effectiveness analysis* (T. Tan-Torres Edejer, R. Baltusse, T. Adam, A. Hutubessy, D. B. Acharya, D. B. Evans, et al., Eds.). Geneva: Author. Retrieved February 10, 2009, from http://www.who.int/choice/publications/p_2003_generalised_cea.pdf

World Health Organization—software tools: http://www.who.int/healthinfo/global_burden_disease/tools_software/en/index.html

DISCOUNTING

Why does a person engage in behaviors, such as eating high-calorie foods or keeping a sedentary lifestyle, that provide an immediate reward over behaviors that offer health benefits in the long run? Understanding the time dimension of health preferences, or intertemporal health preferences, has been an important area of inquiry for medical decision making. The concept of discounting over time has been central to this understanding. Time discounting of preferences refers to the common situation where money, goods, services, and other outcomes are more highly valued when obtained in the present than those occurring in the future. When all things are equal, a given reward is more desirable when obtained sooner than later. The section below provides an introduction to discounting in intertemporal choices and discusses the importance of these concepts in medical decision making.

Preferences for Early Versus Late Rewards

The question as to why money, goods, services, and health are more desirable in the present than in the future has been answered in several ways. Money and some goods can increase in value with time, so that it is better to obtain them in the present to obtain future growth. Having $100 now allows one to invest it and accrue interest over time. Waiting a year to receive $100 means that a year of interest is lost. Also, waiting for a reward may increase the risk of losing it in the future, so it

is better to have it in the present. For example, a person may choose to spend money on a vacation this year over investing in a retirement fund that would allow a vacation in the future. Waiting introduces the risk that one may not be healthy enough to enjoy a vacation during retirement. In medical decision making, the concept of discounting provides a way to understand why people engage in behaviors, such as smoking, that provide immediate gratification but that may contribute to risks to health in the future.

Discounting in Health and Medicine

Consideration of the value of health outcomes over time is critical to decision analysis and in analyses of the costs and benefits of preventive health regimens, diagnostic tests, and medical treatments. The results of decision analyses and cost-and-benefit analyses are important considerations in the development of policy on healthcare. In these types of analyses, discounting rates are used to provide an adjustment of the present value of an outcome for the costs and benefits occurring at different time points. While a variety of discounting rates have been used in these studies to estimate the value of future outcomes, the U.S. Preventive Service Task Force suggests the use of a 3% discount rate for cost-and-benefit analyses with a rate of 5% used for sensitivity analyses. However, discount rates have been shown to vary in studies of time preferences, and higher rates of 40% and 50% have been observed. The selection of discount rates for decision and economic analyses should depend on whether the interest is in group preferences or individual preferences. The lower rates may be reasonable to use for group analyses, but the higher rates may be appropriate to examine individual preferences.

Discounted Utility Theory

Discounted utility theory (DUT) has been used as a framework to understand preferences over time. Similar to expected utility theory (EUT), DUT is a normative decision model. Both models are based on the assumption that choices among alternatives depend on a weighted sum of utilities where decision makers seek to maximize the utility of their choices. While EUT describes preferences in situations of uncertainty, DUT describes preferences in the domain of time. DUT assumes a single discounting rate over time; the discounting rate serves as the utility weights in DUT. DUT also posits a single discount function that is exponential.

The axioms of DUT specify that preferences for outcomes over time are monotonic, complete, transitive, continuous, independent, and stationary. Monotonicity of preferences over time means that if an outcome is preferred at Time A over Time B, then Time A occurs before Time B. Thus, outcomes are more desirable if they occur earlier in time. Based on the propositions of DUT, the same discounting rate should be observed for all choices in time and should be positive in most cases. The axiom of completeness of preferences posits that there are preferences across different points in time. Transitivity of preferences over time means that if Outcome 1 is preferred to Outcome 2 at a later time and if Outcome 2 is preferred to Outcome 3 at a time that is still later, then Outcome 1 will be preferred over Outcome 3. Continuity of preferences assumes that there are points of indifference in preferences for outcomes between an earlier time and a later time, where outcomes are equally preferred. This axiom ensures that there exists a continuous utility function over time. The axiom of independence over time means that the order of preferences for outcomes should not reverse at different points in time. If one outcome is preferred to another at one time, this order of preferences should be preserved over time. Stationarity requires that when preferences are ranked across time, this ranking should not change even if the time interval changes.

Sign Effect

While DUT has proved to be useful in describing intertemporal preferences in a range of situations, violations in the axioms have been observed. For example, based on the assumption of a single discounting rate, DUT would suggest that preferences should be equal over time whether the health outcome is a gain or a loss. However, a number of studies provide evidence that the discount rate for losses is lower than that for gains. In other words, preference for a desirable outcome is discounted more over time than preference to avoid a loss. This has been termed the *sign effect*.

The Value of Health Versus Money

The idea that a single discounting rate can be used to describe preferences in all decisions also has been challenged. These arguments are especially important to medical decision making, where health is the desired outcome rather than money. Both money and health have been found to have relatively large discount rates, especially as compared with what is recommended for use in economic analyses. In contrast to what is predicted by DUT, decision makers appear to use different rates for health as compared with money. This observation has been used to understand why it can be difficult to encourage people to adopt preventive health behaviors to improve future health. A large discounting rate would mean that future health does not seem attractive enough in the present to overcome the desire to engage in behavior that is highly rewarding in the short term, such as smoking.

Choice Sequences

Another violation of DUT has been described with respect to a series of decisions made over time. For example, many health decisions occur in a sequence rather than as single choices. A person diagnosed with cancer may make a series of decisions about surgery, radiation therapy, and adjuvant chemotherapy. A person who is diagnosed with diabetes may face a series of choices about diet, exercise, medication, and self-monitoring. The sequence effect refers to the tendency to observe a negative discount rate when choices occur in a sequence. In other words, people prefer to defer desirable outcomes, to savor the rewards, and to want to hasten undesirable outcomes in order to get them out of the way sooner and reduce dread of an adverse event.

Hyperbolic Discounting

Alternative theories to DUT have been suggested to explain these anomalies. For example, it has been suggested that, as compared with the constant discounting posited by DUT, a hyperbolic model may better describe the preferences reversals over time. A hyperbolic discounting model describes preferences where delayed outcomes are discounted in a way that is inversely related to the time delay between the early review and the late review. Thus, short-term outcomes are discounted more than long-term outcomes. This could happen if the decision maker is more impatient in making judgments about reviews in the short run than in the long run. This might describe the case where a smoker has greater difficulty deferring a cigarette in the short run than in deferring the purchase of cigarettes in the long run.

The Neurobiology of Intertemporal Preferences

Recent work on discounting has been directed toward understanding the neurobiology of intertemporal preferences. These studies often employ functional magnetic resonance imaging to examine the brain activity of research participants who are engaged in a choice experiment. Results of these studies have described two systems relevant to making choices over time. In making intertemporal decisions, humans show several cognitive processes: ones that focus on the present and others that consider the future. These findings—that there may be several cognitive processes that distinguish between events in time—provide some support for the hyperbolic discounting models. These studies, while not conclusive, have offered innovative methods to more fully understand the processes underlying intertemporal choice.

Sara J. Knight

See also Cost-Effectiveness Analysis; Cost-Effectiveness Ratio; Cost-Utility Analysis; Marginal or Incremental Analysis

Further Readings

Cairns, J. A., & Van Der Pol, M. M. (1997). Saving future lives: A comparison of three discounting models. *Health Economics, 6,* 341–350.

Chapman, G., & Elstein, A. E. (1995). Valuing the future: Temporal discounting of health and money. *Medical Decision Making, 15,* 373–386.

Kalenscher, T., & Pennartz, C. M. A. (2008). Is a bird in the hand worth two in the future? The neuroeconomics of intertemporal decision-making. *Progress in Neurobiology, 84,* 284–315.

Kamlet, M. S. (1992). *The comparative benefits modeling project: A framework for cost-utility analysis of*

government health care programs. Washington, DC: Office of Disease Prevention and Health Promotion, Public Health Service.

Khwaja, A., Silverman, D., & Sloan, F. (2007). Time preference, time discounting, and smoking decisions. *Health Economics, 26,* 927–949.

Loewenstein, G., & Prelec, D. (1991). Negative time preference. *American Economic Review, 81,* 347–352.

Loewenstein, G., & Prelec, D. (1992). Anomalies in intertemporal choice: Evidence and interpretation. *Quarterly Journal of Economics, 107,* 573–597.

Loewenstein, G., & Prelec, D. (1993). Preferences for sequences of outcomes. *Psychological Review, 100,* 91–108.

McClure, S. M., Laibson, D. I., Loewenstein, G., & Cohen, J. D. (2004). Separate neural systems value immediate and delayed monetary rewards. *Science, 306,* 503–507.

Redelmeier, D. A., & Heller, D. N. (1993). Time preferences in medical decision making and cost-effectiveness analysis. *Medical Decision Making, 13,* 212–217.

Discrete Choice

A discrete choice experiment (DCE) is a type of stated preference method used to elicit values for goods and services. DCEs rely on the premise that any good or service can be described by its characteristics and the extent to which an individual values a good or service depends on the levels of these characteristics. DCEs have long been used by consumer products companies to design new products to meet customer preferences by measuring the relative importance of different product attributes, but they have only more recently been applied in the context of health and environmental goods. This approach typically provides more detailed, yet substantively different, information compared with traditional stated preference methods, such as contingent valuation or health state utility assessment.

Comparison With Other Stated Preference Methods

Stated preference methods, in which respondents value hypothetical descriptions of products or choices, are useful in valuing nonmarket goods, such as health. They are useful in situations in which the market for a good, or for the full range of attributes of a good, does not exist and "revealed preference" studies cannot be conducted. In a revealed preference study, preferences are estimated by observing the actual choices that have been made in a real-world setting. For example, the relative value of individual attributes of automobiles, such as size, color, make, and model, could be measured by analyzing retrospective data on automobile sales prices along with the specific characteristics of the automobiles sold. For a new model, a revealed preference approach would not be possible since data are not yet available for the new model; instead, a stated preference approach could be used. Other stated preference methods typically used to value health outcomes are health state utility assessment or contingent valuation. Elicitation techniques for these stated preference methods include standard gamble, time trade-off, or willingness to pay. Compared with these methods, DCEs can be used to value health, nonhealth, and process attributes and provide information about the trade-offs between these attributes. DCEs can also be used to value willingness to pay for an attribute, whereas traditional methods provide a single numerical rating for the whole service.

All stated preference methods have the limitation that the valuation task asks about hypothetical choices and, therefore, may not fully predict future choices. Using stated preference methods can often provide a valuable starting point for further research given the difficulty of obtaining preference data on nonmarket goods. All stated preference methods allow data to be collected on programs and interventions while they are still under development, similar to how studies might be conducted to develop new consumer products. Once a program or intervention has been introduced, additional research could combine revealed and stated preference data to provide even more detailed information about user preferences.

Understanding preferences for different aspects of health and health interventions and incorporating these values into clinical and policy decisions can result in clinical and policy decisions that better reflect individuals' preferences and potentially improve adherence to clinical treatments or public health programs.

Terminology

The terms *discrete choice experiments* or *conjoint analysis* are typically used to describe a type of stated preference method in which preferences are inferred according to responses to hypothetical scenarios. These terms are often used interchangeably. Conjoint analysis comes from marketing applications and DCEs from transportation and engineering applications. The common element of DCEs and conjoint analysis is that they both allow the researcher to examine the trade-offs that people make for each attribute, attribute level, and combinations of attributes. They differ in that the term *conjoint analysis* is more generally used to refer to a method whereby the respondent rates or ranks a scenario and DCEs involve a discrete choice between alternative scenarios. DCEs, and the related approach of conjoint analysis, have been successfully applied to measuring preferences for a diverse range of health applications, and the use of these approaches is growing rapidly.

Example of a Discrete Choice Experiment

An example of the attributes used in a DCE designed to identify preferences for a pharmacogenetic testing service is shown in Table 1. The service offers a test to identify a person's risk of developing a side effect (neutropenia) from azathioprine. This example has five attributes, and the attributes have different numbers of levels. Both health and nonhealth outcomes are included in the evaluation of the service.

Table 2 shows an example of one choice question. It is also possible to design discrete choices with more than two options.

Conducting a Discrete Choice Experiment

Conducting a DCE includes proper design, fielding, and analysis. Elements of design include designing the experiment and overall survey development, as the survey should include survey questions in addition to the discrete choice questions.

Designing and Administering a Discrete Choice Experiment

The first step in a DCE is to identify and define the attributes of the health intervention or program. Once the attributes or characteristics have been identified, the levels of each attribute must also be defined, which must be realistic options for the service being valued. Attributes and levels should be developed through an iterative process including literature review, experts in the field, focus groups, and one-on-one interviews.

The second step is to identify the choice task. Discrete choice task options include forced choice, in which the respondent chooses between one or more options. Alternatively, the respondent can be offered an opt-out option, which must then be addressed in the analysis step. The selection of the choice task will have implications about the type of analytic approach that is appropriate.

The third step is to set the experimental design for the DCE. Depending on the numbers of attributes and levels, it may be possible to use a full-factorial design in which all possible combinations of attributes and attribute levels are used to create scenarios. If the number of possible combinations exceeds the likely sample size, then efficient combinations of a subset of choices can be identified through the use of design libraries or other methods. This is called a fractional-factorial design, which uses mathematical properties to ensure nonassociation between the variables in the design (orthogonality). Choice sets must then be created from these scenarios, and again different methods exist, including pairing the scenarios or using fold-over techniques. It is important that key design principles are followed when creating the choice sets from the scenarios to ensure that the main effects—and, if necessary, two-way interactions—can be estimated.

The fourth step is to construct the survey that includes the DCE. A successful DCE survey will include an introductory section to provide the respondents with enough information to understand the choices they are about to be presented in the survey. This will involve a section that describes the attributes and levels and introduces the valuation task. This section should also include a practice question. Key questions for survey design will include mode of administration and sample selection. Mode of administration may determine the number of choices that can be included for each respondent. Depending on the numbers of attributes and attribute levels, choices may need to be divided into *choice sets*. A choice

Table 1 Possible attributes and levels

Attribute	*Level*
Process attributes	
The level of information given to the patient about the test	None Low Moderate High
How the sample is collected	Blood test Mouthwash Finger prick Mouth swab
Who explains the result to the patient	Primary-care physician Pharmacist Hospital physician Nurse
Cost	£0 £50 £100 £250
Health outcome	
The ability of the test to predict the risk of the side effect (neutropenia)	50% 60% 85% 90%
Nonhealth outcomes	
How long it takes before the patient receives a result	2 days 7 days 14 days 28 days

Table 2 Example of a pairwise choice

	Test A	*Test B*
The level of information given to the patient about the test	Moderate	High
The ability of the test to predict the risk of the side effect (neutropenia)	50% accurate	60% accurate
How the sample is collected	Finger prick	Mouth swab
How long it takes before the patient receives the result	28 days	2 days
Who explains the result to the patient	Pharmacist	Hospital doctor
Cost	$50	$250
Tick (✓) one option only	❑	❑

set is a fixed set of choices presented to a single respondent. For example, if the DCE has a total of 64 choices, then the choices may be split into 8 choice sets of 8 choices each to reduce respondent burden. An essential part of survey development is to pretest the questionnaire with respondents one-on-one until the survey instrument is stable. The survey should also include a section on respondent demographics and other characteristics that may relate to choices, such as experience with the health condition or intervention being valued.

Survey administration should follow the recommended approaches for the mode of administration involved. Most DCEs are administered via computer (online) or on paper via a mail survey.

Additional Design Considerations

Additional design considerations include the definition of the value attribute, inclusion of an opt-out option, and internal validity tests. The value attribute is the attribute used to infer value for the program or intervention; in health, it is typically represented by money or time. Other metrics can also be considered, such as risk, but to date there has been little research that explores how respondents value risk estimates in a DCE. The important characteristic of this attribute is that it is a continuous variable and can be analyzed as such. The inclusion of an opt-out option will be appropriate in any situation in which it would be a realistic option for the respondent in a real-world choice situation. In the survey design process, the development of attributes, levels, and choices should aim at keeping the hypothetical situations presented as close to reality as possible while still maintaining the objectives of the study.

Analysis of Discrete Choice Experiments

Analyzing data from a DCE requires the use of discrete choice analysis. The survey should have been designed to have an appropriate number of levels and attributes to produce robust estimates of the value for each attribute/level. For example, using the sample discrete choice question in Table 2, a utility function U is specified for each of the two alternatives of Test A or Test B:

$$U_i = \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \ldots + \beta_k X_{ki} + \varepsilon_i = \beta \mathbf{X}_i + \varepsilon_i, \quad (i = A, B).$$

X_{ki} ($k = 1, \ldots, K$) are the causal variables for alternative i and consist of both the attributes of the alternatives (e.g., effectiveness of the test). Further analyses can explore preferences in subgroups of the population (e.g., race/ethnicity, income). ε_i is the random error.

Assuming utility-maximizing behavior and ε_i iid (independent and identically distributed), extreme value distributed leads to the logit probability of choosing alternative i:

$$P(i) = \frac{exp(\beta X_i)}{\sum_{j=A,B,C} exp(\beta X_i)}.$$

The survey responses are used to make inferences on the coefficients β. This is performed by maximizing the log-likelihood of the sample over the unknown coefficients. Variables that should be statistically tested for inclusion in the model will include all attributes and all levels of each attribute; respondent characteristics such as race/ethnicity, age, sex, and income; appropriate interaction variables if these were included in the design. Other variables that could affect patient choices should also be considered as covariates in the analysis, such as a patient's familiarity with the service in question.

The sign of a statistically significant coefficient provides a decision maker with information about the effect of an attribute. A positive coefficient suggests that improved effectiveness as an attribute of the recommendation would make the program more attractive. Furthermore, the ratio between two coefficients can provide information about the trade-off, or marginal rate of substitution, between the two corresponding variables. For example, the ratio of the coefficient for benefit to the coefficient for cost represents the willingness to pay for a particular level of benefit. Results from a DCE survey can also quantify how an individual's value of effectiveness of an intervention compares with, for example, having a fast turnaround time. The results can provide additional information on which programs are likely to provide the most value to patients and clues on how to improve participation by aligning program characteristics with patient preferences.

In the past, the results of DCEs have been used to place the attributes in relative order of importance according to the size of the coefficient. Analysts have also attempted to compare the results

from DCEs conducted in two populations by directly comparing the size of the coefficients. Both these approaches are problematic because they do not make allowances for the scale parameter.

The analytic approach described above uses standard discrete choice methods. Advanced discrete choice methods can be used to both (a) improve the statistical efficiency of the coefficient estimates (by capturing serial correlations over multiple responses from the same individual) and (b) capture unobserved heterogeneity through estimation of random coefficients.

Future Research

Future research opportunities, some of which are under way, include methodological issues such as identifying optimal design sets, including risk as an attribute, understanding potential bias in using the cost attribute to estimate willingness to pay, estimating individual preferences, accounting for heterogeneity in preferences, and measuring the external validity of DCEs.

Lisa Prosser and Katherine Payne

See also Utility Assessment Techniques; Willingness to Pay

Further Readings

Bridges, J. F. P. (2003). Stated preference methods in health care evaluation: An emerging methodological paradigm in health economics. *Applied Health Economics and Health Policy, 2*(4), 213–224.

Elliott, R., & Payne, K. (2005). Valuing preferences. In *Essentials of Economic Evaluation in Healthcare* (Chap. 11). London: Pharmaceutical Press.

Hensher, D. A., Rose, J. M., & Greene, W. H. (2005). *Applied choice analysis: A primer*. Cambridge, UK: Cambridge University Press.

Lancsar, E., & Louviere, J. (2008). Conducting discrete choice experiments to inform healthcare decision making: A user's guide. *PharmacoEconomics, 26*(8), 661–677.

Ryan, M., & Gerard, K. (2003). Using discrete choice experiments to value health care programmes: Current practice and future research reflections. *Applied Health Economics and Health Policy, 2*(1), 55–64.

Street, D. J., & Burgess, L. (2007). *The construction of optimal stated choice experiments*. London: Wiley.

Discrete-Event Simulation

Discrete-event simulation (DES) is a very flexible modeling method that can be used when the research question involves competition for resources, distribution of resources, complex interactions between entities, or complex timing of events. The main advantage and disadvantage of DES is its large but constrained modeling vocabulary. That is, though there is more to learn initially, there is more freedom regarding the kinds of systems one can model.

DES was originally developed in the 1960s to model industrial and business processes, finding its first home in industrial engineering and operations research. Since then, DES has been used to gain insight into a wide range of research and business questions. Because of its unique strengths, DES began to be applied to healthcare problems in the mid-1980s.

Since its introduction, DES has been used to examine a broad array of healthcare and healthcare-related problems. Areas in which it has been applied have been mental health; disease management; infectious disease; disaster planning and bioterrorism; biology model and physiology; cancer; process redesign and optimization in laboratories, clinics, operating rooms, emergency services, healthcare systems, and pathways of care; geographic allocation of resources; trial design; policy evaluation; and survival modeling. DES is often the preferred simulation method in healthcare when (a) there is competition for resources, (b) systems are tightly coupled, (c) the geographic distribution of resources is important, (d) information or entity flow cannot be completely described a priori, (e) the timing of events may be asynchronous or cannot be modeled on a fixed clock, and (f) entities in the system require memory.

Simulation Modeling

In general, models allow researchers to explicitly explore the elements of a decision/problem and mediate understanding of the real world by rendering it comprehensible. Simulation modeling is any activity where the actual or proposed system is replaced by a functioning representation that approximates the same cause-and-effect relationship

of the "real" system. Simulation allows researchers to generate evidence for decision making or to develop understanding of underlying processes in the real world when direct experimentation (due to cost, time, or ethics) is not possible. Experimentation with simulation models is performed through sensitivity analyses, where the parameters of the system are varied, or through what-if experiments, where the number or types of resources of the system are varied.

Decision trees and Markov models have, to date, been the most common types of computer simulation models used in healthcare. These methods are used to create highly structured representations of decision processes and alternative strategies. This is done by constraining the formulation of these models to a limited vocabulary, essentially three building blocks—decision nodes, chance nodes, and outcome nodes. The main advantage of this type of formulation is that the highly structured format is relatively transparent and easy to interpret. The disadvantage is that the highly structured framework restricts the types or problems that can be articulated, often forcing significant compromises on the model and the modeler. With over 100 building blocks, DES has a much broader vocabulary (than tree models), allowing a broader array of problems to be modeled, with fewer compromises. This means that though there is more to learn initially there, is a greater range of problems one can model.

DES models differ from decision trees and Markov models in several ways. First, unlike tree models, DES allows entities within a system (e.g., patients) to interact and compete with each other. For example, two or more end-stage liver patients may be competing for a newly available donor liver. Second, DES allows for more flexible management of time than in tree models. Unlike simple trees, which handle time in the aggregate, or Markov models, which restrict changes in the system to fixed time intervals (Markov cycles), in DES, the time interval between events can be either fixed or treated as completely stochastic. In DES, each interaction provokes a change in the state of the system. Every interaction of entities with each other or with the resources in the system is an event. Every interaction changes the state of the entity involved and of the system as a whole. The time between events may be handled probabilistically, using fixed time increments, or both depending on the nature of the system being modeled.

There are generally four approaches for managing events in DES platforms: the *process interaction, event-scheduling, activity-scanning*, and *three-phase methods*. The differences are in how the software reacts to or anticipates interactions in the system. Third, every entity in the system can have memory. This means that the modeler can not only have entities interact but also can have the entities carry the memory of the interaction and have this information influence future interactions.

Key Features

The key features of a DES model are entities, attributes, queues, and resources. *Entities* are objects. They can move or be static within the system. They have the ability to interact with other entities. They represent persons, places, or things and so, metaphorically, act like nouns. The types of objects represented are not constrained to physical objects. For example, entities may also represent packages of information, such as phone calls, e-mails, or chemical signals. DES packages have been primarily written in object-oriented computer programming (OOP) languages, and entities may be considered to represent a class of objects.

Attributes are variables local to the entity object. This means that entities may carry information with them describing, for example, their age, sex, race, and health state, acting metaphorically as both the memory of the entity and as an adjective describing the entity. This information may be modified during any interaction within the system and may be used to determine how an entity will respond to a given set of circumstances. In DES, much of the information driving changes in the state of the model are embedded in the entities themselves in the form of attributes. This is in contrast to other modeling methods (e.g., trees, Markovs), where the information and knowledge are embedded in the nodal structure of the model. As a result, entities in DES have potentially many more degrees of freedom in how they transit the system being modeled.

A *resource* is an entity or object that provides a service to a dynamic entity. A service can be described as any activity requiring the simultaneous presence of the active entity. Providing a

service requires time. The number of entities a resource can serve simultaneously is the resource's capacity. For example, a bank with a single cashier can serve one person at a time. A bank with three tellers can serve up to three customers simultaneously. A mobile resource, such as a motorcycle, can transport 2 persons, whereas a school bus can transport 40 people. If a resource is occupied when a new entity seeks its use, the new entity must wait until the resource is free.

A *queue* is any place or list in which an entity waits for access to a resource. If an entity arrives seeking service and the resource is already occupied, it must wait somewhere. Queues have logic. For example, the line at a cashier may follow first-in/first-out (FIFO) logic, getting on or off an airplane may follow last-in/first-out (LIFO) logic, and the waiting room in an emergency department or the waiting list for a transplant may follow highest-value-first (HVF) logic. Queue theory is the mathematical study of queues or waiting lists.

Queue Theory

DES explicitly embeds queue theory. The simplest queuing model is the M/M/1 (Kendall's nomenclature), which translates as Markovian interarrival time/Markovian process time and one server, or M/G/1, which is Markovian interarrival time/general or arbitrary and one server. Simple systems such as these may be solved analytically and give insight into the behavior of more complex systems that cannot be analytically solved. This is important because every system from sufficient distance may be modeled as an M/M/1 system.

Interarrival rate is the rate of entity arrival (λ) ($1/\lambda$ = mean interarrival time). For example, the time between patient arrivals at a clinic may average 1 patient every 10 minutes. This may be stationary or nonstationary; for example, patients may arrive every 5 minutes around lunchtime. The service rate (μ) ($1/\mu$ = mean service time) is the rate at which the resource/server can process entities. The utilization rate (β) is λ/μ. If the average interarrival time = 10 minutes and the average service time = 7.5 minutes, the average utilization = .75. Another way to conceive of utilization is busy time/total time resource available. For example, if a nurse is busy 4 hours out of an 8-hour shift, then the utilization rate is .5. If the interarrival rate is less than the service rate (e.g., patients are arriving at longer intervals than the time required to process them), then the system is stable. If entities arrive faster than the system can process them, then waiting list length rises rapidly. Bottlenecks are temporary or permanent disequilibria between processing capacity and arrival rate at some point in the system—for example, a person calling in sick or an out-of-order elevator serving an apartment complex. Congestion occurs when a stable system has a utilization rate that is very close but slightly less than 1; that is, mean process time is very close to mean arrival rate (e.g., a tunnel or bridge into a major city). These are interesting systems because, first, they are very common and, second, they often experience large variations in behavior over time. Unexpected bottlenecks may occur randomly. These systems generally require longer run times to estimate expected system behavior. The breaks from normal behavior may be more interesting than the typical system behavior.

Flow time is the time from the moment an entity enters a system to the time the entity exits. The average flow time for a simple system may be described as $((\sigma^2_{server}\lambda + \beta^2\lambda)/2(1-\beta)) + \mu$,, where σ = standard deviation of process time, μ = mean process time, and λ = arrival rate. The average wait time is the flow time number minus μ. The average number in queue is $\beta^2(1+\mu^2\sigma^2_{server})/2(1-\beta)$.

Measures of Performance

In addition to the standard outputs, such as quality-adjusted life years and cost, DES also provides operational outcome measures, such as throughput, utilization, flow time, and wait time. *Flow time* is usually defined as time from entry into the system to time of exit. *Wait time* is usually defined as time from entry into the system to time of receipt of service. *Throughput* is usually defined as total system production over measurement period. *Utilization* is usually defined as total busy time of a resource over total time resource available. Queue theory allows researchers to predict or approximate these measures.

Software

There are many software packages available for conducting DES. However, most of these are

custom built for specific purposes. The Institute for Operations Research and the Management Sciences provides an extensive list of vendors on its Web site. Currently, some of the most commonly used general-purpose DES packages are GPSS, Arena/SIMAN, AutoMod, Extend, ProModel, Simu18, and Witness. There is also freeware available on the Internet, although these tools generally require a higher degree of computing skill to use.

James Stahl

See also Decision Trees, Construction; Markov Models

Further Readings

Banks, J. (1998). *Handbook of simulation.* New York: Wiley.

Davies, H., & Davies, R. (1987). A simulation model for planning services for renal patients in Europe. *Journal of the Operational Research Society, 38*(8), 693–700.

Institute for Operations Research and the Management Sciences: http://www.informs.org

Nance, R., & Sargent, R. (2002). Perspectives on the evolution of simulation. *Operations Research, 50*(1), 161–172.

Discrimination

In statistics, discrimination is the ability of a prediction (judgment scheme, statistical model, etc.) to distinguish between events and nonevents (or cases from controls, successes from failures, disease from nondisease, etc.). In the simplest form, a prediction scheme focuses on a single event with two possible states and assigns some estimate of the chance that one state will occur. This prediction comes from the set of cues and other factors, both measurable and immeasurable, available to the researcher.

Whether it is meteorologists forecasting the weather, business analysts predicting the rise and fall of the stock market, bookmakers predicting the big game, or physicians diagnosing disease, predictions have some degree of "correctness" relative to the actual occurrence of some unknown or future event. In medicine, we commonly predict the presence or absence of disease (diagnosis) or the likelihood of development of disease progression (prognosis). Measures have arisen to gauge the quality of a given set of predictions and to quantify prediction accuracy.

Multiple methods for forming these predictions exist, and each has associated strengths and weaknesses. One aspect is the "difficulty" that is set by nature. Outcome index variance is a measure of this difficulty. In addition, calibration addresses the relationship of the subgroup-specific predictions to the subgroup-specific observed event rate. The part of prediction accuracy that is often of highest interest is discrimination. The task of discrimination is to determine with some degree of certainty when the event will or will not occur. It measures the degree to which the prediction scheme separates events from nonevents. Discrimination is therefore influenced by variation in the predictions within the event/nonevent groups. Discrimination strength is related to the degree to which a prediction scheme assigns events and nonevents different probabilities—in other words, how well a scheme separates events into distinct "bins" (e.g., alive vs. dead or first vs. second vs. third). The sole focus of discrimination is this ability to place different events into different categories. The labels placed on those categories are somewhat arbitrary.

"Perfect" discrimination will occur when each appropriate category contains 100% or 0% of events. Perfect nondiscrimination, or nil discrimination, occurs when the group-specific event rate is the same as the overall percentage of events (also called the prevalence or mean base rate). In this case, the prediction scheme is no better than chance, and the groups are essentially assigned at random. One can, however, do worse than this by predicting groups in the wrong direction. However, this is, in a sense, still better than nil discrimination, but it is classifying groups incorrectly. Any discrimination that is better than the overall event prevalence improves the discrimination. In this case, simply reversing the labels of event and nonevent associated with the predictions can improve the discrimination.

Discrimination Types

Discrimination can be thought of in three distinct ways, each of use in different situations. These three types of discrimination arise from thinking

of discrimination like types of data. Data can be *nominal*, having no order but simply labels (e.g., color or gender). *Ordinal* data have associated order (e.g., mild/moderate/severe) but no measurable distance between groups. *Continuous* data have a distance between two groups that can be measured. Discrimination can be thought of along a similar continuum: nominal, ordinal, and continuous.

The simplest conceptualization is to partition items into similar event groups—that is, strictly labels without an associated order. Separation occurs through labeling similar predicted event groups with the same name or probability. These labels have no rank or relative position. The labels have no intrinsic meaning and serve no purpose other than to form bins to place the groups into. Discrimination is measured by the degree to which these bins are used for distinct event types. This level of discrimination can be measured when no probability measurement is assigned to the groups. Observations need only be assigned to differing groups with a similar likelihood of event occurrence (e.g., Group A vs. Group B, red vs. blue, or common vs. rare).

The Normalized Discrimination Index (NDI) is typically used to measure this type of discrimination and is most often found in the meteorological literature. A hypothetical example of looking at the NDI would be a study comparing the ability of two new screening exams to separate cancerous lesions from noncancer lesions. All other things being equal, it would be favorable to choose the new test that has the higher NDI.

Next, one can conceive a measure of rank order discrimination—for example, when only rank order predictions are available. In this case, the available information separates groups into situations where "A will have a higher event rate than B." With rank order discrimination, a group of events can be placed in terms of least to most likely. Rank order discrimination occurs when the events have predictions consistently higher (or lower) than the nonevents. Rank order discrimination measures the ability of a judge to correctly assign the higher likelihood of occurrence when the outcome of interest actually occurs. This is similar to nominal discrimination, but bins now have an associated rank, thus requiring at least ordinal predictions.

The area under the receiver operating characteristic (ROC) curve, or *C* statistic, best measures this sort of discrimination and is the most used method in medicinal research. The *C* statistic ranges from 0 (*perfect discrimination, wrong labels*) to 1.0 (*perfect discrimination*), with *nil discrimination* at .50. Returning to the hypothetical example, we would look to the *C* statistic if instead of the lesion being declared cancerous versus noncancerous, the screening tests returned a four-level scale (e.g., high, medium, low, or no risk of being cancerous).

Finally, actual probability estimates used can be compared among the groups. By comparing actual probabilities, the focus is on a continuous discrimination, drawing an arbitrary cut-point where separation is most distinct. Continuous discrimination determines how far apart groups are on a probability scale, and it requires continuous predictions to be calculated. The difference between the mean probabilities assigned to events and nonevents defines the slope index (SI), the primary measure of this type of discrimination. If we had two screening models that returned the exact probability of the lesion being cancerous, we could use the SI to compare the models' discrimination ability.

Discrimination Measurements

Except for the *C* statistic, these measures are on a –1 to +1 scale, where 1 is *perfect discrimination*, 0 is *nil discrimination*, and –1 indicates *perfect discrimination with the wrong labels*. The *C* statistic can be transformed into Somer's *D* by subtracting 0.5 and doubling the result. Somer's *D* is on the same –1 to +1 scale as NDI and SI. Unfortunately, no rule of thumb exists to define "weak" or "strong" discrimination. Since what might be "strong" discrimination in one area might be "weak" in another, discrimination strength is relative to the scientific area of interest; thus, this entry is reluctant to provide a rule of thumb for good versus poor discrimination. Whether a .7 score (70% of the total scale) is a strong discrimination really depends on the situation under study. In areas where little is known and any relationship is of value, a smaller amount of discrimination might be more important than in an area where much is understood and the research is trying to distinguish between degrees of perfection.

Any given prediction scheme will have degrees of the three types of discrimination. So long as predictions are given in terms of probabilities, for example, the results of a logistic regression, all three of these measures can be calculated. By creating ordered bins of probability of some fixed width, the SI, Somer's *D*, and NDI can all be calculated. This can be especially useful when discrimination ability of one type is capped and the goal is to determine tests that are "more perfect," or stronger but on a differing scale. For example, when taking a drug from the ideal conditions of a randomized clinical trial and using it in day-to-day practice, "ROC shrinkage" or a slightly less effective test is often observed. Examinations of continuous discrimination can help gauge the degree of ROC shrinkage, that is, the reduction in rank order discrimination expected to be observed, when variation in predictions increases.

Matthew Karafa

See also Calibration; Diagnostic Tests; Logistic Regression

Further Readings

Dawson, N. V. (2003). Physician judgments of uncertainty. In G. B. Chapman & F. A. Sonnenberg (Eds.), *Decision making in health care: Theory, psychology, and applications* (p. 211). Cambridge, UK: Cambridge University Press.

Yaniv, I., Yates, J. F., & Smith, J. E. K. (1991). Measures of discrimination skill in probabilistic judgment. *Psychological Bulletin, 110*(3), 611–617.

Yates, J. F. (1990). *Judgment and decision making* (p. 56). Englewood Cliffs, NJ: Prentice Hall.

Yates, J. F. (1994). Subjective probability accuracy analysis. In G. Wright & P. Ayton (Eds.), *Subjective probability* (pp. 381–410). New York: Wiley.

Zhou, X.-H., Obuchowski, N. A., & McClish, D. K. (2002). *Statistical methods in diagnostic medicine.* New York: Wiley.

Disease Management Simulation Modeling

Simulation is a general term describing a method that imitates or mimics a real system using a *model* of that system. Models vary widely and may be a physical object, such as a mannequin used in training healthcare providers, or a conceptual object, such as a supply-demand curve in medical economics. This entry is confined to computer models that are based on a logical or mathematical/statistical structure and use the computer to examine a model's behavior. Models can represent various types of healthcare systems that are engaged in disease management of patients, allowing, for example, examination and comparisons of alternative clinical decisions for patient care, insurance coverage policies, or the processes for delivering safe, effective, and efficient preventive or therapeutic care.

The best practices for disease management use evidence-based medicine, such as the outcomes from observational and experimental human studies, including clinical trials. However, such studies are not always possible and may be impractical. Consider, for example, studies that seek to determine the most effective (balancing risks and benefits) and cost-effective (balancing costs and effectiveness) strategies for colon cancer screening. An effective strategy might include ignoring small polyps in low-risk people, but a prospective human study that includes such a component might never be approved. Determining the best age to initially screen for colon cancer would require an experiment that tested perhaps 25 different ages. Determining the frequency and type of follow-up testing based on a person's family history, biological and social profile, and past test results, including the size and type of past polyps, would require such a large study over such a long period of time as to be essentially impossible. In engineering and in the physical sciences, computational models have been frequently used to complement and to substitute for direct experimentation.

Key Components of Simulation Models

Simulation models can be used to integrate evidence from observational and experimental human studies and extend insights into the consequences of different disease management strategies. The fundamental concept involves constructing a model of the natural history of the disease in an individual patient from a specific patient group. The model can be simulated on the computer to produce the

experience of many patients with this disease over their lifetimes. Then the model is altered to represent a medical strategy of care that includes an intervention, such as a screening test, a diagnostic test, medical therapy, or a surgical procedure. The population of patients with the intervention is simulated using the new model, and the results for the new model are compared with the results from the baseline model. Statistical comparisons can readily be made across myriad clinical strategies.

Validating the Simulation Model

A model is a representation of reality, not reality itself. As a representation, it attempts to replicate the input and the essential logical structure of the real system. A valid model can be exercised and the results inferred to the real world being studied. Consider Figure 1.

Here, the real world is the experience of real patients. The modeled world is the simulated experience of those patients. In addition to the proper representation of the logical and temporal relationships among the patients and their disease and the accurate description (including higher-order moments beyond the mean) of the probabilities of events and the importance of the outcomes of the various events, a third important key to any successful modeling activity is its validation. For purposes of assessing strategies for disease management, *construct validity* is supported by including model elements, relationships, and data derived from the published literature and assessed as appropriate by clinical experts. *Criterion-based validity*, comparing a model's output with real-world data when input conditions are held similar, provides significant assurance of the overall validity of the model for the purposes for which it is intended.

Notably, all models have limitations in their ability to represent the totality (complete detail) of actual patients or systems of care. Nonetheless, the simulation model should ideally represent patient- and/or system-level experiences that are indistinct from real patients or systems of care. For example, when computer-generated (simulated) histories are compared with a set of real patient histories (at the same level of detail), physicians (or other medical personnel) should not be able to tell them apart. Ultimately, the level of detail should depend on the questions being asked of the model, and sufficient detail should be included to allow model validation and provide for useful results.

Performance Measures From the Simulation

The amount of detail in a disease management simulation model is a direct reflection of the purpose for doing the simulation, such as comparing health outcomes or healthcare system performance. For example, the model can record the

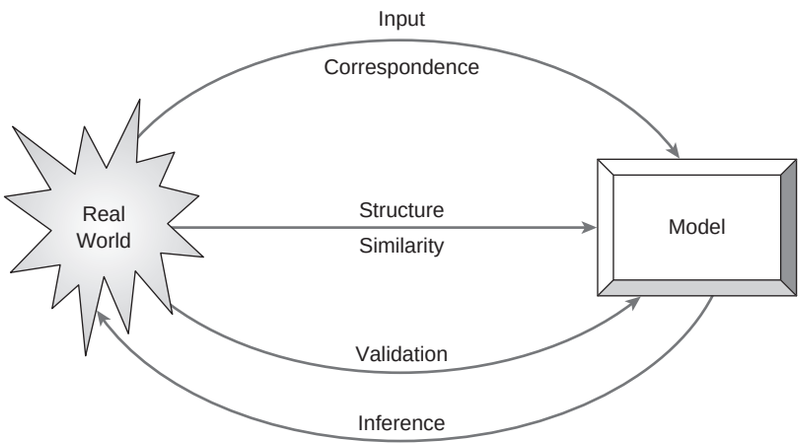


Figure 1 Relationship of model to real world

time a simulated patient spends in various health states (precancerous, healthy, cancerous, etc.), with each health state representing a different level of quality of life. A summary estimate from many simulated patients of the overall *quality-adjusted life years* (QALYs) can then be computed. Costs could be collected as a patient moves through the care system and summarized for many simulated patients to assist in the assessment of cost-effectiveness. Various healthcare resource utilization metrics may also be collected, such as the number of screening or diagnostic tests, laboratory procedures, or days of hospitalization. Performance measures need to be delineated during the construction of the simulation model so that appropriate data are being collected during execution of the model and summarized for analysis.

Data collections relative to performance measures are implemented in the simulation as patients traverse the care system. These performance measures are generally statistically presented at the end of the simulation. By sampling a statistically sufficient number of people, adequate statistical confidence intervals for the averages can be constructed. Fortunately, since individual patients are being individually simulated, the performance measures for the patients provide independent and identically distributed observations, which make traditional statistical analysis applicable. Advanced statistical methods related to statistical design of experiments can also be applied to simulation models.

Modeling the Care Cycle

A disease management simulation model should be a model of the care cycle for that disease, namely, from onset of the disease to eventual resolution or death. For example, a woman who died from colon cancer may have experienced the medical timeline shown in Figure 2.

In this case, two undetected adenomas occur (A1 and A2), each of which would eventually result in invasive cancer (C1 and C2). Surgery is performed to remove the invasive cancer from C1 at time CO. But because of late detection, she dies from this colon cancer at CD, which occurs before the second adenoma develops into invasive cancer (C2). Had A1 been detected and removed with a screening strategy, she would have survived to natural death (i.e., from some cause other than colon cancer). Each of these occurrences is considered an "event" since each changes the health status of the individual.

A valid model of this "natural history" must recognize that almost all the causes for these events that affect the timeline need be described by *random variables*. A random variable is a variable whose exact value is unknown but is described by a probability distribution or probabilistic process. The time for an undetected adenoma to become invasive cancer is an example of a "time until" random variable, whereas the incidence of undetected adenomas is a random process. Also, note that there is an "order" or pathway to the events in that an invasive cancer cannot occur without first being an undetected adenoma.

So a comprehensive model of the natural history for this case would include the following: (a) incidence of the disease based on an overall risk modified by the individual risk, (b) the disease pathways, (c) the rate and trajectory of

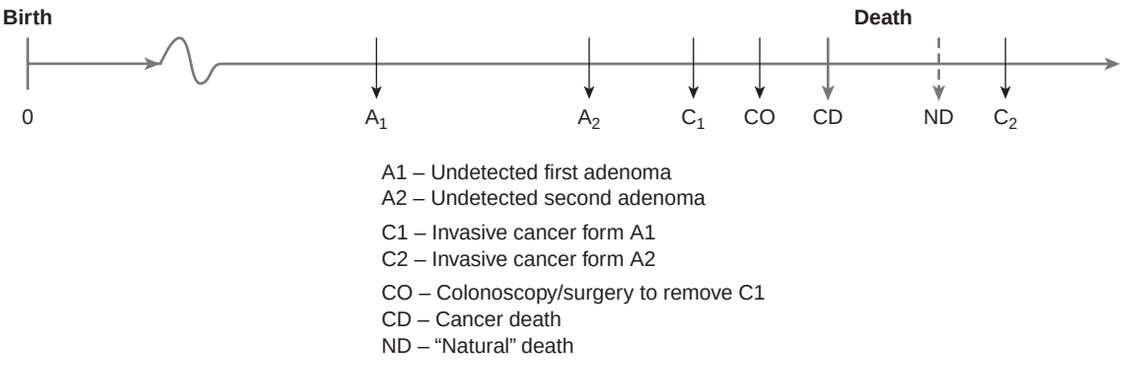


Figure 2 Medical timeline for a colon cancer death

progression, (d) the state and progress of the end disease, and (e) the time until natural death. Corresponding to the intervention (e.g., a screening strategy), the model changes would potentially alter any of the processes to produce longer life or more quality-adjusted life years or different costs in relation to quality-adjusted life years.

Modeling Details

There are several possible approaches to modeling these kinds of disease processes. One approach would be a state-based model, such as a Markov process. The modeling approach would identify patient states and the potential transitions between states. Figure 3 is an example of a Markov model.

Here, the "ovals" represent states relative to cancer, and the "arrows" represent the possible transitions between states at fixed points in time. Transitions are usually probabilities and may be a function of patient characteristics, such as age, gender, and genetic profile. The arrows that are directed back to the same state indicate that one possible transition is to remain in the same state. An alternative is to model "time in state" as a random variable. If the arrows represent transition probabilities, then the total probability "out" for each state must sum to 1.

Markov models are usually used to compute change in the state of a particular population. Time is incremented, and the transitions are applied, yielding populations in different states. For example, if 100 people start with no cancer, then after one time step, some people may move to the "Death" state, while others may move to the "Local cancer" state. Yet, depending on the probabilities, most will remain in the "No cancer" state. When a Markov model is used to model individual experiences, the simulation must be manipulated to employ tracker variables to report the events that a patient has suffered over time. In general, when the model employs random variables to describe state transitions and time-in-state variables, the simulation is called a Monte Carlo simulation. One of the popular uses of Monte Carlo simulations is in *probabilistic sensitivity analysis*, in which the parameters of a decision model are represented by random variables and the decision model is examined by sampling from these uncertain parameters.

A generalization of the Monte Carlo simulation model is the *discrete-event simulation*. Discrete-event simulations typically focus on the experiences of individuals throughout a process (such as healthcare delivery) and statistically aggregate the individual experiences to a population at risk. The description of a discrete-event simulation begins with the identification of *events*. Events are points in time when the individual changes health state—namely, when a patient experiences something that moves him or her from one state to another. The simulation operates by maintaining a calendar of future events, removing one event at a time, updating time to the time of that event, and executing all the processes associated with that event. Execution

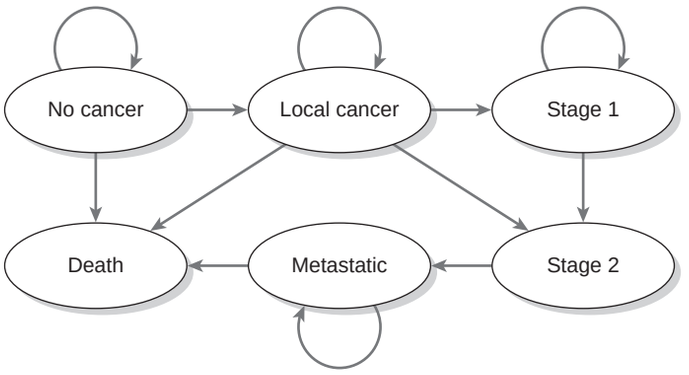


Figure 3 A Markov model

of these processes may add new events to the calendar of future events. What makes a discrete-event simulation most flexible is that it deals with only the immediate event. The immediate event may change other potential future events, so that, for example, more than one adenoma can be represented in the colon and each has its own characteristics. Diseases with multiple precursors and multiple consequences are readily included in the model. A discrete model can be visualized as an event diagram, as shown in Figure 4.

In this diagram, the "boxes" represent events, while the arrows represent the event-scheduling requirements. For example, suppose a "Nonvisible adenoma" event occurs. From this event, three possible new events are scheduled: (1) possibly another nonvisible adenoma event, (2) an advanced adenoma event if the progression type is progressive, and (3) a cancer event when the progression type is progressive or if the cancer is immediate. Note that an event graph is not a flowchart, since it only schedules future events. Furthermore, the time when a future event is scheduled may be described by a random variable. Thus, time to the next event is also part of the "arrow." It is the scheduling of future events that distinguishes the discrete-event simulation from its Monte Carlo counterpart.

Choosing to Do a Simulation

Simulation modeling provides a very flexible and powerful method to represent the evolution of disease and the management of its treatment. However, part of the power of the technique is derived from the detailed data input requirements, which is also a challenge when using the method. While a simulation of the natural history of a disease may employ local and national databases, it is often the case that critical information related to the stochastic nature of the disease and treatment process must be estimated or inferred from experience and nominal group or survey techniques.

Disease management simulation models provide a viable method to synthesize the complex natural history of a disease replete with the stochastic and statistical elements that describe real experiences. Interventions in the process make it possible to consider alternative management choices quantitatively.

Stephen D. Roberts and Robert S. Dittus

See also Discrete-Event Simulation; Markov Models

Further Readings

Caro, J. J. (2005). Pharmacoeconomic analyses using discrete event simulation. *PharmacoEconomics, 23*(4), 323–332.

Banks, J., Carson, J. S., Nelson, B. L., & Nicol, D. M. (2005). *Discrete-event simulation* (4th ed.). Upper Saddle River, NJ: Prentice Hall.

Doubilet, P., Begg, C. B., Weinstein, M. C., Braun, P., & McNeil, B. J. (1985). Probabilistic sensitivity analysis using Monte Carlo simulation: A practical approach. *Medical Decision Making, 5*(2), 157–177.

Freund, D. A., & Dittus, R. S. (1992). Principles of pharmacoeconomic analysis of drug therapy. *PharmacoEconomics, 1*(1), 20–31.

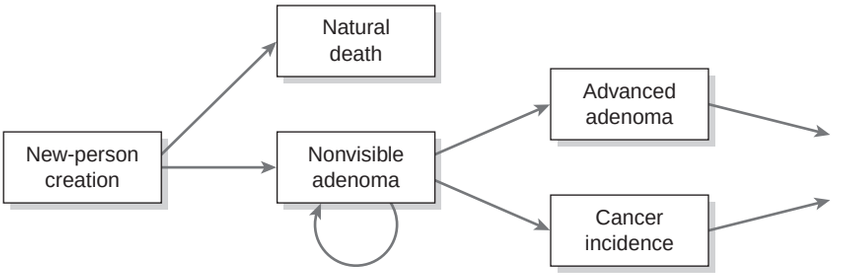


Figure 4 An event diagram for discrete-event simulation

Gold, M. R., Siegel, J. E., Russell, L. B., & Weinstein, M. C. (Eds.). (1996). *Cost-effectiveness in health and medicine*. New York: Oxford University Press.

Puterman, M. (1994). *Markov decision processes*. New York: Wiley.

Roberts, S., Wang, L., Klein, R., Ness, R., & Dittus, R. (2007). Development of a simulation model of colorectal cancer. *ACM Transactions on Modeling and Computer Simulation, 18*(4), 1–30.

Sonnenberg, F. A., & Beck, J. A. (1993). Markov models in medical decision making: A practical guide. *Medical Decision Making, 13*(4), 322–339.

Distributions: Overview

In medical decision making, distribution functions are used for two main purposes. The first is to model variability in data at the individual observation level (often subjects or patients). The second is to model uncertainty in the parameter estimates of decision models.

Distributions for Modeling Variability

Distributions that are used to model variability can be either discrete or continuous. Examples of discrete distributions include the binomial distribution, commonly used to model the occurrence or not of an event of interest from a total sample size, and the Poisson distribution, commonly used to model counts of events. Examples of continuous distributions that are used to model data variability are the normal distribution and gamma distribution. The modeling of variability is particularly important for discrete-event simulation (DES) models, which are often employed to look at service delivery methods that involve queuing problems. For example, patients might be assumed to arrive at an emergency room and queue up to see the receptionist before waiting to see a physician. Arrival times could be modeled as random while following an underlying exponential distribution, and different methods of organizing the procedures for receiving and attending to patients could be modeled to maximize throughput and minimize waiting time. More generally, individual patient simulation models describe medical decision models that model an individual's pathway through disease and treatment. Monte Carlo simulation is typically used to represent the stochastic nature of this process and is termed "first-order" simulation when the focus is on variability in the patient experience rather than uncertainty in the parameters.

Distributions for Modeling Parameter Uncertainty

The use of probability distributions to represent parameter uncertainty in decision models is known as probabilistic sensitivity analysis. Distributions are chosen on the basis of the type of parameter and the method of estimation. Monte Carlo simulation is then used to select parameter values at random from each distribution, and the model is evaluated at this set of parameter values. By repeating this process a large number of times, the consequences of uncertainty over the input parameters of the model on the estimated output parameters is established. In contrast to modeling variability, only continuous distributions are used to model parameter uncertainty. Monte Carlo simulation used in this way is termed "second order" to reflect the modeling of uncertainty of parameters. Probability parameters are commonly modeled using a beta distribution, since a beta distribution is constrained on the interval 0 to 1. Parameters such as cost of quality-of-life disutility, which are constrained to be 0 or positive, are often modeled using the lognormal or gamma distributions since these distributions are positively skewed and can only take positive values. Relative-risk parameters are often used as treatment effects in decision models and can be modeled using a lognormal distribution, reflecting the standard approach to the statistical estimation of uncertainty and confidence limits for these measures.

Central Limit Theorem

The normal distribution is of particular note for two reasons. First, it turns out that many naturally occurring phenomena (such as height) naturally follow a normal distribution, and therefore, normal distributions have an important role in modeling data variability. Second, the *central limit theorem* is an important statistical theorem that states that

whatever the underlying distribution of the data, the sampling distribution of the arithmetic mean will be normally distributed with sufficient sample size. Therefore, the normal distribution is always a candidate distribution for modeling parameter uncertainty, even if the parameters are constrained (in technical terms, if there is sufficient sample size to estimate a parameter, the uncertainty represented as a normal distribution will result in negligible probability that a parameter will take a value outside its logical range).

Statistical Models

Decision models in the medical arena often include statistical models as part of their structure. For example, a multivariate logistic regression may be used to estimate the probability of an event, or an ordinary least squares regression model may be used to explain how quality-of-life disutility is related to a particular clinical measure. Statistical regression models are of interest in that they simultaneously assume a distribution for the data and for the parameters of interest. For example, suppose that a transition probability in a Markov model is to be estimated from a survival analysis of time to event. A common parametric distribution for the time-to-event data themselves might be a Weibull distribution, which is capable of modeling time dependency of the underlying hazard function of the event of interest. However, the parameter uncertainty relates to the estimated coefficients from the regression of how the (log) hazard depends on patient characteristics. Since the scale of estimation in survival analysis is the log hazard scale and since a multivariate normal distribution of regression coefficients is assumed, this means that the underlying distribution of any particular parameter (coefficient) of the model is lognormal.

Bayesian Interpretation

As a brief aside, it is perhaps worth noting that the use of parametric distributions to represent uncertainty in decision models underlies the fundamentally Bayesian nature of medical decision making. The classical approach to probability does not allow uncertainty in the parameters themselves; rather, uncertainty relates to the estimation process and the likelihood that the true (but unobserved) parameter takes a particular value given the data. The Bayesian paradigm more naturally allows distributions to be chosen to reflect not only the data (equivalent to the frequentist data likelihood) but also the parameter itself.

Effective Modeling

Many distributions exist and have potential applications in medical decision making. Nevertheless, the appropriate distribution typically depends on the purpose of the model, on the constraints on the data or parameter, and on the method of estimation. Careful consideration of the appropriate distribution is required for effective modeling. Typically, such careful consideration will reduce a wide set of all possible distributions to a small set of candidate distributions for a particular application within the model.

Andrew H. Briggs

See also Bayesian Analysis; Decision Trees: Evaluation With Monte Carlo; Decision Trees: Sensitivity Analysis, Basic and Probabilistic; Decision Trees: Sensitivity Analysis, Deterministic; Discrete-Event Simulation; Managing Variability and Uncertainty; Parametric Survival Analysis

Further Readings

Briggs, A., Claxton, K., & Sculpher, M. (2006). *Decision modelling for health economic evaluation*. Oxford, UK: Oxford University Press.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (1995). *Bayesian data analysis*. London: Chapman & Hall.

Johnson, N. L., Kemp, A. W., & Kotz, S. (2005). *Univariate discrete distributions* (3rd ed.; Wiley Series in Probability and Statistics). Hoboken, NJ: Wiley.

Johnson, N. L., Kotz, S., & Balakrishnan, N. (1995). *Continuous univariate distributions* (Vol. 2; Wiley Series in Probability and Statistics). New York: Wiley.

Kotz, S., Balakrishnan, N., & Johnson, N. L. (1994). *Continuous univariate distributions* (Vol. 1; Wiley Series in Probability and Statistics). New York: Wiley.

Parmigiani, G. (2002). *Modeling in medical decision making*. Chichester, UK: Wiley.

Distributive Justice

Distributive justice is the branch of theories of justice that is concerned with the distribution of available resources among claimants. Theories of distributive justice all accept the central claim of formal justice that "equal claims should be handled equally" but differ in terms of what features they accept as relevant to the judgment of equality in claims and in outcomes. Considerations of distributive justice in health and healthcare are important because being healthy is a prerequisite for full or equal participation in a whole range of social activities, from employment to politics.

Within healthcare, considerations of distributive justice are especially important in the context of priority setting or rationing of healthcare services. This raises specific problems because one of the main outcomes of interest, namely, health, cannot be distributed directly. Health is not fungible and not detachable from the person who is healthy. Many discussions about justice in health therefore focus on access to healthcare, on the social determinants of health, or on justice in aggregate outcomes for groups of claimants. A further issue of current debate is whether justice in health or healthcare can be detached from much more general questions of social justice.

The main approaches to distributive justice that are of relevance to medical decision making are egalitarianism, maximization, equal opportunity, and procedural.

Egalitarianism

The most stringent theory of distributive justice is egalitarianism, which simply states that the resource in question should be distributed equally among claimants. Egalitarianism also implies that if complete equality cannot be obtained, then the distribution must at least reduce any preexisting inequalities.

In the healthcare context, most egalitarians hold that what determines the strength of a healthcare claim is exclusively the size of the healthcare need. The further from complete health a person is, the stronger is his or her claim to healthcare. However, the concept of health need probably also contains elements related to urgency and to the possibility of intervention.

Most people hold egalitarian views of varying strengths, and such views probably underlie the common idea that healthcare resources should be allocated primarily to those who have the greatest need, that is, those who are most ill.

Egalitarianism is open to two significant counterarguments: (1) the leveling-down objection and (2) a potential conflict between equality and making people better off. The leveling-down objection is simply that one of the ways of making people (more) equal is to take something away from those who have the most without redistributing it. In the medical context, health equality could be increased by taking health away from those who are completely healthy. A strict egalitarian would have to claim that that is an improvement, but that judgment seems highly counterintuitive. A world where some people have been harmed and no one benefited cannot be ethically better than the one before the change.

The potential conflict between equality and making people better off arises in its starkest form when one is contemplating a change that will improve the situation of everyone while also widening inequalities. A strict egalitarian would have to say that the position after the intervention is worse than before, and this is again strongly counterintuitive, at least in cases where the improvement for the least well-off is significant.

Maximization

Another prevalent theory of justice is the theory held by consequentialists. For consequentialists, what matters is the maximization of good outcomes from a distributive decision, not whether the distribution increases or decreases equality. The strength of a claim is thus not based on need but on how much good can be generated if the claim is met. In the healthcare context, this view underlies or is at least compatible with many health economic approaches to resource allocation—for instance, the quality-adjusted life year (QALY) approach. This compatibility with health economics is partly a result of an isomorphism between consequentialism and economic theory in their approach to maximization.

Most people hold maximizing views of some strength, and this underlies the general belief that a healthcare system has a strong, but perhaps not

overriding, obligation to allocate resources in order to get the largest benefit possible.

The most significant counterarguments to understanding distributive justice as the maximization of good outcomes are (a) that it is completely need independent, unless need is redefined as "possibility of benefiting," and (b) that maximization may in some cases be achieved by distributive decisions that take away resources from those who are a priori worst off and thus increase inequalities.

Prioritarianism

Prioritarianism is a recent revision of the consequentialist maximizing approach, aimed at dealing with some of the counterintuitive distributive effects of maximization. Prioritarians argue that when good consequences are assessed and added up, benefits to the worse off should count for more than comparable benefits to the better off. This will have the effect of strengthening the claims of those who are worse off and make it less likely that resources are taken away from them to benefit others. In the healthcare context, a prioritarian approach will thus lead to more resources being directed to those who are most ill than nonprioritarian maximization.

Equal Opportunity

A third approach to distributive justice argues that what is important is not equality in or maximization of outcomes but equality in initial opportunities or capabilities. If we distribute so that everyone has an equal starting point in relation to whatever resources are important for success in a given area, then we have distributed justly. This implies that the claims of those who are worst off in terms of opportunities or capabilities should be given priority. In the healthcare context, this would, for instance, imply that public health interventions aimed at socioeconomically deprived groups should be given priority; and this view also has significant implications for the distribution of resources for healthcare research.

The main counterargument against the equal opportunity approach is that it is often difficult to define who is worst off in a given situation because a person may be worst off on one parameter and not worst off on another. For instance, should priority be given to the claim of the poor person with a minor illness or the rich person with a major illness?

A further problem arises in the healthcare area because there are good reasons to believe that this is an area where it is impossible to create equality of opportunity. There are people with such severe disabilities or illnesses that there is no intervention that can improve their health to such a degree that they have equal opportunities with respect to health status.

The equal opportunities approach is very similar to an approach toward compensation for bad outcomes, including bad health outcomes that depend on a distinction between "brute luck" and "option luck." Brute luck refers to those outcomes that are not dependent on the choices of the person; it is, for instance, a matter of brute luck whether a person is born with a disability. Option luck refers to those outcomes that are dependent on a person's prior choices; it is, for instance, a matter of option luck if a smoker develops chronic obstructive lung disease. On the basis of this distinction, an argument can be made that persons should be compensated for large differences in brute luck, but that differences in option luck do not justify compensation or redistribution.

Procedural Approaches

It has long been recognized that there are situations where we know what the just outcome is but do not have any easy way of achieving it except by devising a procedure leading as close to the just outcome as possible. If X children are to share a birthday cake they should each get $1/X$ of the cake, but given the practical difficulties in cutting cakes, this is difficult to achieve. The procedural solution is to let one child cut the cake, knowing that he or she will be the last to pick a piece.

In healthcare it has recently been argued that the situation is even more complex. Not only do we not know how to bring about the just outcome (e.g., equal access to tertiary services if we believed that that was what justice required), there are many cases where we cannot fix the just outcome with any degree of precision. We may be able to identify clearly unjust distributions of resources but find it difficult to identify which of the remaining distributions is the most just. This argument has led many to shift focus from further elaboration of the

details of the theories of distributive justice to procedural approaches that can ensure that our distributive decisions are (a) not clearly unjust and (b) legitimate to those who are affected. These developments have been linked with the more general developments of ideas concerning deliberative democracy.

A number of different procedural approaches have been developed, but all aim at ensuring that (a) all stakeholders have a voice, (b) all reasonable arguments are put on the table, and (c) the decision processes are transparent.

The currently most popular and well-researched procedural approach is the so-called accountability for reasonableness (or A4R) approach. It has four distinct components: publicity, relevance, appeals, and enforcement. In conjunction, these four components emphasize reason giving and create a process with successive opportunities for all interested parties to challenge priority decisions.

Publicity is a call for explicitness. *Relevance* entails a requirement for reasonableness in priority setting. That is, priority decisions must be made in accordance with reasons that stakeholders will agree are relevant and adequate. The *appeals* component is an institutional mechanism that provides patients with an opportunity to dispute and challenge decisions that have gone against them. Finally, *enforcement* entails public or voluntary regulation of the decision process to ensure that the three other components are maintained. Proper enforcement of the decisions that are made through agreement on fairness will ensure that reasoning is decisive in priority setting and not merely a theoretical exercise.

Søren Holm

See also Bioethics; Rationing

Further Readings

Daniels, N. (1985). *Just health care*. Cambridge, UK: Cambridge University Press.

Daniels, N. (2000). Accountability for reasonableness: Establishing fair process for priority setting is easier than agreeing on principles. *British Medical Journal, 321,* 1300–1301.

Nozick, R. (1977). *Anarchy, state and Utopia*. New York: Basic Books.

Rabinowicz, W. (2001). *Prioritarianism and uncertainty: On the interpersonal addition theorem and the priority view*. Retrieved August 14, 2008, from http://mora.rente.nhh.no/projects/EqualityExchange/ressurser/articles/rabinowicz2.pdf

Rawls, J. (1999). *A theory of justice* (Rev. ed.). Harvard, MA: Belknap Press.

Sen, A. (1992). *Inequality reexamined*. Oxford, UK: Clarendon Press.

Temkin, L. S. (1993). *Inequality*. Oxford, UK: Oxford University Press.

DISUTILITY

Where utility reflects the positive value of a health state to an individual (its desirability) and is expressed as the fraction of perfect health it entails, disutility reflects the complement of this fraction (its undesirability), 1 minus the utility. Thus, if a disability state is assigned a utility of .85, its disutility, relative to good health, is .15. Disutility is mostly used in comparative contexts, where states are compared relative to one another. In these cases, disutility is the difference in the average utility reported by persons with a given problem compared with those without the problem. An example is that of a treatment for menopausal symptoms that is 80% effective. If the utility for the health state "living with menopausal symptoms" is .6, a way to calculate utility with treatment is the following: The disutility of the remaining symptoms will be (1 – Effectiveness of treatment) × Disutility from symptoms = .20 × .40 = .08, and thus, the utility for "living with the remaining menopausal symptoms" will be 1 – .08 = .92.

Expected Utility Theory

The utility of a health state is a cardinal measure of the strength of an individual's preference for particular outcomes when faced with uncertainty, on a scale from 0 to 1, where 0 generally reflects *death* and 1 reflects *perfect health*. A distinction is usually made in the decision-making literature between utilities, or strengths of preferences under uncertainty, and values, strengths of preferences under certainty. This concept of utilities dates back to 1944, when John von Neumann and Oskar

Morgenstern developed a normative model for decision making under uncertainty—expected utility theory. This model calculates the utility that can be expected from each option in terms of the desirability of its outcomes and the probability with which they will occur. For most decisions in healthcare, outcomes may occur with a certain probability, and the decision problem is thus a problem of choice under uncertainty. Decision analysis is indeed firmly grounded in expected utility theory, and the most common use of utilities is in decision analyses. In decision analyses, the strategy of preference is calculated by combining the utilities of the outcomes with the probabilities that the outcomes will occur.

Anne M. Stiggelbout

See also Utility Assessment Techniques; Values

Further Readings

Franks, P., Hanmer, J., & Fryback, D. G. (2006). Relative disutilities of 47 risk factors and conditions assessed with seven preference-based health status measures in a national U.S. sample: Toward consistency in cost-effectiveness analyses. *Medical Care, 44,* 478–485.

Hunink, M., Glasziou, P., Siegel, J., Weeks, J., Pliskin, J., Elstein, A., et al. (2001). *Decision making in health and medicine. Integrating evidence and values*. Cambridge, UK: Cambridge University Press.

Morgenstern, O., & von Neumann, J. (2004). *Theory of games and economic behavior*. Princeton, NJ: Princeton University Press. (Original work published 1944)

DOMINANCE

In cost-effectiveness analyses, costs and effectiveness of different decision alternatives are estimated. They can then be presented in the two-dimensional cost-effectiveness plane (Figure 1).

A decision alternative A is called strongly dominated (or dominated) by a different alternative B if the costs and effectiveness of Alternative B are at least as favorable as those of Alternative A:

$$\text{Effect}_A \leq \text{Effect}_B \text{ and } \text{Cost}_A \geq \text{Cost}_B,$$

with strict inequality for either effectiveness or costs. In Figure 1, all alternatives in the light gray top-left area are strongly dominated by Alternative B.

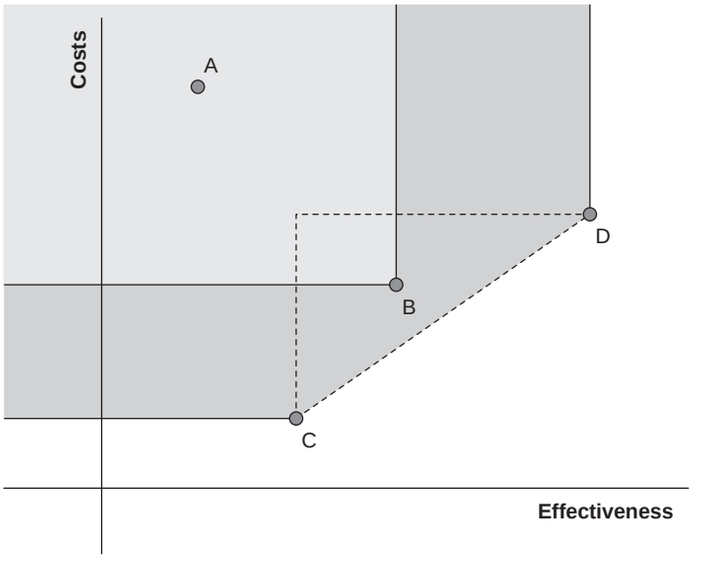


Figure 1 Cost-effectiveness plane

Note: The light gray area is (strongly) dominated by Alternative B. The dark gray area is weakly dominated by Alternatives C and D together.

Decision alternative A is called weakly dominated by two different alternatives C and D if Alternative A is strongly dominated by a mixture of those alternatives C and D:

$$\text{Effect}_A \leq \alpha \times \text{Effect}_C + (1 - \alpha) \times \text{Effect}_D$$

and

$$\text{Cost}_A \geq \alpha \times \text{Cost}_C + (1 - \alpha) \times \text{Cost}_D, \quad \text{for some } 0 \leq \alpha \leq 1,$$

with strict inequality for either effectiveness or costs. Mixtures can be thought of as if one alternative is applied to a fraction α of the patients and the other to a fraction $(1 - \alpha)$ of the patients. All such mixtures together form a straight line segment between alternatives C and D. In Figure 1, all alternatives in the dark gray area are weakly dominated by Alternatives C and D.

The main difference between strong dominance and weak dominance is illustrated by the dashed triangle in Figure 1. Decision alternatives in this triangle (like Alternative B) are not strongly dominated by Alternative C or by Alternative D, but they are weakly dominated by Alternatives C and D together. Strong and weak dominance are also referred to as strict and extended dominance.

Preference

Dominance is closely related to preference. Which alternative is preferred, in general, depends on the properties of the utility function on effectiveness and costs. In Figure 1, Alternative A has lower effectiveness and higher costs than Alternative B. Nevertheless, an individual is free to prefer A over B. However, it is more reasonable to assume that higher effectiveness and lower costs are preferred. The utility function $U(\cdot)$ is then strictly increasing in effectiveness and strictly decreasing in costs:

$$\text{Effect}_A < \text{Effect}_B \Rightarrow U(\text{Effect}_A) < U(\text{Effect}_B),$$

$$\text{Cost}_A > \text{Cost}_B \Rightarrow U(\text{Cost}_A) < U(\text{Cost}_B).$$

If Alternative A is strongly dominated by Alternative B, then this strict monotonicity of the utility function is sufficient for B to be preferred over A:

$$U(\text{Effect}_A, \text{Cost}_A) \leq U(\text{Effect}_B, \text{Cost}_A) \leq U(\text{Effect}_B, \text{Cost}_B),$$

with strict inequality for either of the inequalities.

For weak dominance, strict monotonicity of the utility function is not sufficient to determine preference. For example, in Figure 1, Alternative B may be preferred over Alternatives C and D. A common stronger assumption is that the utility function is linear in costs and effectiveness—that is, the utility function equals the net benefit:

$$U(\text{Effect}, \text{Cost}) = \text{WTP} \times \text{Effect} - \text{Cost},$$

where the positive WTP stands for the willingness to pay in monetary terms for one unit of effectiveness. For this linear utility function, a weakly dominated alternative cannot be preferred; if A is weakly dominated by Alternatives C and D, then A is less preferred than the hypothetical mixture (due to strong dominance) and the hypothetical mixture is not more preferred than the better of alternatives C and D (due to the linearity of the line segment and the utility function). This reasoning holds regardless of whether the decision alternatives C and D are actually divisible into mixtures. Therefore, weakly dominated alternatives are not the most preferred in standard cost-effectiveness analyses (i.e., with linear utility functions and positive WTP).

Example

Consider the numerical example presented in Table 1 and Figure 2. Alternative A is strongly dominated by B, C, and D because A has lower effectiveness and higher costs. Therefore, Alternative A is not preferred if the utility function is strictly increasing in effectiveness and decreasing in costs.

Alternative B is not strongly dominated by any of the alternatives. If the utility function were nonmonotone, then Alternative B could be the most preferred alternative. However, Alternative B is weakly dominated by Alternatives C and D, since it is above the line segment through C and D. The cost-effectiveness ratio of Alternative D compared with Alternative C is (\$20,000 – \$5,000)/(.8 – .3) = \$30,000 per QALY. For WTP below \$30,000 per QALY, B may be preferred over D but B is not preferred over C. For WTP above

Table 1 Example of cost-effectiveness analysis results, presented numerically

Decision alternative	A	B	C	D	E
Effectiveness (in QALYs)	.2	.4	.3	.8	.9
Costs	$25,000	$15,000	$5,000	$20,000	$35,000

Note: QALYs, quality-adjusted life years.

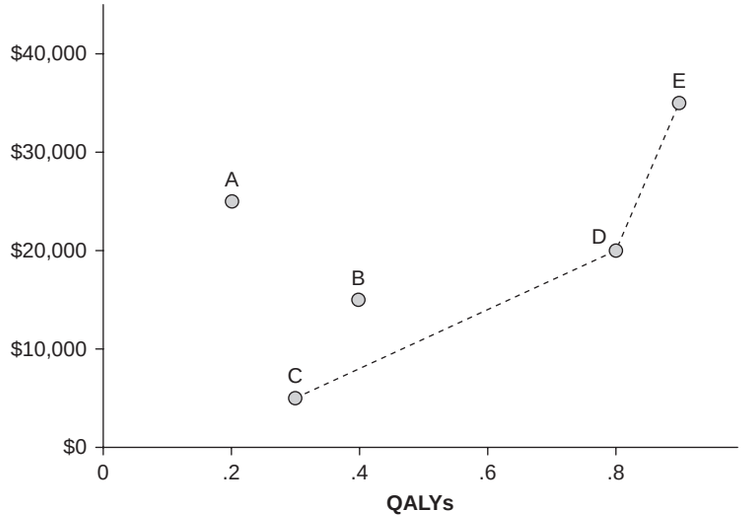


Figure 2 Example of cost-effectiveness analysis results, presented graphically

$30,000 per QALY, B may be preferred over C but B is not preferred over D. Regardless of the WTP, B is not preferred over both C and D. Therefore, Alternative B is not the most preferred alternative if the utility function is linear (with positive WTP per QALY).

Alternatives C, D, and E are neither strongly nor weakly dominated by any of the alternatives. Even if the utility function is linear, depending on the WTP, C, D, or E can be preferred. The cost-effectiveness ratios for Alternative C compared with D and for Alternative D compared with E are $30,000 and $150,000 per QALY, respectively. Alternative C is preferred for low WTP (up to $30,000 per QALY), Alternative B is preferred for intermediate WTP (between $30,000 and $150,000 per QALY), and Alternative C is preferred for high WTP (above $150,000 per QALY).

Wilbert van den Hout

See also Cost-Effectiveness Analysis; Economics, Health Economics; Net Monetary Benefit

Further Readings

Cantor, S. B. (1994). Cost-effectiveness analysis, extended dominance, and ethics: A quantitative assessment. *Medical Decision Making, 14,* 259–265.

Drummond, M. F., O'Brien, B. J., Stoddart, G. L., & Torrance, G. W. (1997). *Methods for the economic evaluation of health care programmes* (2nd ed.). New York: Oxford University Press.

Gold, M. R., Siegel, J. E., Russell, L. B., & Weinstein, M. C. (1996). *Cost-effectiveness in health and medicine*. New York: Oxford University Press.

Postma, M. J., de Vries, R., Welte, R., & Edmunds, W. J. (2008). Health economic methodology illustrated with recent work on chlamydia screening: The concept of extended dominance. *Sexually Transmitted Infections, 84,* 152–154.

Dual-Process Theory

Dual-process theories of cognition (also referred to as "two-system" theories) posit two distinct systems of judgment operating in parallel. Dual-process theories have been described since 1975, with a variety of different names for the two processes. Since 2000, however, the two processes have been conventionally referred to as System 1 and System 2.

System 1 is an intuitive judgement system that shares many features with the perceptual system. It operates by tacitly encoding and retrieving associations between perceived cues in the environment. System 1 is fast, holistic, and automatic and underlies pattern recognition, prototypicality judgments, and heuristic processing. Because it is driven by associations acquired through experience, it is sensitive to the features of learning context and environmental exposure. It is also influenced by the emotional state of the judge and the emotional content of the judgment.

In contrast, System 2 is a rule-based system for forming judgments. It is slow, effortful, and analytic and applies rules in an emotionally neutral manner. When appropriate data are available, System 2 yields the most normatively rational reasoning, but because it is relatively difficult and demanding, it is easily disrupted by high cognitive load or time pressure. The figure, reproduced from the psychologist Daniel Kahneman's Nobel Prize lecture on dual-process theories, compares the attributes of the two judgmental systems and the perceptual system.

A key feature of dual-process theories is that System 1 and System 2 operate simultaneously and in parallel. Because System 1 is considerably faster, System 1 judgments typically emerge first and serve as additional inputs to System 2. If a System 1 judgment does not emerge, the judge must resort to System 2 alone; similarly, if a lack of time or cognitive resources curtails System 2, the judge must resort to System 1 alone.

The two systems can interact in several ways. When a System 1 judgment has been made, System 2 may endorse the judgment, may use the System 1 judgment as an anchor and adjust the judgment on the basis of other situational features, or may identify the System 1 judgment as incompatible with a subjectively valid rule and block it from overt expression. Because System 1 processing itself (as distinct from the judge's response) cannot be suppressed, judges often feel drawn to the System 1 judgment even when they recognize that it is incorrect.

Kahneman has illustrated this effect with the bat-and-ball problem: "A baseball bat and a ball together cost one dollar and 10 cents. The bat costs one dollar more than the ball. How much does the ball cost?" Most people who hear this problem initially conclude that the ball costs 10 cents, but

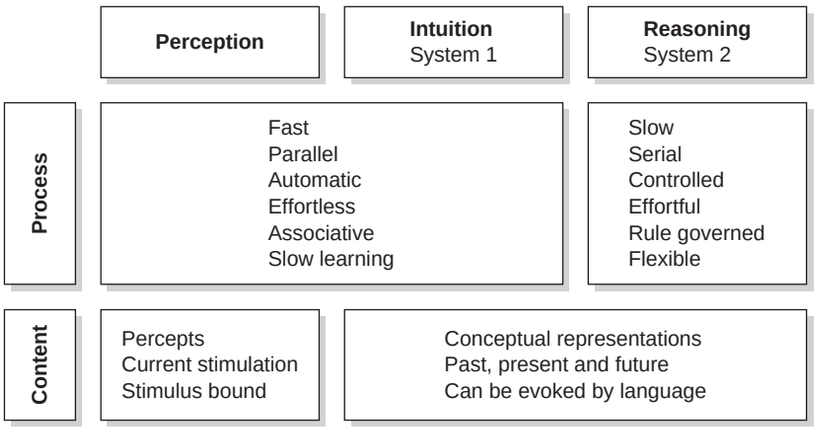


Figure 1 A comparison of the features of the human perceptual system with human judgment Systems 1 and 2

Source: Kahneman, D. (2003). Maps of bounded rationality: A perspective on intuitive judgment and choice. In T. Frangsmyr (Ed.), *Les Prix Nobel 2002* [The Nobel Prizes 2002]. Stockholm: Almqvist & Wiksell International. © The Nobel Foundation 2002. Reprinted with permission.

they realize, after a moment of reflection, that this (System 1) answer is incorrect and, in many cases, suppress the response in favor of the System 2 answer (5 cents), which emerges later.

Paul Slovic provides a more distressing example of the power of System 1 and the need for System 2 regulation. He reviews extensive research on willingness to provide life-saving interventions and argues that because the perceptual basis of System 1 is attuned to small changes at the margins, it can lead to increasing disregard for absolute numbers of lives saved. Saving a small number of lives is highly valued; saving 10,000 times as many lives (as in the case of preventing genocide), while more valuable, is intuitively treated as much less valuable than a factor of 10,000.

Dual-process theories posit System 1 as the source of heuristics in judgment; when the results of such heuristics produce normatively incorrect judgments, they are referred to as biases. However, Gerd Gigerenzer and colleagues have argued extensively for the adaptive nature of "fast and frugal" System 1 heuristics. Fuzzy-trace theory specifically argues that gist processing (a System 1 function) represents the apex of the development of reasoning.

Alan Schwartz

See also Bias; Fuzzy-Trace Theory; Heuristics; Intuition Versus Analysis; Judgment

Further Readings

Gigerenzer, G., Todd, P. M., & ABC Research Group. (2000). *Simple heuristics that make us smart*. Oxford, UK: Oxford University Press.

Hogarth, R. M. (2005). Deciding analytically or trusting your intuition? The advantages and disadvantages of analytic and intuitive thought. In T. Betsch & S. Haberstroh (Eds.), *The routines of decision making* (pp. 67–82). Mahwah, NJ: Lawrence Erlbaum.

Kahneman, D. (2003). Maps of bounded rationality: A perspective on intuitive judgment and choice. In T. Frangsmyr (Ed.), *Les Prix Nobel 2002* [The Nobel Prizes 2002] (pp. 449–489). Stockholm: Almqvist & Wiksell International.]

Slovic, P. (2007). "If I look at the mass I will never act": Psychic numbing and genocide. *Judgment and Decision Making, 2*(2), 79–95.

Stanovich, K. E., & West, R. F. (2000). Individual differences in reasoning: Implications for the rationality debate? *The Behavioral and Brain Sciences, 23*(5), 645–726.

Dynamic Decision Making

Most real-life decisions occur in multiple stages—individuals experience a series of actions and their consequences over time. Medical decision making, in particular, often involves anticipating uncertain future consequences where subsequent decisions are contingent on (or constrained by) the outcomes of patients' earlier choices. There has been a great deal of research on the basic principles underlying single-stage decision making but very little work on multistage decisions. Decision field theory has begun to examine many of the principles underlying multistage decision making and could be used to inform real-life choices involving both uncertainty and multiple stages.

Decision Field Theory

Most psychological research focuses on static decisions in isolation—a single decision followed by a single outcome. Dynamic decisions involve a sequence of decisions in which the choices and outcomes available at a later stage depend on the choices and outcomes that occurred earlier. Decision field theory tries to quantify such multistage decisions. A typical study asks subjects to take an initial gamble, and then they are given the option of taking a second gamble; this process is repeated.

Subjects are instructed to make three kinds of decisions:

1. A *planned decision:* Before subjects begin, they are asked to predict what they will decide at the end of the decision-making task, contingent on both winning and losing the initial gambles.
2. A *final decision:* This is what the subjects actually decide once they have gone through to the end of the task.
3. An *isolated decision:* All the initial gambles are eliminated so that only the final gamble remains.

Subjects are then asked to make the same final decision as above but without the experience of going through the previous decision-making tasks.

The normative procedure for selecting a strategy for these three decisions involves three consistency principles: dynamic, consequential, and strategic. The first requires the decision maker to follow through with his or her plans to the end, the second requires the decision maker to focus solely on future events and final outcomes given the current information available, and the third is the conjunction of the first two.

Decision field theory predicts (and has found that) there will be

1. a difference between planned and final decisions—a violation of dynamic consistency, that is, the plan for action differs from the final choice taken;
2. *no* difference between isolated and final decisions—*no* violation of consequential consistency; and
3. a difference between planned and isolated decisions—a violation of strategic consistency.

Dynamic Inconsistency

Two types of dynamic inconsistencies can be found: (1) Subjects who planned to take the second gamble but then won the second gamble (i.e., experienced a gain) become risk-averse and reverse their original plan; that is, they now want to play it safe and keep their winnings, so they choose not to take the gamble. (2) Subjects who planned *not* to take the second gamble but then lost the second gamble (i.e., experienced a loss) become risk seeking and reverse their original plan; that is, they now want to recoup their losses, so they are willing to take the risk and gamble.

The explanation for this reversal of preference was a change in the reference point. The planned decision was made against a reference point of zero (i.e., nothing gained or lost yet), but the final decision was made by incorporating the outcome of the first gamble. Consequently, the reference point was shifted such that the gamble seemed more or less risky, as shown above.

An alternative explanation is that the planned decision was made in a "cold" or rational state, whereas the final decision was made during the actual decision-making task, when subjects were in a "hot" or emotional state. Therefore, the final decision may be based more on immediate hedonic and affective processes, leading subjects to make a different choice in the "heat of the moment" from what they had planned to do.

Consequential Consistency

The consequential consistency finding is supported by the goal-gradient hypothesis. This hypothesis comes from approach-avoidance theories, which state that a decision anticipated from a distance feels very different from the decision one experiences as one gets closer to actually having to make a choice. Therefore, the hypothesis argues that the decision maker faces identical consequences from the same distance in both the final- and the isolated-decision conditions and, therefore, the two choices should also be identical.

Multistage Medical Decision

An important multistage medical decision individuals commonly face today involves cancer screening tests. Prostate and breast cancer are the most commonly occurring cancers in U.S. men and women and the second leading cause of cancer deaths. However, both tests are surrounded by controversy.

Prostate-specific antigen (PSA) testing has led to both overdiagnosis of and unnecessary treatment for prostate cancer. It is estimated that 75% of early-stage prostate cancers detected through PSA testing would never have become clinically significant. Therefore, men may be exposed to unnecessary prostate cancer treatment and suffer from the side effects of impotence and incontinence needlessly. Even professional organizations disagree about whether PSA screening is more beneficial than harmful.

Increased mammography screening has quadrupled the diagnosis of ductal carcinoma in situ (DCIS). Neither the prognosis nor the treatment for DCIS is known, and it is not necessarily a precursor to invasive breast cancer. However, this diagnosis has led some young women to undergo

prophylactic mastectomies that are potentially unnecessary in order to avoid developing cancer.

Therefore, the multistage decision individuals face for these screening tests is (a) whether to have prostate or breast cancer screening tests and, if so, when to have them done; (b) whether to undergo an invasive diagnostic procedure after a potentially false-positive test result; and (c) how to proceed if something is detected that may not lend itself to standard treatment options. A dynamic decision theory could be used to guide patients through this process.

Strategies for Success

According to dynamic decision theories, two tasks are crucial for success: goal setting and information collection. The most accomplished dynamic decision makers are able to integrate the goals of the decision-making task with the current state of the environment in order to identify tactics that have worked in analogous situations from their past. If no such situations exist, they are able to generate strategies using problem-solving techniques. Second, they systematically gather information relevant to achieving their goals. Third, they continually evaluate their advancement toward their goals.

Those who are less successful tend to shift from one goal to another or focus too narrowly on a single goal. To improve performance, dynamic decision theories suggest three strategies. First, constrain information processing. This may be accomplished by asking patients to focus on the next two or, at most, three goals rather than thinking about everything at once, which is the tendency of many newly diagnosed patients. Second, encourage a more focused information-gathering strategy, perhaps by pointing patients toward specific educational materials or online resources, as the nearly endless amount of medical information available both in print and online can quickly become overwhelming.

Finally, if patients do not have relevant past experiences to inform their decision making, introduce them to more experienced others from whom they can learn, such as former patients who have successfully completed their treatment. This may help patients envision what it is like to face the decisions they are contemplating and to experience the outcomes. By doing so, they may be better able to anticipate the "hot" or more emotional state of mind they are likely to be in as they get closer to making their treatment choice.

Julie Goldberg

See also Biases in Human Prediction; Decision Psychology; Decisions Faced by Patients: Primary Care; Gain/Loss Framing Effects; Gambles; Hedonic Prediction and Relativism; Managing Variability and Uncertainty; Preference Reversals; Prospect Theory; Risk Attitude; Value Functions in Domains of Gains and Losses

Further Readings

Barkan, R., & Busemeyer, J. R. (2003). Modeling dynamic inconsistency with a changing reference point. In "Time and decision" [Special issue]. *Journal of Behavioral Decision Making, 16*(4), 235–255.

Busemeyer, J. R., & Townsend, J. T. (1993). Decision field theory: A dynamic-cognitive approach to decision making in an uncertain environment. *Psychological Review, 100*(3), 432–459.

Busemeyer, J. R., Weg, E., Barkan, R., Li, X., & Ma, Z. (2000). Dynamic and consequential consistency of choices between paths of decision trees. *Journal of Experimental Psychology: General, 129*(4), 530–545.

Johnson, J., & Busemeyer, J. R. (2001). Multiple-stage decision-making: The effect of planning horizon length on dynamic consistency. *Theory and Decision, 51*(2–4), 217–246.

Roe, R. M., Busemeyer, J. R., & Townsend, J. T. (2001). Multialternative decision field theory: A dynamic connectionist model of decision making. *Psychological Review, 108*(2), 370–392.

Dynamic Treatment Regimens

A dynamic treatment regimen (DTR) is a sequence of individually tailored decision rules that specify whether, how, and when to alter the intensity, type, or delivery of treatment at critical decision points in the medical care process. DTRs operationalize sequential decision making with the aim of improving clinical practice. Ideally, DTRs realize this goal by flexibly tailoring treatments to patients when they need it most, thereby improving the efficacy

and effectiveness of treatment and reducing inappropriate variance in treatment delivery. DTRs can be used to develop clinical guidelines, including clinical decision support systems. All the following are types of DTRs: (a) structured treatment interruptions in the HIV/AIDS literature; (b) clinical strategies, treatment strategies, or treatment algorithms in the psychiatric disorders literature; (c) adaptive therapy or multiple-treatment courses in the cancer literature; and (d) adaptive treatment strategies, stepped-care models, or continuing-care models in the alcohol and other substance abuse treatment literature. A variety of statistical methods exist to inform the development of DTRs.

Structure

A DTR consists of four key ingredients. The first ingredient is a *sequence of critical decision points* in the medical care process. These decision points may represent time in the form of patient visits to the clinic (first visit, second visit, and so on); or if critical decisions are to be made on a monthly basis, the decision points may represent calendar time in months since disease diagnosis. More generally, though, the sequence of critical decision points is not required to be aligned with a prespecified set of discrete time points. For example, critical decision points may, instead, be defined by patient events, such as the point at which a patient fails to respond to prior treatment.

The second ingredient is a set of one or more *treatment options* at each critical decision point. Possible treatment options may be switch medication, augment medication, or continue medication; or there may be more complex options, such as any of the three-way combinations of treatment type (medication, physical therapy), treatment intensity (high, medium, low), and treatment delivery (specialty clinic, general clinic). The set of potential treatment options may differ at different decision points. For example, initially, the emphasis may be on treatment suitable for an acute episode of the illness, whereas subsequent decisions may involve options for intensifying or augmenting treatment for nonresponding patients or transitioning to lower-intensity treatments or monitoring for responding patients.

The third ingredient is a set of one or more *tailoring variables* at each critical decision point. The tailoring variables form the set of key measures that will determine subsequent treatment. For example, tailoring variables may include patient severity, number and type of comorbidities, side effects resulting from prior treatment, treatment preference, adherence to prior treatment, and, perhaps most important, response to prior treatment. Tailoring variables can also be summary measures over the full course of prior treatment; for example, subsequent treatment could depend on the rate of improvement in symptoms during prior treatment or the pattern of nonadherence to prior treatment. The set of tailoring variables may differ at different time points; for instance, history of comorbidities or genetic background may be used to choose from the options for initial treatment, while the choice of subsequent treatment might be based on response to the present treatment and the type of present treatment.

The final ingredient in a DTR is the specification of a *decision rule* at each of the critical decision points. For every patient and at each time point, the decision rule inputs values of the tailoring variables and outputs one or more recommended treatments from the set of treatment options. Importantly, the decision rules specify recommended treatment(s) for every feasible level of the tailoring variables. In the context of treatment for alcohol abuse, for example, a decision rule may state that as soon as the patient incurs 2 or more heavy drinking days following initiation of the medication, augment the medication with one of a set of cognitive behavioral therapies; otherwise, if the patient incurs less than 2 heavy drinking days during the 8 weeks following initiation of the medication, then keep the patient on medication and provide telephone disease monitoring.

The full set of decision rules over all of the critical decision points, taken together, constitutes one DTR. From the patient's point of view, a DTR is a sequence of treatments over time. This sequence of treatments is dynamic and patient specific because it is tailored in response to the patient's variable and evolving clinical status.

Clinical Settings

DTRs can be used to enhance clinical practice in any clinical setting in which sequential medical decision making is essential for the welfare of the

patient. In settings in which treatment response is widely heterogeneous and/or patients are insufficiently responsive to any one treatment, clinicians must often consider a series of treatments to achieve a desired response. Furthermore, in settings in which relapse rates are high, treatment decisions during the acute phase of the disease are often followed by decisions concerning the best suitable treatment to prevent subsequent relapse. The treatment of many chronic disorders such as cardiovascular disease, HIV/AIDS, cancer, diabetes, epilepsy, obesity, substance abuse disorders, mental disorders, and behavioral disorders require these types of sequential decisions. Furthermore, since chronic disorders are often characterized by a waxing-and-waning course, it is important to reduce treatment burden by reducing treatment intensity whenever possible. DTRs are ideally suited for these settings because they can be designed to respond over time to the changing course of a patient's illness.

Development

Currently, DTRs are formulated using a combination of expert opinion, clinical experience, and biological/behavioral theory. Either by scientific consensus or by relying on more quantitative methods (e.g., meta-analyses), scientists using this approach rely on summarizing the results of separate randomized trials to inform their view about DTRs. This strategy does not involve research designs or data-analytic methods designed explicitly for the purpose of developing DTRs.

A variety of statistical methods currently exist that can be used to inform the development of DTRs. These methods can be used either with longitudinal data arising from specialized trials designed to inform their development or with existing longitudinal data sets. These tools are used in conjunction with clinical experience and biological/behavioral theory to arrive at recommended DTRs for implementation in clinical practice.

Basic Structure and Sources of Data

Data Structure

To be useful for developing a DTR, a data set must have both treatment measures and potential tailoring measures (or time-varying covariates) observed at each of the critical decision points. In addition, the data set should have (possibly time varying) measures that define a clinically meaningful primary outcome measure. The choice of the primary outcome is crucial because the DTR will be developed explicitly to improve (or optimize) this outcome variable. In most cases, the primary outcome is a summary measure of response to treatment over time. For example, the outcome variable may be the percentage of time in remission over the full (dynamic) treatment course, or the outcome may involve a measure of functionality or may even be a composite measure involving cost and patient burden.

Existing Longitudinal Data

Longitudinal data sets having the characteristics described above are commonly collected as part of observational studies or can be extracted from large medical databases. In addition, longitudinal data sets of this type may arise from experimental studies. These include intervention studies that randomize patients to one of two (or more) single-shot treatments at baseline and follow them repeatedly over time, measuring the actual receipt of the assigned treatment and other treatments as well as a variety of other outcomes and time-varying covariates.

One of the primary challenges with using existing longitudinal data sets is the likely existence of unknown or unobserved, fixed or time-varying variables that affect both actual treatment receipt and the primary outcome. These variables confound (bias) the comparisons of different treatment regimens and present an important obstacle in data analyses aimed at informing the development of DTRs.

Sequential Multiple-Assignment Randomized Trials

Sequential multiple-assignment randomized trials (SMARTs) have been proposed explicitly for the purpose of developing new DTRs or refining already established ones. The key feature of a SMART is that patients are randomized multiple times over the course of the trial; that is, they are randomized at each critical decision point among feasible treatment options. Randomizing patients multiple times in this fashion ensures comparability among patients assigned to different treatment

options at each time point, thereby resolving the problem of confounding described earlier.

Statistical Methods

A variety of statistical models and methods are currently available that allow researchers to compare the effectiveness of different decision rules, examine the effect of different timing and sequences of treatments, and discover the important tailoring measures for use in a DTR. These methods can be used with data arising from a SMART or with existing longitudinal data sets. They include the marginal mean model and the structural nested mean model, and adaptations of them; Bayesian methods; methods to discover DTRs connected with time-to-event outcomes; and methods designed explicitly for discovering optimal DTRs. Recently, as well, methods and models from computer science, called *reinforcement learning algorithms*, are emerging as viable options for informing the development of DTRs.

Susan A. Murphy and Daniel Almirall

See also Decision Rules; Dynamic Decision Making; Evaluating and Integrating Research Into Clinical Practice; Evidence-Based Medicine; Expert Systems

Further Readings

Gaweda, A. E., Muezzinoglu, M. K., Aronoff, G. R., Jacobs, A. A., Zurada, J. M., & Brier, M. E. (2005). Individualization of pharmacological anemia management using reinforcement learning. *Neural Networks, 18,* 826–834.

Lavori, P. W., & Dawson, R. (2004). Dynamic treatment regimens: Practical design considerations. *Clinical Trials, 1,* 9–20.

Lunceford, J., Davidian, M., & Tsiatis, A. A. (2002). Estimation of the survival distribution of treatment regimens in two-stage randomization designs in clinical trials. *Biometrics, 58,* 48–57.

Murphy, S. A. (2003). Optimal dynamic treatment regimens. *Journal of the Royal Statistical Society, Series B, 65*(2), 331–366.

Murphy, S. A. (2005). An experimental design for the development of adaptive treatment strategies. *Statistics in Medicine, 24,* 1455–1481.

Pineau, J., Bellemare, M. G., Rush, A. J., Ghizaru, A., & Murphy, S. A. (2006). Constructing evidence-based treatment strategies using methods from computer science. *Drug and Alcohol Dependence, 88,* S2, S52–S60.

Robins, J. M. (2004). Optimal structural nested models for optimal sequential decisions. In D. Y. Lin & P. Haegerty (Eds.), *Proceedings of the 2nd Seattle symposium on biostatistics* (pp. 189–326). New York: Springer-Verlag.

Thall, P. F., Logothetis, C., Pagliaro, L. C., Wen, S., Brown, M. A., Williams, D., et al. (2007). Adaptive therapy for androgen-independent prostate cancer: A randomized selection trial of four regimens. *Journal of the National Cancer Institute, 99,* 1613–1622.

Thall, P. F., Millikan, R. E., & Sung, H. G. (2000). Evaluating multiple treatment courses in clinical trials. *Statistics in Medicine, 19,* 1011–1028.

Wahed, A. S., & Tsiatis, A. A. (2004). Optimal estimator for the survival distribution and related quantities for treatment policies in two-stage randomization designs in clinical trials. *Biometrics, 60,* 124–133.

E

Economics, Health Economics

Health economics investigates how scarce resources are used, or should be used, to satisfy health wants. Although in high-income countries 10% or more of wealth is spent on healthcare, resources are still scarce compared with the potentially unlimited want for physical, psychological, and social health.

The health market is not a competitive market in which price setting resolves differences between demand and supply, among others because insurance interferes with the relation between price and demand; new healthcare providers are often not free to enter the market and patients do not have perfect information about their needs. As a result, active decision making may be necessary to prevent supply of inefficient healthcare. Health economics intends to provide information on the economic aspect of such decision making.

Health economics includes several fields such as organization, management, finance, and insurance. Most relevant for medical decision making is the field of economic evaluation, which investigates whether costs of different medical decisions are justified by the value of associated effectiveness. Economic evaluation is a component of the wider research field of health technology assessment, which also includes evaluation of ethical, social, and legal aspects. It has its roots in evidence-based medicine, in trying to derive conclusions from explicit and judicious use of the best available evidence.

Types of Analysis

Economic evaluation in healthcare requires that costs and effectiveness of interventions are somehow measured and analyzed. Different types of analysis can be distinguished, depending on how costs are related to effectiveness. Two types of cost analysis that do not compare decision alternatives are cost price analysis and cost of illness analysis. Cost price analyses estimate the costs of a particular intervention. They are an essential starting point for economic evaluations, but by themselves they are often only part of the picture. Cost of illness analyses estimate the costs associated with a particular illness or condition, without comparing decision alternatives. As a result, they are largely irrelevant to decision making: How high costs are is not necessarily linked to whether these costs are justified.

For medical decision making, analyses that explicitly compare decision alternatives are more relevant: cost-minimization analysis (CMA), cost-consequence analysis (CCA), cost-benefit analysis (CBA), cost-effectiveness analysis (CEA), and cost-utility analysis (CUA). These types of analyses differ in how costs are compared with effectiveness. CMA only looks at which alternative is the least expensive, without considering the effectiveness. It is therefore only applicable when effectiveness is known to be equal for all alternatives, for example, when different ways to provide the same care are compared. CCA provides a list of both cost and effectiveness outcomes, but without explicitly combining these outcomes: The overall judgment is left to the decision maker. CBA, CEA, and CUA

do explicitly combine costs and effectiveness, to suggest which decision alternative provides best value for the money. They differ in how effectiveness is quantified. CBA measures effectiveness by its monetary value, for example, by asking patients how much they would be willing to pay for effectiveness. Converting effectiveness to money is problematic, but does facilitate a direct assessment of whether the value of effectiveness exceeds the costs. CEA measures effectiveness in physical units, rendering cost-effectiveness ratios such as the costs per identified cancer patient, costs per prevented death, or costs per gained life year. CEAs can be used to compare the relative efficiency of interventions with the same goal, but are not useful for a more general framework for economic assessment across the wide field of healthcare. For that purpose, CUAs are advocated. CUAs are a special case of CEAs, measuring effectiveness in terms of quality-adjusted life years (QALYs). QALYs measure the two general goals of healthcare: to prolong life and to improve life.

Measuring Value of Effectiveness

Effectiveness of medical interventions can be measured in many ways. Intermediary outcome measures, such as cholesterol levels, bone density, and cancer recurrence, are relatively easy to measure and are essential to understanding how interventions work. However, to assess whether an intervention actually helps improve patients' health requires measures for disease burden, which includes both survival and quality of life. In addition, economic evaluation requires measuring the *value* of improving survival and quality of life.

Key to measuring value in economic evaluations is the concept of utility, which is the value of quality of life at a particular moment in time. Utility is measured on a scale anchored at 0 (*as bad as death*) and 1 (*perfect health*). It may even be less than 0 (for quality of life worse than being dead). Since utility tries to aggregate the multifaceted concept of health into a single index, measuring utility is not without problems. The simplest approach is to ask respondents to indicate the overall value of their quality of life on a visual analog scale (Figure 1). Other, more complicated, utility assessment techniques (such as the time trade-off and the standard gamble) are considered more valid ways to directly assess utility, because they value quality of life compared with some other commodity (lifetime and mortality risk, respectively). These direct methods can be used to assess utility from the patients' perspective. Indirect utility measures, such as the EQ5D, HUI, and SF6D, ask the respondents not to value their health but to describe their health on a classification system. An existing formula is then used to assign a utility value to that description. Such formulae reflect the general public's valuation of the health described by the patient, which is preferred for economic evaluations from a societal perspective.

Life expectancy has long been an accepted measure of health. QALYs combine life expectancy with utility, to obtain a single generic value measure for both survival and quality of life. QALYs measure the value of a patient's health over a period of time by the product of the length of that period and the average utility during that period. This is equivalent to measuring the area under the utility curve. This way, both prolonged life and improved quality of life lead to higher QALYs. Conceptually, QALYs are very similar to DALYs (disability-adjusted life years) and Q-TWiST (quality-adjusted time without symptoms or toxicity).

As a schematic example of QALYs, consider the course of life shown in Figure 2. This person had depression from age 20 to age 40, contracted cancer at age 74, and died at age 80. The depression led to an average 25% utility loss over a 20-year period, which corresponds to a loss of 5 QALYs. The cancer period of 6 years has an associated QALY loss of 3 QALYs. Therefore, adjusted for quality of life, the 80 life years correspond to 72 QALYs. Measured in terms of QALYs, postponing cancer and death by 2 years would gain 2 QALYs. Reducing the severity of depression by 50% would gain 2.5 QALYs.

Measuring Costs

In economic evaluation, costs represent the monetary value of the investments that are associated with a particular medical decision. An important first step is to determine the relevant perspective of the cost evaluation, which can be the healthcare perspective, the societal perspective (including productivity and patient costs), or a particular institution (such as a hospital or insurer). The perspective determines not only which cost categories should

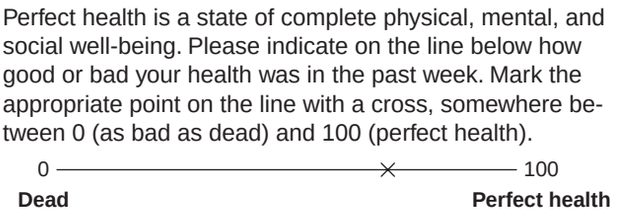


Figure 1 Visual analog scale (indicating a 70% utility)

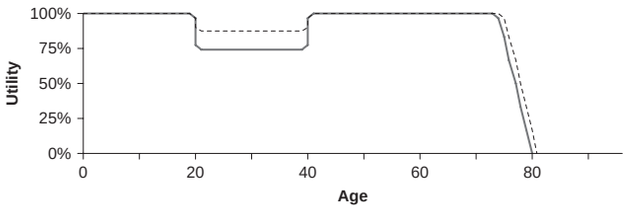


Figure 2 A schematic example of utility throughout life

be included but also how they should be valued. In the end, it is the differences in costs between the decision alternatives that need to be estimated, so all cost categories that can be expected to show an appreciable cost difference should be measured.

An essential part of a cost evaluation is usually the costs of a primary intervention. For this intervention, a detailed cost price analyses should be performed, including costs of personnel, equipment, materials, housing, and overhead. For other cost categories, standard prices or cost estimates from the literature can be used. For evaluations from an institutional perspective, charges may be relevant, but it should be realized that charges are not necessarily good approximations of costs. For example, costs of radiotherapy are partly proportional to the number of sessions and partly fixed per treatment, whereas charges for radiotherapy are often either fixed per session or fixed per treatment. When the number of sessions per treatment is changed, then estimating costs from charges per session or from charges per treatment will, respectively, overestimate and underestimate the impact on costs.

For many types of costs, costs can be distinguished as the product of volumes and prices. Volumes, such as the number of GP (general practitioner) visits or days absent from work, are more generalizable to other settings than costs. Patients can be asked to report volumes, using diaries, questionnaires, or interviews. They are aware of all the care they receive but may have difficulty in accurately remembering less salient types of care. Providers of care can rely on the accuracy of information systems but can only report on care that they themselves are involved in.

Study Designs

Typically, two types of study designs are used for economic evaluations. On the one hand, there is research measuring costs and effectiveness in one single patient population. On the other hand, there are modeling studies, aggregating data from different sources.

In patient research, data should ideally originate from research in which patients are first selected and then randomly allocated to the different decision alternatives. This procedure ensures that the decision alternatives are all applied to the relevant patient population, without selection bias. Measuring costs and effectiveness in a single patient population is important to provide internal validity of the research. For external validity, it is important to use a pragmatic design with conditions that are close to those in practice, in how treatments are provided and to which patients. Typical for pragmatic trials is that it is more relevant to study whether and how much a treatment helps than why.

For many reasons, performing patient research to compare decision alternatives may not be feasible. The number of alternatives may be too large (e.g., when evaluating follow-up strategies), the differences between alternatives may be too small (to be demonstrated with the number of patients available), one of the decision alternatives may be generally considered unethical (obstructing new research), or the time to make a decision may be too limited (more limited than the duration of patient follow-up). In such situations, mathematical models may help evaluate decision alternatives and to aggregate effectiveness and cost data, obtained from different sources. Models can have varying degrees of detail, ranging from aggregate epidemiological models to patient-level models for day-to-day disease progression, and can be evaluated with techniques ranging from spreadsheet calculations and regression models to microsimulation. The use of models allows for sensitivity analysis to see how model parameters influence the conclusions, in order to validate the model's reliability for supporting decision making.

Cost-Effectiveness Analysis

Once costs and effects of different decision alternatives have been determined, CEA is used to decide which decision is optimal. CEA is intrinsically two-dimensional. When comparing two decision alternatives, one option is clearly preferred over the other alternative if it has lower costs and better effectiveness. The decision becomes difficult when one option is preferred based on better effectiveness and the other is preferred based on lower costs. In that case, a trade-off needs to be made between costs and effectiveness, to decide whether the more expensive decision alternative is justified by its better effectiveness.

Because of their two-dimensional nature, cost-effectiveness results are best presented graphically. Figure 3 shows costs and effectiveness for five different decision alternatives. Alternative A is said to be (strongly) dominated by Alternatives B, C, and D, because A has higher costs and lower effectiveness. As a result, Alternative A will not be the optimal decision, at least with respect to the economic aspect.

Alternative B is not dominated by any of the other alternatives, but it is dominated by a mixture of Alternatives C and D. This type of dominance is called weak, or extended, dominance. If Alternatives C and D were both applied to half of the patient population, then overall effectiveness and costs would be (.3 + .8)/2 = .55 and ($5,000 + $20,000)/2 = $12,500. Alternative B is (strongly) dominated by this 50:50 mixture of Alternatives C and D. The straight line CD between Alternatives C and D depicts the results that would be obtained by all possible mixtures of Alternatives C and D. The lines CD and DE together form the so-called efficient frontier. All alternatives above or to the left of this frontier are strongly or weakly dominated. All possible optimal alternatives are on the efficient frontier.

Which alternative on the efficient frontier is optimal depends on how much one is willing to pay to improve effectiveness. Cost-effectiveness should always be considered incrementally, that is, compared with the next best alternative. Compared with Alternative C, Alternative D provides .5 additional units of effectiveness and $15,000 additional costs, with a cost-effectiveness ratio of $15,000/.5 = $30,000 per unit. Similarly, the cost-effectiveness ratio comparing Alternatives D and E is $30,000/.1 = $300,000 per unit. The improvement by Alternative E is 10 times more expensive than the improvement by Alternative D, but without specifying the effectiveness measure, it is impossible to say which alternative is optimal. If effectiveness measures prevented mortality, then $300,000 per prevented death is likely to be acceptable and Alternative E would be optimal. If effectiveness measures prevented days with the flu, then $30,000 per day is unlikely to be acceptable and Alternative C would be optimal.

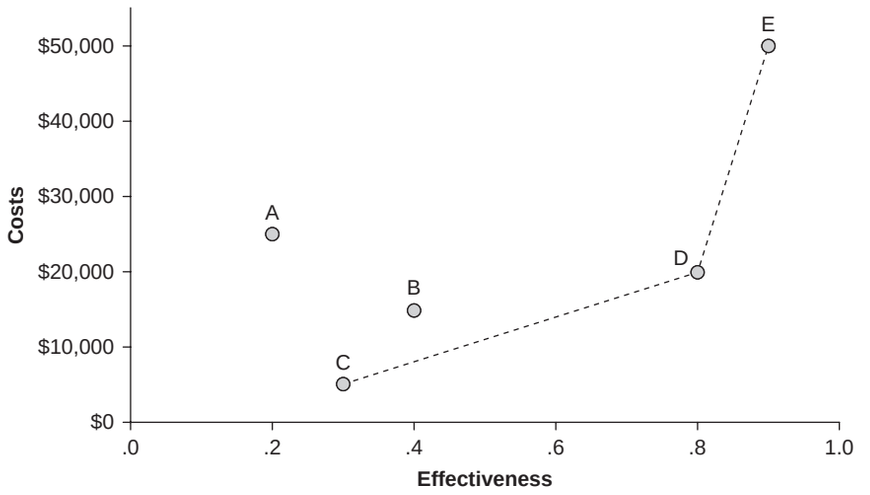


Figure 3 Cost-effectiveness plane

The economic aspect of a decision is rarely the only relevant aspect for decision making. Therefore, no strict thresholds exist for how much improved effectiveness is allowed to cost. Nevertheless, for effectiveness measured in terms of QALYs, there is some consensus on the rule of thumb that costs are definitely acceptable below $20,000 per QALY, are acceptable up to $50,000 per QALY, and are possibly acceptable up to $100,000 per QALY. According to this rule, Alternative D would be optimal: It provides good value for the money compared with Alternative C, and the costs of Alternative E would be too high.

Wilbert van den Hout

See also Cost-Effectiveness Analysis; Disability-Adjusted Life Years (DALYs); Evidence-Based Medicine; Marginal or Incremental Analysis, Cost-Effectiveness Ratio; Quality-Adjusted Life Years (QALYs); Quality-Adjusted Time Without Symptoms or Toxicity (Q-TWiST); Randomized Clinical Trials; Utility Assessment Techniques

Further Readings

Briggs, A., Sculpher, M., & Claxton, K. (2006). *Decision modelling for health economic evaluation*. New York: Oxford University Press.

Drummond, M. F., O'Brien, B. J., Stoddart, G. L., & Torrance, G. W. (1997). *Methods for the economic evaluation of health care programmes* (2nd ed.). New York: Oxford University Press.

Gold, M. R., Siegel, J. E., Russell, L. B., & Weinstein, M. C. (1996). *Cost-effectiveness in health and medicine*. New York: Oxford University Press.

Editing, Segregation of Prospects

In medical decision making, a prospect can be a medical treatment that will yield different outcomes with different probabilities. When patients are offered multiple treatment options, they will have to make a decision and follow one treatment that they think is the best. This selection process includes editing and segregation of the different prospects. A prospect $(x_1, p_1; \ldots; x_n p_n)$ is a contract that yields outcome x_i with probability p_i, where $p_1 + p_2 + \ldots + p_n = 1$. To simplify this notation, we omit null outcomes and use (x, p) to denote the prospect $(x, p; 0, 1 - p)$, which yields x with probability p and 0 with probability $1 - p$. The riskless prospect that yields x with certainty is denoted by (x).

Within prospect theory, there are two distinct phases in the decision-making process: an early phase of editing and a subsequent phase of evaluation. The editing phase consists of a preliminary analysis and process of the offered prospects, which often yields a simpler representation of these prospects. In the second phase, the edited prospects are evaluated, and the prospect of highest value is chosen.

The objective of the editing phase is to organize and reformulate the options so as to simplify subsequent evaluation and choice. Editing is a mental process that transforms the probabilities of the various prospects. In medical decision making, the patient will edit the prospect of every treatment and then segregate the obvious undesirable treatments from the others according to the preliminary results from the editing phase.

Editing can be divided into six separate phases: (1) coding, (2) combination, (3) segregation, (4) cancellation, (5) simplification, and (6) detection of dominance.

Coding

Patients normally perceive the treatment outcomes as gains and losses, rather than as final states of health or life quality. This coding process will primarily rely on the choice of reference point. The reference point usually corresponds to the patients' current health level, in which case gains and losses can be interpreted as improvement or deterioration of their current health level. For example, consider a percentage as an indicator of people's health level, with 100% as *healthy* and 0% as *dead*. Also, consider the case of two patients, with health levels of 20% and 70%, respectively. Both patients are offered a treatment that provides an 80% chance of achieving an 80% health level and a 20% chance of decreasing to a 10% health level (.8, .8; .1, .2). Although the treatment is the same, the two patients would code this prospect differently. The first patient (current health level of 20%) will regard this as a good, acceptable gain choice, as his reference level is low and this treatment will increase his

health level remarkably with a high probability. However, the second patient (current health level of 70%) would code this as a losing choice because she has a relatively high reference point. Therefore, she can increase her health level only trivially, while facing a 20% probability of losing most of her health.

In real medical decision-making cases, patients' reference points are usually influenced and shifted because of their expectation of the treatment, which comes from the prediagnosis, and their adaptation to the prognosis, which will also change their coding results over time.

Combination

Prospects can sometimes be simplified by combining the probabilities associated with identical outcomes. For example, a single treatment will result in four health outcomes with probabilities .1, .2, .3, and .4. The four health outcomes result in a life expectancy of 5, 10, 5, and 10 years, respectively. The patients will then combine the prospect into simply 5 years of life expectancy with probability .4 and 10 years of life expectancy with probability .6 (5, .4; 10, .6).

Segregation

Some prospects contain a riskless component along with an uncertain component. Patients can mentally segregate the risky part and simplify the decision making by addressing only the risky part. For example, a treatment has only two outcomes: (1) increase in life expectancy to 20 years with a probability of .3 and (2) increase in life expectancy to 30 years with a probability of .7. This can be naturally decomposed into a sure gain of 20 years of life expectancy and a risky prospect of 10 years of life expectancy with probability of .7 (10, .7).

Cancellation

Most patients tend to discard or exclude some components that are shared by the offered prospects. They rely on the cancellation of the common parts of the two prospects to help them make decisions. For example, consider two treatments. Both of them have a probability of success of .25. After the treatment is successful, Treatment A has an 80% chance of increasing life expectancy by 30 years (30, .8) while Treatment B has a 100% chance of increasing life expectancy by 20 years (20, 1.0). When facing this choice, most patients will ignore the precondition that both treatments have a 25% success rate, which is shared by both prospects. After they edit the two prospects, most of the patients will choose Treatment B. Interestingly, however, if we edit the prospects differently and apply the precondition, it will be a choice between (30, .20) and (20, .25), in which case most patients will choose Treatment A.

Simplification

This refers to the simplification of prospects by rounding probabilities or outcomes. For example, the prospect (101, .49) is likely to be recoded as an even chance to achieve 100.

Detection of Dominance

Many prospects may be dominated by the others. This can be mentally detected so that the dominated prospects will be segregated from the potential choices of patients. For example, Treatment A will achieve 10 years of life expectancy with a probability of .3 and 20 years of life expectancy with a probability of .7 (10, .3; 20, .7). Treatment B will achieve 8 years of life expectancy with a probability of .5 and 15 years of life expectancy with a probability of .5 (8, .5; 15, .5). It is obvious to the decision maker that Treatment A dominates Treatment B. Therefore, B is discarded without any further consideration.

One thing that needs to be stressed is that the editing process will vary between people, as well as within a single decision maker. The process is dynamic and highly dependent on the context of the problem, thus producing different results.

After editing the prospects, people will form a mental representation of all the existing prospects, which will segregate the prospects that they think are undesirable from the prospects that they would like to further evaluate. This editing and segregation phase simplifies the process of decision making and preclude some prospects so that the decision makers will reach their decision much more easily.

Lesley Strawderman and Yunchen Huang

See also Expected Utility Theory; Probability; Prospect Theory

Further Readings

Emma, B. R., Kevin, P. W., & Kevin, A. S. (2005). Can prospect theory explain risk-seeking behavior by terminally ill patients? *Medical Decision Making, 25,* 609–613.

Jonathan, R. T., & Leslie, A. L. (1999). Health value and prospect theory. *Medical Decision Making, 19,* 344–352.

Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica, 47,* 263–291.

Effect Size

In statistics, an effect size is a measure of the magnitude of a treatment effect. It is an indicator of how important an obtained effect is. Unlike statistical significance tests, effect size does not depend on the sample size of an underlying study. It is helpful to report the effect size, not just the statistical significance, when assessing the effectiveness of a specific intervention in medical studies as well as studies in other sciences. It has been also widely used in meta-analysis, which combines and compares estimates from different but relevant studies.

In medical studies, such as comparison of a new treatment with other traditional ones, the following question is often asked: How well does the new treatment work? In answering this question, the researchers are actually trying to quantify the difference between the effect of the new treatment and those of the traditional ones. Similar things happen in social studies and studies in educational and behavioral sciences. Effect size is a simple way of answering the question, and it has many advantages over the use of tests of statistical significance alone. Effect size measures directly the size of the difference rather than confounding this with the sample size of the study. It is easy to calculate and to interpret, and it can be applied to any measured outcome of medical, social, and educational sciences to quantify the effectiveness of a particular intervention in comparison with others.

Effect Size for Two Independent Groups With Continuous Outcomes

Let us consider comparing the outcome of two groups, the experimental group (the one for which a new treatment is going to be applied) and the control group (the one for which a traditional treatment is going to be applied). The outcome of the study is a kind of continuous measurement. The effect size in such a case is defined as the standardized difference in means between the two groups. In other words,

$$\text{Effect size} = \frac{[\text{Mean of experimental group}] \quad [\text{Mean of control group}]}{\text{Standard deviation}}.$$

It is very natural to take the difference of two group means when comparing the two groups of measurements. The standard deviation in the denominator, which is a measure of the spread of a set of values, is to standardize this difference. The same value of difference may represent totally different meanings when the standard deviations are different. It could be explained as a huge difference if the standard deviation is small, such that the two groups of values are completely separated; whereas if the corresponding standard deviation is large, the two sets of values might be well overlapped and the same value of difference might mean just nothing. The difference in means is standardized when it is divided by the standard deviation. In practice, however, the standard deviation is not known. It can be estimated either from the control group or from a pooled value of both groups.

The above-defined effect size is exactly equivalent to the z score of a standard normal distribution. For example, an effect size of 1 means that the average of the experimental group is 1 standard deviation higher than that of the control group. With the assistance of a graph of standard normal distribution curve, one can observe that the average of the experimental group, which is 1 standard deviation higher than that of the control group, is indeed the 84th percentile of the control group. In other words, 84% of the measurements of the control group are below the average of the experimental group. The value of 84% is calculated from the

standard normal distribution as the probability that the standard normal random variable is less than or equal to 1. In case the effect size takes different values, the underlying effect size will replace the value 1 in the calculation. Another percentage rather than 84% will be obtained correspondingly. This provides an idea of how the two groups overlap with each other.

Effect Size for Experiments With Dichotomous Outcomes

Effect size can be defined differently within different settings of the studies. Another commonly used effect size in medical studies is the odds ratio. When the experimental outcome is dichotomous—for instance, success versus failure, or survival versus death, the comparison of a new treatment with a control experiment can be conducted based on the odds ratio. If success and failure are the only two possible outcomes, the odds of success is defined as the ratio of the probability of a success to that of a failure. For each group, an odds can be calculated that equals the ratio of the number of successes to the number of failures in the group. The odds ratio is then defined as the ratio of the two odds. Let $n_{S,exp}$ denote the number of successes in the experimental group, $n_{F,exp}$ the number of failures. Let $n_{S,con}$ and $n_{F,con}$ denote the numbers of successes and failures in the control group, separately, then the odds ratio can be calculated as

$$\mathrm{OR} = \frac{\mathrm{ODDS}_{\mathrm{experimental}}}{\mathrm{ODDS}_{\mathrm{control}}} = \frac{n_{S,exp}/n_{F,exp}}{n_{S,con}/n_{F,con}}.$$

If the treatment effect is remarkable, the odds ratio should be much greater than 1. Otherwise, it should be very close to 1.

Examples

Example 1

Suppose that there was a study conducted to investigate the weekly weight gain of 3-month-old infants fed with different formulae. There were 20 infants randomly assigned to Group A and 30 infants assigned to Group B. The infants in Groups A and B were fed with Formulae A and B, separately. Formula B is newly developed. The infants were followed up for 4 weeks, and their individual average weekly weight gains (oz.) were recorded as

Group A:

10.41 10.38 9.16 10.01 11.07 10.47 10.18 9.59 8.77 9.75 8.37 8.95 10.27 9.70 11.61 9.43 8.36 10.23 10.23 9.69

Group B:

13.13 13.88 14.04 10.48 13.06 10.13 11.49 11.03 10.17 10.71 11.05 12.42 13.03 11.67 11.50 11.06 10.86 12.26 11.95 13.49 11.18 13.43 11.91 13.43 13.53 12.58 12.02 11.43 11.65 12.62

The mean of Group A is $\mu_A = 9.83$ oz. The mean of Group B is $\mu_B = 12.04$ oz. Group A has standard deviation $\sigma_A = .84$. Group B has standard deviation $\sigma_B = 1.13$. The pooled standard deviation of both groups is

$$\sigma_{\mathrm{pooled}} = \sqrt{\frac{(n_A-1)\sigma_A^2 + (n_B-1)\sigma_B^2}{(n_A+n_B-2)}} = \sqrt{\frac{(20-1)(.84)^2 + (30-1)(1.13)^2}{20+30-2}} = 1.03.$$

The effect size is then

$$\mathrm{ES} = \frac{\mu_B - \mu_A}{\sigma_{\mathrm{pooled}}} = \frac{12.04 - 9.83}{1.03} = 2.15.$$

If we treat Group B as the experimental group and Group A as the control group, the effect size of this treatment (Formula B) is 2.15. Looking at the standard normal distribution table, the value 2.15 corresponds to a probability of .9842. This tells us that about 98% or 19 of the 20 values observed from Group A are below the mean of Group B. This is a big effect size.

Example 2

Suppose that there was a medical study investigating the effect of a newly developed medicine. There were 100 patients assigned to a group where the new treatment (medicine) was applied, and 80 were assigned to the control group, where the placebo was applied. Within 2 weeks of this experiment, 80 patients from the experimental group and 45 patients from the control (placebo) group had been cured.

$$\text{OR} = \frac{\text{ODDS}_{\text{experimental}}}{\text{ODDS}_{\text{control}}} = \frac{80/20}{45/35} = 3.11.$$

This effect size is much greater than 1. It tells us that the patients assigned to the experimental group have a much better chance to be cured, or that the new medicine is very effective.

Alternative Measures of Effect Size

A number of statistics were proposed as alternative measures of effect size, other than the standardized mean difference and odds ratio. In studies that employ linear statistical models to analyze the experimental outcome, the effect size can be defined as the square of the correlation coefficient of the two involved variables, denoted by R^2. This measure is the proportion of variance in one variable accounted for by the other. It can extend automatically to the case of multiple regression models.

It can be shown that the effect size measured by standardized mean difference is sensitive to the assumption of normality of data. For this reason, many robust alternatives were suggested. Peter Tymms and colleagues proposed a method for calculating effect sizes within multilevel models. José Cortina and Hossein Nouri discussed the effect sizes in analysis of covariance designs and repeated measures designs. To understand different effect size measures under different models, the monograph of Robert Grissom and John Kim gives a comprehensive discussion on effect sizes.

Xiao-Feng Wang and Zhaozhi Fan

See also Meta-Analysis and Literature Review; Odds and Odds Ratio, Risk Ratio; Sample Size and Power; Statistical Testing: Overview

Further Readings

Bausell, R. B., & Li, Y. F. (2002). *Power analysis for experimental research: A practical guide for the biological, medical and social sciences*. New York: Cambridge University Press.

Cliff, N. (1993). Dominance statistics: Ordinal analyses to answer ordinal questions. *Psychological Bulletin, 114*(3), 494–509.

Cortina, J. M., & Nouri, H. (2000). *Effect size for ANOVA designs*. Thousand Oaks, CA: Sage.

Grissom, R. J., & Kim, J. J. (2005). *Effect sizes for research: A broad practical approach*. New York: Lawrence Erlbaum.

Murphy, K. R., & Myors, B. (2004). *Statistical power analysis*. Mahwah, NJ: Lawrence Erlbaum.

Rubin, D. B. (1992). Meta-analysis: Literature synthesis or effect-size surface estimation. *Journal of Educational Statistics, 17*(4), 363–374.

Tymms, P., Merrell, C., & Henderson, B. (1997). The first year at school: A quantitative investigation of the attainment and progress of pupils. *Educational Research and Evaluation, 3*(2), 101–118.

Efficacy Versus Effectiveness

The terms *efficacy* and *effectiveness* refer to different concepts and are not interchangeable. In general, efficacy refers to whether an intervention works under ideal conditions for a specific outcome. Effectiveness refers to a broader view of the usefulness of an intervention in the routine care of patients in the day-to-day practice of medicine. Efficacy is measured using controlled clinical trials, using specific outcome measures, such as prespecified changes in rating scales or laboratory parameters. Examples of efficacy studies are medication registration trials testing drug versus placebo. Effectiveness is measured by a variety of methods, including synthesis of efficacy and tolerability clinical trial data, clinical trials that incorporate broad outcomes such as quality of life, longitudinal prospective naturalistic studies, and retrospective studies using large-scale clinical, pharmacy, and administrative databases. Examples of effectiveness studies are studies examining all-cause discontinuation in the use of antipsychotics for the treatment of schizophrenia.

Efficacy

Efficacy refers to whether an intervention works under ideal conditions for a specific outcome. Regulatory agencies such as the U.S. Food and Drug Administration require that medications demonstrate efficacy prior to their approval for commercialization. These premarketing studies are referred to as drug registration trials and generally

aim to show superiority of the proposed agent versus placebo. This superiority is measured using a very specific outcome, such as reduction of symptoms using a rating scale designed and validated for that purpose, or a reduction in a laboratory measure, such as decrease in blood cholesterol levels. These clinical trials can be very large, enrolling multiple hundreds of patients across many study centers in several countries. Attempts are usually made to ensure a homogeneous test population. Intervention choice is randomized and subjects are followed double-blind. These clinical trials also monitor for adverse events, usually relying on spontaneous reporting but also including safety scales when certain tolerability problems are anticipated, such as extrapyramidal symptoms encountered with the use of antipsychotics. Clinical registration trial reports include information on both efficacy and tolerability under these artificial study conditions, but the aim of these reports is not to provide a synthesis for clinical guidance but to prove that the intervention is efficacious. Whether or not the intervention is efficacious and effective in a routine clinical practice is not certain. This is especially problematic when the patients who receive the intervention in clinical practice are unlike the subjects who received the intervention under controlled conditions. A good example of this are the registration trials of intramuscular antipsychotics for the treatment of agitation associated with schizophrenia or bipolar mania. Patients in these trials were required to provide informed consent and may represent a population that is very different from the agitated patient involuntarily brought to an emergency department by the police in terms of level of cooperation, degree of agitation, comorbid medical conditions, and presence of active alcohol or drug use. Perhaps the biggest objection to the use of registration trial data is that the comparator of placebo is not appropriate for the clinician whose main interest is to know how the new intervention compares with the old established one.

Effectiveness

Effectiveness is a term that refers to the broad utility of an intervention under the usual conditions of care. This utility includes efficacy (whether or not the intervention reduces the symptoms and signs of the disease), tolerability (whether or not the adverse events intrude on the well-being of the patient), and adherence (whether the patient complies with the treatment as prescribed). These three components are necessary for the intervention to be effective in the "real world." Efficacy is a necessary but not sufficient condition for an intervention to be useful.

Effectiveness can be estimated by the pooling together of clinical trial data that include information on both efficacy and tolerability. However, the predicted adherence or acceptability of the intervention in general clinical populations cannot be directly ascertained from this synthesis of efficacy studies. To accurately identify drug effects under the conditions of routine clinical care, different methods are needed. These methods include clinical trials that incorporate broad outcomes, longitudinal prospective naturalistic studies, and retrospective studies using large-scale clinical, pharmacy, and administrative databases. The subjects in effectiveness studies are usually more heterogeneous than those in a medication registration study, and this can facilitate the comparison of different active treatments.

An example of a controlled double-blind effectiveness trial is the Clinical Antipsychotic Trials of Intervention Effectiveness (CATIE) study for schizophrenia, where patients were initially randomized to one of five antipsychotics for up to 18 months. This is in direct contrast to the usual efficacy trial of an antipsychotic whose design is to compare a drug with a placebo over a relatively brief period ranging from 3 to 8 weeks. CATIE's primary outcome measure was time to all-cause treatment failure marked by discontinuation of the medication. The assumption was that if a medication was continued to be prescribed, then it was thought to be of acceptable value by both the patient and the clinician. The three principal reasons for discontinuation were patient decision, lack of adequate efficacy, or poor tolerability. The study included three main phases that allowed for switching from one antipsychotic to another. When enrolled, patients were made aware that these switches were possible. This mirrors clinical practice in that switching of antipsychotics is not uncommon. Moreover, unlike registration trials, subjects were not excluded if they had psychiatric comorbidities such as substance use disorders.

Effectiveness studies such as CATIE can answer questions that registration trials cannot, but there are several practical limitations to conducting large-scale effectiveness trials, including their length, size, and expense. Informed consent is also required, limiting generalizability. This patient selection bias can be extreme when studying chronic mental disorders such as schizophrenia, where impaired decisional capacity is not unusual. The use of naturalistic data from large-scale clinical and administrative databases produced by the ordinary, day-to-day operations of healthcare delivery systems is another option. Data for very large numbers of patients (thousands and tens of thousands) are available. Advantages include generalizability (the whole population across multiple diagnoses can be studied as they receive routine care). Multiple interventions or sequences of interventions can be assessed. The major limitation is the lack of randomization and the presence of substantial treatment selection biases (e.g., more chronically ill patients may receive different and/or multiple medications). Another criticism is that the retrospective analysis of databases is prone to data mining, where many outcomes are evaluated but only a select few are ever reported.

Evidence-Based Medicine

Clinicians often struggle to find interventions that make a difference in the well-being of their patients. It is not always easy to discern whether or not a study result should actually change clinical practice. Evidence-based medicine (EBM) is a philosophy that can help answer a clinical question that a practitioner may have about two different interventions for an individual patient. Clinical judgment and clinical expertise are still required to make the best decision possible, but the ability to formulate the question, seek out clinical trial evidence, appraise this evidence, and then to apply it and assess the outcome forms the nucleus of EBM. The evidence base can vary in quality, from anecdotal reports that are subject to bias, and hence of lower value, to the gold standard of randomized clinical trials and systematic reviews of randomized clinical trials. Both efficacy and effectiveness studies can help answer the clinical questions, but the limitations of each approach need to be understood. The clinician will need to identify evidence that can quantify the differences between treatments, ensuring that there are clinically significant differences. A discussion of effect sizes, such as the number needed to treat (NNT), is beyond the scope of this discussion but is integral to the clinical interpretation of efficacy and effectiveness studies.

Leslie Citrome

See also Confounding and Effect Modulation; Hypothesis Testing; Randomized Clinical Trials

Further Readings

Guyatt, G. H., & Rennie, D. (2001). *Users' guides to the medical literature: A manual for evidence-based clinical practice.* Chicago: AMA Press.

Jaffe, A. B., & Levine, J. (2003). Efficacy and effectiveness of first- and second-generation antipsychotics in schizophrenia. *Journal of Clinical Psychiatry, 64*(Suppl. 17), 3–6.

Lieberman, J. A., Stroup, T. S., McEvoy, J. P., Swartz, M. S., Rosenheck, R. A., Perkins, D. O., et al. (2005). Effectiveness of antipsychotic drugs in patients with chronic schizophrenia. *New England Journal of Medicine, 353,* 1209–1223.

Efficient Frontier

Efficient frontier is an economics term commonly used in performance measurement, although it has more recently also been applied to decision analysis. Another term for it is *production possibilities curve*. It shows the maximum output attainable from various combinations of inputs: the boundary between what is possible with the given resources and technologies and what is not.

Performance of Firms

Economists speak of *firms* or decision-making units, which convert a variety of inputs (materials, capital, and labor) into outputs. These outputs can be goods and/or services, and the firms may be public, for-profit, or not-for-profit. A variety of methods have been used to measure the performance of firms. One common measurement is the

productivity ratio, which is related to concepts of efficiency.

Defining Productivity and Efficiency

Productivity is defined as the ratio of outputs to inputs, both weighted by their prices. If it is possible to define the maximum output attainable from each input level, analysts can use this information to draw a production frontier. If a firm is operating on that frontier, they are classified as *technically efficient*; conversely, if they are beneath the frontier, they are technically inefficient. However, it may still be possible to move along the production frontier. One common way is to take advantage of economies of scale (or avoid diseconomies of scale). Technological changes may also shift the entire production frontier, allowing greater productivity for a given level of input.

Since most firms use multiple inputs, and produce multiple outputs, these must often be aggregated. Analysts may examine the productivity of particular inputs (e.g., labor productivity), but this can be misleading, particularly if substitution is possible. Total factor productivity refers to the productivity of all inputs used to produce the given outputs, each weighted by its price.

Productivity does not incorporate the costs of production but considers only the volume of outputs producible. *Allocative efficiency* for a given quantity of output is the term used to select the mix of inputs that will produce those outputs at minimum cost; this procedure assumes that the prices for the inputs are known and requires incorporating information about all firms that might produce the desired outputs. Depending on how broadly the outputs are defined, this may require assessment of the mix to produce a particular service (e.g., renal dialysis treatment), services within a particular sector (e.g., hospital care), or services across sectors within a society (e.g., trade-offs between education and healthcare). Data requirements increase with scope, such that determining allocative efficiency for an economy is extremely challenging. Total economic efficiency (also referred to as productive efficiency) must consider both technical and allocative efficiency. Economists will usually incorporate considerations of Pareto efficiency, defined as requiring that no alternative allocation of goods is possible without causing a net loss to one or more consumers.

These concepts are similar, but not identical, to cost-effectiveness. Unlike allocative efficiency, cost-effectiveness does not fix the desired output levels and mix; instead, it looks at the marginal cost to produce an additional marginal unit of benefit.

Measuring Productivity and Efficiency

Although some authors use the terms *productivity* and *efficiency* interchangeably, others stress that they have slightly different meanings and different operational definitions.

Economists have devised a number of methods for measuring efficiency and productivity. Many require computing index numbers, to allow analysts to compute productivity compared with a reference case. These relative measures of performance may look at how much more output could be produced for a given level of inputs (the output-oriented measures) or, conversely, at how little input would be required to produce a given level of outputs (input-oriented measures), as compared with the reference case.

One problem with using these efficiency measures is the high data requirements. They assume that the production functions of maximally efficient firms are known; this is rarely the case. Various approaches for estimating these functions using various techniques have been suggested; good reviews can be found in Coelli et al. and Worthington.

One approach is to use statistical techniques to construct a deterministic frontier, which is taken to represent the most efficient approach. Accordingly, any deviation from this frontier is assumed to represent inefficiency. This approach assumes that there is no noise and no measurement error. It also requires a large sample size (often not available) and sufficient spread of the observations throughout the distribution. Accordingly, it is used less commonly than the alternatives noted below.

Another family of approaches, stochastic frontiers, uses econometric models to estimate the frontier. These resemble the deterministic models in that they use parametric models that require assumptions as to the functional form but differ in introducing a disturbance term to allow for measurement error and noise. A third family, Data

Envelopment Analysis (DEA), uses linear programming techniques and is classified as nonparametric. Because this approach does not include stochastic (i.e., random) components, it assumes that all deviations from the frontier represent inefficiency. DEA, however, is more flexible than the alternatives in its data requirements and in how models are specified. Note that all these modeling approaches differ in the underlying assumptions made (e.g., whether it is assumed that firms are fully efficient), the ability to deal with noise and outliers, the assumptions about functional forms, and the data requirements. All are also subject to omitted variable bias.

Applications to Healthcare

Frontier efficiency has been applied to a wide range of firms, typically within public or quasi-public sectors. More recently, some efforts have been made to use these techniques to study the productivity of various healthcare organizations, including hospitals, nursing homes, and physician practices. An ongoing issue has been whether efficiency measures should be incorporated into reimbursement schedules and, if so, whether these approaches might be helpful in determining them.

One set of issues is defining what is meant by outputs. Technical efficiency may be used to refer to intermediate outputs such as the number of patients treated or their waiting time. It may also be defined in terms of health outcomes, such as mortality or life expectancy. Because health outcomes are related to many factors, often outside the healthcare system, analysts may have difficulty in defining the production function linking particular interventions to overall outcomes. This dilemma becomes even more pronounced if efforts are made to aggregate outputs (e.g., to look at the performance of a healthcare system, as opposed to the results of a particular drug or surgical procedure).

In 2004, Worthington identified 38 studies that applied frontier efficiency approaches to the study of healthcare organizations. Over half referred to organizations in the United States, although examples were found for Spain, Sweden, the Netherlands, Finland, Taiwan, and the United Kingdom. Most studies (68%) analyzed the performance of hospitals, with other examples examining nursing homes, health maintenance organizations, local-area health authorities, and other settings.

More recently, these approaches have been applied to decision analysis through the construction of a cost-effectiveness frontier. This analysis equates cost efficiency with the production of technically efficient combinations of inputs and outputs at the least cost. If it is possible to create a cost function, one can construct a production frontier that represents the best currently known production techniques. Accordingly, Eckermann and colleagues have recommended shifting the two-dimensional representation of cost-effectiveness from the commonly accepted incremental cost-effectiveness (which plots difference in effectiveness against difference in cost) to a production function approach. This application shares advantages, and disadvantages, with the previously noted efforts to use these methods.

Cautions

As Worthington cautions, this approach may not always be appropriate. One problem is how to ensure that studies do not compare apples with oranges. One way to ensure homogeneous outcomes is to aggregate; studies have accordingly categorized outputs in terms of age or type of treatment. As Newhouse noted, such aggregation can be problematic. Frontier techniques appear to be designed for homogeneous outputs, which is rarely true in healthcare. It is particularly difficult to capture variations in quality unless these lead to unambiguous impacts on the chosen measure (e.g., mortality). In general, many important outputs will not be included, and their omission is likely to distort the findings. Similarly, many inputs may be omitted (e.g., capital, physicians), and case-mix controls are likely to be inadequate. Hospitals treating sicker patients may thus be seen as being inefficient rather than as delivering a different mix of services.

Despite these caveats, frontier analysis is being more widely used by policy makers seeking to increase accountability in the use of public funds. This has been particularly evident in the United Kingdom. These techniques are being used as an alternative to the "performance indicator" movement; they seek to aggregate multiple indicators into a single measure of efficiency, based on the

difference between observed performance and that which would be predicted from the best case. A 2005 review by Jacobs and Street concludes that the approach is still not ready to be used to inform policy but recommends further research.

One key limitation to all these approaches is that they are not intended to deal with whether particular outputs are worth producing. Efficient markets assume that anything demanded should be produced as long as there are willing buyers and sellers. In contrast, appropriateness is a major concern for many healthcare services, and there is a widespread agreement that services that are not needed should probably not be provided, regardless of how efficiently they can be produced.

Raisa Deber and Audrey Laporte

See also Cost-Effectiveness Analysis; Cost-Identification Analysis; Cost-Minimization Analysis; Economics, Health Economics; Value Functions in Domains of Gains and Losses

Further Readings

Coelli, T., Rao, P. D. S., & Battese, G. E. (2002). *An introduction to efficiency and productivity analysis*. Boston: Kluwer Academic.

Eckermann, S., Briggs, A., & Willan, A. R. (2008). Health technology assessment in the cost-disutility plane. *Medical Decision Making, 28*(2), 172–181.

Jacobs, R., & Street, A. (2005). Efficiency measurement in health care: Recent developments, current practice and future research. In P. C. Smith, L. Ginnelly, & M. Sculpher (Eds.), *Health policy and economics* (pp. 148–172). Berkshire, UK: Open University Press.

Newhouse, J. P. (1994). Frontier estimation: How useful a tool for health economics? *Journal of Health Economics, 13*(3), 317–322.

Worthington, A. C. (2004). Frontier efficiency measurement in health care: A review of empirical techniques and selected applications. *Medical Care Research and Review, 61*(2), 135–170.

Emotion and Choice

It is increasingly recognized that emotions can have an important impact on judgment and decision making. However, in many respects, it remains an undeveloped area of judgment and decision making, particularly in medicine. First, emotions are difficult to characterize or define. Second, the causal mechanisms by which emotions influence decisions—independent of purely cognitive interactions—are poorly understood. Third, the circumstances in which emotions are most important in changing decisions are only partially understood. Finally, most of the well-controlled empirical data on emotions and decision making are outside the field of medicine, rarely involving physicians and patients. Each of these limitations is important when describing the role emotions play in medical decision making, so the sections that follow address each of these points in turn.

Defining Emotions

Clear definitions are crucial for outlining the role of emotions in judgment and decision making. Definitions or characterizations of emotion range across multiple disciplines. The philosopher Paul Griffiths proposes dividing what we commonly call emotions into two categories of mental phenomenon: lower-level "affect programs" and higher-level "irruptive emotional states." The first category of emotions consists of automated, stereotypical reactions that provide rapid responses to stimuli, seem rooted in evolutionarily justified patterns, are cross-cultural, and are correlated with survival needs in all higher animals. They are represented by the "lower" emotions of fear, anger, happiness, sadness, surprise, and disgust. The second category of emotions consists of those with complex mixtures of cognitive and emotional elements that occur more passively and interrupt other cognitive processes and tie together our mental lives in the long run. They are characterized by emotions such as love, guilt, envy, jealousy, and pride. They remain separate from other, more diffuse dispositional or visceral states referred to most accurately as "moods," such as anxiety, depression, and elation.

The political scientist Jon Elster presents a cluster of "features" that are robustly associated with human emotions, but none of which are essential to them. These features include being unbidden in occurrence, possessing cognitive antecedents, having intentional objects, being arousing, leading to action tendencies, and having specific valence. He

specifically distinguishes human emotions from emotions that have a sudden onset, brief duration, and characteristic expressions. These correspond to the affect programs Griffiths describes, which we largely share with other animals and across human societies and cultures.

The psychologists Reid Hastie and Robin Dawes define emotions as reactions to motivationally significant stimuli and situations that usually include three components: (1) a cognitive appraisal, (2) a signature physiological response, and (3) an accompanying phenomenal experience. This captures, at least operationally, the features of emotions that are most relevant for decision making.

Overall, there seems to be agreement that there are two groups of emotions. The first group consists of those that are more basic and stereotypical, are rooted most obviously in evolutionary survival, and suddenly interrupt ongoing cognition to cause different behavior. The second group consists of more complex cognitive states, with cognitive antecedents, less obviously tied to our evolutionary roots and less obviously interrupting other cognitive states. A third category, which is left aside here, consists of moods, which are more diffuse mental states that seem to be predispositions or precursors to other states and less obviously tied to specific actions.

Impact of Emotions on Decisions

There are two methodological approaches to decision making: the economic approach and the psychological approach. The economic approach emphasizes rationality, response to incentives, and maximization of utility (i.e., benefits) subject to constraints (i.e., costs). Such an approach minimizes the role emotions play in decision making, treating them as inputs to valuation or utility. Choice is fundamentally cognitive and rational, with a dispassionate consideration of costs and benefits. The psychological approach focuses on two mental operations, judgments and decisions, both of which can be, and often are, influenced by emotions. Psychologists identify persistent exceptions to rational behavior, showing how systematic biases shape human behavior. The maturing field of behavioral economics brings psychological realism and attention to human biases, including the impact of emotions, to the rational utility-maximization approach of economics. It is the approach taken here.

A causal framework for understanding the role of emotions in decision making from a behavioral economics perspective has been advanced by Loewenstein and Lerner. The framework highlights how emotions can influence decisions through two pathways: (1) immediate emotions and (2) expected emotions. Immediate emotions are experienced at the time a decision is made. These emotions can influence decision making in two ways: directly or indirectly. They can *directly* affect a decision as it is being made. For example, a patient might feel fearful at the time of choosing a treatment and therefore decline a riskier option, even if it has a better possible outcome. Immediate emotions can also *indirectly* influence a decision by altering expectations of the probability or desirability of an anticipated future outcome. In this case, a patient who is feeling happy may optimistically expect a good outcome from risky therapy and "go for it," even it is it riskier. The second pathway of influence, expected emotions, are cognitive appraisals about the emotional consequences of decisions rather than the emotions currently being experienced. These are possible emotions that one considers when making a current decision. An example of an expected emotion's impact on decision making is a patient with prostate cancer projecting how he might feel if he developed impotence as a result of surgery, then choosing watchful waiting to avoid the undesired emotional consequences of that surgical outcome.

This general framework can be understood in specific circumstances based on the particular emotions involved and the context of the decision. Affect-program emotions—those that are most immediate, universal, and disruptive to current actions—can strongly influence immediate emotions. Consider, for example, a physician heading to the office after an unresolved spousal argument. Anger is a negative emotion, an activating one, and one that leaves one feeling less in control. The source of the anger is not relevant to the medical decisions that will be made that day; yet it is probable that those decisions will be more negative, aggressive, and definite, independent of the relevance of these features to the calculation of what is best for the patients. The important feature about affect-program emotions is that they can

have big impacts on decisions with relatively little input from cognition.

When considering the other type of more complex emotions, the framework for application becomes more complicated. Other higher-level emotions seem to have longer-standing cognitive underpinnings that accompany them, making it more difficult to see their specific causal role. For example, emotions such as love, envy, vengeance, and empathy, to name a few, are quite different in character from those of the affect programs. Because they are accompanied by underlying, preceding thought processes that influence the emotions, it is more difficult to assign specific influences regarding decisions to these more complex emotional states. In the medical context, empathic physicians are thought to provide better care for their patients, all other considerations being equal, through more thoughtful decision making. However, characterizing the influence of empathy on decision making is very difficult.

Another reason emotions can be difficult to study with regard to choices is that many of the precursors of emotional responses are unconscious. Therefore, people are unaware that their decision processes are affected by these emotions, making them difficult to assess accurately. This is particularly true of the affect-program types of emotions. Evolutionarily, they are believed to protect the organism by causing certain actions to avoid specific situations. These more basic, universal emotions use neural pathways such as the amygdala that bypass other, more cognitive pathways such as the frontal cortex. Using animal models and functional magnetic resonance imaging (fMRI), neuroscientists have done an impressive job outlining the relevant neural pathways and showing how they bypass higher centers. For example, if one "sees" a snake in one's path and immediately reacts to get away, that might be important for survival from an evolutionary perspective. However, if that snake turns out to be a harmless stick, one has responded to a false judgment. These responses to fear, anger, and happiness still exist, but they can lead to false judgments and decisions in the modern world. A patient's fear about a disease such as cancer or Alzheimer's disease may derail his or her ability to consider rationally the probabilities involving a treatment decision.

In a similar vein, it has also been shown that damage to specific brain areas that disconnects our emotional responses from cognitive assessment can profoundly affect decision making. Damage to prefrontal cortex areas seems to disconnect our emotional centers from our more cognitive ones, leading those with such damage to become excessively risk taking and unable to conduct straightforward cost-benefit calculations. The mechanism seems to be a loss of the normal emotional response to losses, which makes undamaged individuals loss-averse. Patients with dementia are also prone to such behavior, and they may be unable to make decisions regarding their own care.

Typical Circumstances

The role of emotions in decision making is best understood in decisions that involve risky and/or uncertain choices over time. In the economic, utility-based conception of choice, risk and uncertainty are modeled with the assumption of expected utility represented by a "risk preference." Choices over time are modeled by discounted utility models in which future values are assumed to be worth less than the current values. However, a number of examples have been found showing that both of these models have important, persistent exceptions.

Under expected utility models of risky choice, risk is conceived as an outcome's expected value, the product of its likelihood of occurring and the subjective value of that outcome. People are said to have risk preferences if, when the expected values of competing outcomes are equal, they prefer one based on the distribution of it occurring. This explanation has been invoked to explain people's general willingness to purchase health insurance because they are risk-averse. However, there are a number of empirical situations in which people prefer riskier options in some situations and safer options in other situations, something that is inconsistent with expected utility. This is the result of the hedonics of valuing—people generally dislike a loss much more than they like the same-sized gain. As a result, people tend to avoid risks in a gain situation but to accept the same risks in a loss frame, an effect called loss aversion. One of the explanations of such findings is that emotions regarding risk change valuation.

Considering time-based decision making, the standard economic model is the discounted utility model. This model assumes that future values are

worth less than current values at a constantly decreasing discount rate. Once again, a number of exceptions to this model have been found. For example, when asked to give the preferred time for a kiss from a chosen movie star, people choose 3 days from now rather than immediately, to "savor" the anticipation of the event. Once again, emotions are likely explanations for the failure of the economic models and the need for alternative explanations.

Emotions are thought to have important effects on decisions. However, characterizing the exact role emotions play in choices is very difficult. The best characterized emotions are the affect-program emotions, such as anger, fear, disgust, and happiness. These basic, evolutionarily preserved, and universal emotions appear to bypass the usual neural pathways and influence choices by disrupting other cognitive inputs. This is most important when the emotions are immediate but unrelated to the decisions being made, thereby deviating most strongly from balanced cost-benefit assessments. Other, more complex emotions have more cognitive underpinnings, and their effects on behavior are more indirect. Choices that involve risk, uncertainty, and "distance" are most likely to be influenced by emotions. There remains much work to be done to characterize how these emotions affect medical decisions.

William Dale

See also Decision Making and Affect; Decision Psychology; Fear; Mood Effects; Risk Perception

Further Readings

Elster, J. (1999). *Strong feelings: Emotion, addiction, and human behavior*. Cambridge: MIT Press.

Griffiths, P. E. (1997). *What emotions really are*. Chicago: University of Chicago Press.

Hastie, R., & Dawes, R. M. (2001). *Rational choice in an uncertain world: The psychology of judgment and decision making*. Thousand Oaks, CA: Sage.

LeDoux, J. (1996). *The emotional brain: The mysterious underpinnings of emotional life*. New York: Touchstone.

Loewenstein, G., & Lerner, J. (2003). The role of affect in decision making. In R. J. Davidson, K. R. Scherer, & H. H. Goldsmith (Eds.), *Handbook of affective sciences*. New York: Oxford University Press.

EQUITY

Equity in medical decision making is an area that has received little attention. One strategy to reduce disparities in care that often arise during the medical encounter, and thus increase equity, is shared decision making between providers and patients. The shared decision-making model includes a number of critical factors that can improve care: better communication; patient-centered, culturally competent care; and patient involvement in deliberations and decisions. Each of these elements can mitigate the sociopolitical factors that have been institutionalized in medicine through the unbalanced relationship between physician and patient. This model appears to be a powerful tool that could reduce disparate care and improve overall health outcomes for minority patients.

Background on Disparities in Healthcare

Disparities in healthcare in the United States are widespread and well documented. Multiple studies show that minorities are less likely to receive important healthcare services, including preventive services and regular physicals, as well as clinically appropriate interventions. They are also more likely to receive care from providers with fewer resources, lower qualifications, and less experience than whites. This disparate care results in less satisfaction with care, lower compliance with prescribed treatments, and poorer health outcomes for many minority Americans.

The poor quality of care provided to minority groups can be explained in part by failures in the healthcare system. Insurance status is a powerful predictor of healthcare use and type of provider seen, and minority groups are more likely to be uninsured than whites. Access problems, including geographic proximity to care and linguistic and cultural barriers, also hinder minority patients' ability to seek out high-quality care.

Disparities are not, however, solely a consequence of these system-level factors. Differences in care persist even when controlling for insurance status and access issues. Some researchers suggest that lower-quality care for minorities may be explained in part by patient preference, but the

evidence is inconsistent, and the effect has been found to be small.

Given that patient preference cannot adequately explain disparities in care, researchers have begun to examine whether disparities emerge from the medical encounter and the process that physicians and patients go through to make important decisions about patients' health and healthcare. Provider bias in decision making, for example, can lead to disparate care for minorities. While providers resist believing that they provide disparate care, studies suggest that intentional and unintentional stereotyping and bias by race, ethnicity, and gender influence clinical decisions and lead to inferior care for minorities.

Poor communication, lack of information, and mistrust between patient and provider can influence patients' understanding of their health and the decisions they make regarding their care. Care for minority patients is often less patient centered than care for white patients, particularly when the patient-physician relationship is not racially or ethnically concordant. Minorities are less likely than whites to report that their physicians engage in participatory care and patient-centered communication and more likely to report that their physicians treat them with disrespect. Misunderstandings and a lack of culturally competent care on the part of the provider also contribute to disparate care.

A physician-patient interaction in which there is poor communication, bias, and mistrust is likely to result in uninformed decision making. Understanding and improving the decision-making process may mitigate some of these effects and substantially improve care for minority patients.

The Decision-Making Process and Disparities in Care

In their seminal research on shared decision making in the medical encounter, Cathy Charles et al. identify different theoretical approaches to medical decision making and espouse the benefits of shared decision making over the more traditional paternalistic model. In the paternalistic model, patients defer all decisions to the physician, who has the professional authority, training, and experience to make the "right" decisions for the patient. In each of the three stages of decision making—exchanging information, deliberating options, and deciding on a treatment—the physician controls the process, with little to no input from the patient.

In this model, information exchange is restricted, flowing largely in one direction from the provider to the patient. During the deliberation stage, the physician alone or in consultation with other physicians considers the risks and benefits of alternative treatment options. Finally, the decision of the most appropriate treatment is made solely by the physician.

Within the context of social, economic, and political inequities experienced by minorities, the power asymmetry of the medical encounter is fraught with tension; the paternalistic model of decision making perpetuates this imbalance when the relationship is not racially or ethnically concordant. The one-way direction of information exchange (Phase 1) from physician to patient controls not only the amount but also the content of information shared with the patient. Minority patients might experience this information imbalance as a form of coercive authority and a way of dismissing patients' desire to be involved in their own care decisions. In fact, studies suggest that physicians' communication style is more verbally dominant with black patients than with white patients and that black patients have significantly less participatory visits with their physicians than white patients, particularly when the physician is white.

Without information provided by the patient, the physician may be unaware of important clinical concerns that could inform his or her treatment decisions and recommendations for certain interventions. For example, minority patients may experience illness differently than white patients; they may also have different expectations of the role of healthcare in their lives. In a paternalistic model, however, these issues are unlikely to emerge or be considered during the care process.

The deliberation and decision phases of the paternalistic model (Phases 2 and 3) also exclude the patient. For minorities, this is an especially important issue when their physician relationships are not concordant, and the physician is unlikely to be familiar with or knowledgeable about the unique racial and cultural contexts of their minority patients. Physicians who infer patients' needs and preferences rarely get them right, which can result in misunderstandings, confusion regarding care, lower patient compliance with treatments,

distrust of the system, and overall dissatisfaction with care.

In the paternalistic model, the physician will provide the patient with a recommendation based on what he or she believes to be the "best" course of treatment despite imperfect information. Physicians' personal biases and stereotypes are reinforced as assumptions regarding what the patient wants and needs go largely unchallenged. Preconceived notions can very likely influence physicians' treatment recommendations and convey messages regarding minority patients' competence, self-efficacy, and deservingness.

The paternalistic model of decision making may still be the prevalent mode in which most physicians practice. Given its potential for inequitable practices and outcomes, however, a more patient-centered model of decision making should be considered. A model of shared decision making could be used to mitigate many of the negative factors inherent in the paternalistic model.

In the shared decision-making model, the patient (sometimes including the patient's family) and the physician work together through all three stages of the decision-making process. At the core of this model is the concept of patient-centered care. Through communication, information exchange, and partnership in the deliberation and final decision processes, the physician and the patient together identify the patient's needs and preferences and incorporate them into their decision.

In a patient-physician relationship that is not racially or ethnically concordant, this model of decision making is critical to developing mutual trust and understanding of how the patient's social and cultural context influences his or her presentation of illness and compliance with care. Providers must approach minority patients in a manner that is appropriate and respectful of their cultural mores. This requires that the physician and patient participate in open communication, exchange information, and develop a relationship where both parties are partners in the decision-making process.

In this model, providers must also try to recognize their own limitations. For example, physicians must use interpreters and seek help in understanding racial and ethnic groups' styles of communication. They must learn about their patients' backgrounds and value systems to understand better the most appropriate course of action for their minority patients—one that will be accepted and followed. Finally, physicians must be frank with themselves about their assumptions and beliefs regarding racial and ethnic groups and understand that they may be intentionally or unintentionally reinforcing disparate behavior based on stereotypes.

The exchange of information is critical to this process (Phase 1). In the shared decision-making model, the responsibility of exchanging information falls on both participants in the medical encounter. The provider is expected not only to provide information but also to elicit information from the patient regarding his or her needs, preferences, and values. The patient is also expected to share his or her experiences and expectations. If both participants are clear about their expectations and share their knowledge and values, then the decision-making process can be used to eliminate many of the inequities that may emerge from the patient-physician encounter.

As part of this process, the physician and patient should clearly establish the preferences of the patient regarding the roles each will play in decision making. It may be that not all patients want to take a participatory role in their own care process. For example, recent studies have suggested that black patients want information and full disclosure regarding medical tests and procedures but are hesitant to have autonomous decision-making power and prefer to follow the recommendations of their providers. For some ethnic groups, decision making is a family-centered process, including multiple family members. Understanding that some patients may prefer to delegate final responsibility of the treatment decision to others, including the physician, is part of the shared decision-making process.

Information exchange and patient involvement in this model of medical decision making may be able to reduce provider assumptions, improve communication, and achieve congruence in perspectives of health and approaches to treatment. It also sets the stage for the deliberation and final treatment decision phases of the care process. When engaged in shared decision making, the physician helps the patient weigh different treatment options with a better understanding of that patient's unique cultural context. When a decision regarding the best course of action is agreed on, the physician

can probe the patient to ensure that he or she fully understands the implications of their (the physician and the patient's) choice. Understanding the patient's perspective is critical for the provider to fully comprehend the patient's experience of illness, how he or she perceives risks and benefits of treatment, and how he or she might accept and comply with medical intervention.

Katherine Mead and Bruce Siegel

See also Cultural Issues; Discrimination; Shared Decision Making

Further Readings

Aberegg, K., & Terry, P. B. (2004). Medical decision-making and healthcare disparities. *Journal of Laboratory and Clinical Medicine, 144*(1), 11–17.

Brian, D., Smedley, A. Y., & Nelson, A. R. (Eds.). (2002). *Unequal treatment: Confronting racial and ethnic disparities in health care*. Washington, DC: Institute of Medicine.

Burgess, D. J., Van Ryn, M., & Fu, S. S. (2004). Making sense of the provider role in promoting disparities. *Journal of General Internal Medicine, 19,* 1154–1159.

Charles, C., Whelan, T., & Gafni, A. (1999). What do we mean by partnership in making decisions about treatment? *British Medical Journal, 319,* 780–782.

Cooper-Patrick, L., Gallo, J. J., Gonzales, J. J., Vu, H. T., Powe, N. R., Nelson, C., et al. (1999). Race, gender, and partnership in the patient-physician relationship. *Journal of the American Medical Association, 282,* 583–589.

Johnson, R. L., Roter, D., Powe, N. R., & Cooper, L. A. (2004). Patient race/ethnicity and quality of patient-physician communication during medical visits. *American Journal of Public Health, 94*(12), 2084–2090.

Kaplan, S. H., Gandek, B., Greenfield, S., Rogers, W., & Ware, J. E. (1995). Patient and visit characteristics related to physicians' participatory decision-making style. *Medical Care, 33*(12), 1176–1187.

Murray, E., Pollack, L., White, M., & Lo, B. (2007). Clinical decision-making: Patients' preferences and experiences. *Patient Education and Counseling, 65*(2), 189–196.

Stewart, M., Brown, J. B., Weston, W. W., McWhinney, I. R., McWilliam, C. L., & Freeman, T. R. (1995). *Patient-centered medicine: Transforming the clinical method*. Thousand Oaks, CA: Sage.

Suurmond, J., & Seeleman, C. (2006). Shared decision-making in an intercultural context: Barriers in the interaction between physicians and immigrant patients. *Patient Education and Counseling, 60*(2), 253–259.

Torke, A. M., Corbie-Smith, G. M., & Branch, W. T. (2004). African American patients' perspectives on medical decision-making. *Archives of Internal Medicine, 164*(5), 525–530.

Equivalence Testing

Frequently, the objective of an investigation is not to determine if a drug or treatment is superior to another but just equivalent. For instance, it is often of interest to investigate if a new drug, with say fewer side effects or lower price, is as efficacious as the one currently used. This situation occurs when new or generic drugs are evaluated for approval by the Food and Drug Administration (FDA).

In standard hypotheses testing, equivalence (i.e., equality) is the null hypothesis, and the alternative is the nonequivalence hypothesis. One problem with using this procedure, and determining equivalence when the null is not rejected, is that the test is designed to reject the null hypothesis only if the evidence against it is strong (e.g., $p < .05$). In other words, the burden of proof is in nonequivalence. The correct procedure to establish equivalence reverses the roles of null and alternative hypotheses so that the burden of proof lies in the hypothesis of equivalence. Consequently, the Type I error is tantamount to favoring equivalency when the drugs are not equivalent. This is the error that the FDA wants to minimize, and its probability is controlled at a low level (e.g., .05 or lower).

Some issues arise when testing for equivalence. A critical one is that perfect equivalence is impossible to establish. This problem is solved by introducing limits of equivalence that establish a range within which equivalence is accepted. Frequently, these limits are symmetric around a reference value. An example should help clarify the situation.

Suppose that a new drug for eliminating (or reducing to a prespecified level) a toxin in the blood is being evaluated. It has fewer side effects and the manufacturer is interested in proving that

it is as efficacious as the currently used drug. Let p_C and p_N be the true (population) proportion of patients who respond to the current and the new drug, respectively. The problem consists of testing

H_0: $|P_C - P_N| \geq \delta$ (nonequivalency)
H_1: $|P_C - P_N| < \delta$ (equivalency).

A more informative way to write the alternative is H_1: $p_C - \delta < p_N < p_C + \delta$, which states that the efficacy of the new drug is within δ units from that of the current drug. The role of δ is crucial, and its value should be chosen with great care. Clearly, the probability of favoring equivalency increases as δ increases, so its value should be based on acceptable levels of deviation from perfect equivalence. The value of δ should be determined based on sound medical and biological considerations, independently of statistical issues. For example, if the potential benefits (fewer side effects) of the new drug are high, a larger value of δ could be justified. When the effect of the current drug is well established, the value of p_C is fixed and the test becomes a one-sample equivalence test.

Using data from the National Immunization Survey (NIS), in 2002, Lawrence Barker and colleagues investigated whether vaccination coverage was equivalent between children of three minority groups and white children. Since the NIS data for 2000 were supposed to detect coverages at the 5 percentage point level, δ was chosen to be 5. Thus, the alternative hypothesis was H_1: $-5 < p_M - p_W < 5$, where p_W and p_M are the coverage for white and minority children, respectively. The equivalence of the coverage was to be established if the data provided enough evidence to support H_1.

Procedure

An intuitive method to test equivalency is known as the two-one-sided test (TOST) procedure. At an α level, the TOST procedure will accept the hypothesis of equivalence if a $(1 - 2\alpha) \times 100\%$ confidence interval (CI) for the difference in proportions is contained in the interval $(-\delta, \delta)$. If either limit is outside the interval, nonequivalency cannot be rejected (i.e., equivalency cannot be established). The TOST procedure can be used in situations that involve other parameters (i.e., means, medians, odds ratios, etc.). It is important to note that, even though the TOST procedure is two sided, it achieves an $\alpha = .05$, using a 90% CI.

Barker and colleagues found that the vaccination coverage for the 3-DTP vaccine was 95.0% for whites and 92.1% for blacks, with a 90% CI for the difference of (1.5, 4.3). Since this interval is included in the interval (–5, 5), equivalence was established at a .05 level. It is important to note that this interval does not include 0, so the standard procedure would have found a significant difference between the coverages (i.e., a lower coverage for black children). A contradiction also occurs when the CI includes 0, but is not within $(-\delta, \delta)$. In fact, Barker and colleagues found that contradictions occurred in 9 out of 21 comparisons (three minority groups and seven vaccines). In 7 of the 9 cases, the TOST procedure favored equivalency in contradiction with the standard procedure; in 2 cases the results were reversed.

In some cases, symmetric limits of equivalency are not appropriate. That would be the case when the "costs" of erring in either direction are not the same. In such a case, the procedure would be based on whether the CI is contained in an interval (δ_1, δ_2).

Sample-Size Considerations

The main problem in equivalence testing is that the samples needed to achieve acceptable levels of power are, frequently, fairly large. Using the TOST procedure with $\alpha = .05$, samples of $n = 2{,}122$ per group are needed to achieve a power of .95 to establish equivalence when $p_N = .4$, $p_C = .3$, and $\delta = .15$. Under the same circumstances, a standard procedure requires $n = 589$ per group to reject the null hypothesis of equivalence and conclude (incorrectly) nonequivalence. As mentioned earlier, larger values of δ increase the power to detect equivalence and thus reduce the required sample size. For example, if $\delta = .2$, the sample size needed to establish equivalence in the previous situation is $n = 531$.

Testing equivalence is particularly applicable in public health, where the sample sizes are usually large. The NIS contains millions of records of children nationwide, yielding a high power for any test. However, in clinical studies, large samples are hard

to obtain, thus limiting the application of equivalence testing. In this respect, Stefan Wellek states,

> In equivalence testing power values exceeding 50% can only be obtained if either the equivalence range specified [δ] by the alternative hypothesis is chosen extremely wide or the sample size requirements are beyond the scope of feasibility for most if not all applications. (p. 63)

Noninferiority Testing

Noninferiority, or one-sided equivalence, testing is appropriate when the objective is to establish that one arm is not inferior to another (and possibly superior). Actually, the example of the toxin-reducing drug might be better suited for a noninferiority test. That is, the objective is to establish that the new drug is not inferior in efficacy to the drug in current use. As before, let p_C and p_N be the true efficacy of the current and the new drug, respectively. The test of interest is

H_0: $P_C - P_N \geq \delta$ (inferiority)
H_1: $P_C - P_N < \delta$ (noninferiority).

It is informative to write the noninferiority hypothesis as H_1: $p_N > p_C - \delta$, which states that the efficacy of the new drug is not more than δ units lower than the efficacy of the current drug. In cases where higher values imply inferiority, the alternative hypothesis becomes H_1: $p_N - p_C < \delta$.

Noninferiority testing differs from a standard one-sided test only by the use of the "offset" term δ. Thus, noninferiority is established at an α level if the upper limit of a $(1 - 2\alpha) \times 100\%$ CI for the difference $p_C - p_N$ is less than δ. Significant confusion in the medical literature is caused by the fact that the upper limit of a $(1 - 2\alpha) \times 100\%$ CI is also the upper limit of a $(1 - \alpha) \times 100\%$ one-sided CI. That is, the upper limit of a 90% CI is also the upper limit of a 95% one-sided CI.

Warfarin prevents ischemic stroke in patients with nonvalvular atrial fibrillation, but dose adjustment, coagulation monitoring, and bleeding limit its use. In 2005, SPORTIF (Stroke Prevention Using an Oral Thrombin Inhibitor in Atrial Fibrillation) was created to conduct a study to compare ximelagatran with warfarin for stroke prevention. Ximelagatran has a fixed oral dosing, does not require coagulation monitoring, and has few drug interactions. The objective was to establish noninferiority of ximelagatran with respect to stroke prevention. An absolute margin of $\delta = 2\%$ per year was specified. Therefore, if p_X and p_W are the yearly stroke rates for ximelagatran and warfarin, respectively, the noninferiority hypothesis was H_1: $p_X < p_W + 2$. The observed yearly event rates were 1.62% and 1.17% for ximelagatran and warfarin, respectively. The difference was .45%, and the 95% upper limit of the CI for the difference was 1.03%. Since it was less than 2%, noninferiority was established. Note that for $\delta = 1\%$, the data do not support noninferiority.

Other Situations

Equivalence testing, just as standard testing, can be applied to a variety of problems. This includes differences or ratios of measures of location (e.g., means, proportions, medians) and dispersion (e.g., standard deviations). Wellek describes parametric and nonparametric tests of equivalence for dependent observations, multiple samples, linear models, survival times, hazard rates, and bioequivalence.

A natural application of the equivalence concept is in lack of fit where the objective is to determine if an observed distribution is equivalent to another. In the standard chi-square test for lack of fit, the null hypothesis is of equivalence, and thus, it is not designed to establish equivalence to the reference distribution and tends to favor equivalence too frequently. Wellek presents an equivalence test for this situation that has the desired properties.

p Values

Reporting *p* values in testing equivalence is not done routinely. This is unfortunate because *p* values are not difficult to calculate. In the noninferiority case, the *p* value is obtained from a standard test with an offset value of δ. This procedure can be carried out with any standard statistical program. To calculate the *p* value in the case of equivalence, one uses the fact that establishing equivalence is tantamount to establishing nonsuperiority and noninferiority, simultaneously. Thus, the *p* value for equivalence is the larger of the two *p* values.

Final Thoughts

Testing equivalence or noninferiority is the appropriate procedure for many biological and medical situations in which the objective is to compare a new therapy with a standard. The procedures to perform these tests are simple modifications of those used in standard testing but in many cases result in completely different conclusions. The margin of equivalence is critical and sometimes is the most critical issue. In spite of their apparent simplicity, there is still considerable confusion in the medical literature on how to perform and interpret equivalence and noninferiority tests. In a recent study, Le Henanff and colleagues found that out of 162 published reports of equivalence and noninferiority trials, about 80% did not justify the choice of the equivalence margin. They also observed that only about 50% of the articles reported a p value, and only 25% interpreted it.

The main obstacle in the application of these methods is the large samples needed to achieve acceptable levels of power. In the study by Le Henanff and colleagues, the median number of patients per trial was 333. However, 28% of the studies reviewed did not take into account the equivalence margin, so it is likely that many were underpowered to detect equivalency. Finally, the decision between testing equivalence or noninferiority involves similar issues as in choosing between a two-sided or a one-sided alternative in standard testing. This decision should be based on the objectives of the study and not on the observed data.

Esteban Walker

See also Efficacy Versus Effectiveness; Hypothesis Testing

Further Readings

Barker, L. E., Luman, E. T., McCauley, M. M., & Chu, S. Y. (2002). Assessing equivalence: An alternative to the use of difference tests for measuring disparities in vaccination coverage. *American Journal of Epidemiology, 156,* 1056–1061.

Barker, L., Rolka, H., Rolka, D., & Brown, C. (2001). Equivalence testing for binomial random variables: Which test to use? *The American Statistician, 55,* 279–287.

Le Henanff, A., Giraudeau, B., Baron, G., & Ravaud, P. (2006). Quality of reporting of noninferiority and equivalence randomized trials. *Journal of the American Medical Association, 295,* 1147–1151.

SPORTIF Executive Steering Committee. (2005). Ximelagatran vs. Warfarin for stroke prevention in patients with nonvalvular atrial fibrillation. *Journal of the American Medical Association, 293,* 690–698.

Wellek, S. (2003). *Testing statistical hypotheses of equivalence.* Boca Raton, FL: Chapman & Hall.

Error and Human Factors Analyses

The study of human error in diverse sociotechnical systems may be conducted by examining the human factors that contribute to error. A human error may be broadly defined as failure to take required action, failure to meet a performance standard for that action, or performing the wrong action. In the medical domain, human error may or may not adversely affect the patient. Patient safety is a medical providers' principal concern; thus, much attention has been placed on uncovering errors that have the potential to cause patient injury. Any patient injury (i.e., adverse medical event) that is attributable to human error is described as a preventable adverse event. Two prolific studies conducted in New York, Colorado, and Utah suggested that between 2.9% and 3.7% of hospitalizations produce adverse events. The proportion of these adverse events that was attributable to error (i.e., preventable adverse event) was between 53% and 58%. The New York study estimated that preventable adverse events in hospitals causes approximately 44,000 to 98,000 deaths annually when extrapolated to the 33.6 billion hospital admissions the United States experienced in 1997. Even if these statistics underestimate the magnitude of the problem, as some have argued, they would still place preventable adverse events among the leading causes of death in the United States, ranking higher than motor vehicle accidents, breast cancer, and AIDS. Results of these studies prompted the Institute of Medicine to produce the report *To Err Is Human: Building a Safer Health System,* which strongly recommended that the healthcare community look to other high-risk industries, such as nuclear power

and aviation, for ways to improve their own record on the quality and safety of healthcare delivery. The authors of that report endorsed the study and application of human factors analyses to measure and improve human-system performance.

Human factors analyses are used to study how humans interact psychologically and physically with their particular environment (i.e., system). This includes the study of both human-human and human-system interactions. The objective of human factors analyses is to understand the nature of these interactions in order to improve system performance and human well-being. The field of human factors originated in aviation, but its current scope is as broad as it is deep, with specialization ranging from safety and human error to aging and virtual reality. The results of many years of research from the human factors community have yielded valuable insights into crucial aspects of human performance that can be affected by the design of work in light of human capabilities and limitations and the specialized nature of work involved in different aspects of healthcare.

Human factors analyses can be used to create assessments of the requirements of work involving demands on human performance in the physical, cognitive, and social domains. This includes the ways in which task demands, workload, and situational awareness are likely to interact to influence the performance of individuals. This knowledge can be translated to work process and system design such that human performance is optimized and system safeguards prevent human error from translating to patient injury. Formal human factors methods and measurement techniques are available, and many demonstrations of their uses exist to guide their application to healthcare. Each provides potentially useful information about factors that can limit or improve human performance. Measuring and accounting for human factors associated with healthcare delivery is likely to lead to safer and more reliable patient care.

The human factors analyses methods described in this entry are a subset of the methods most relevant to the study of error in medicine. This entry focuses on error in medicine and how human factors analyses may be applied to study and improve the quality and safety of healthcare delivery.

Human Error

Human error is often classified into one of two groups. The first group describes errors of *omission*—failing to perform a task or failing to act within a time period required by the situation. In this case, something that should have been done was not done because it was either skipped or not performed in time. The second group encompasses errors of *commission*—performing the wrong action or performing the right action incorrectly. Regardless of whether an action was performed or not, if it fails to meet an established performance standard, then it may be said to result in a state of error. Both errors of omission and commission describe a failure to achieve prescribed results in such a way that action was not performed or not performed to an acceptable standard. This includes both actions that are intentionally undertaken and those that are unintentionally committed.

Exposure to Human Error in Medicine

Medical systems function either directly or indirectly under the control of humans. This ranges from frontline medical care to the management and administrative work needed to support frontline care. For the patient, this encompasses a range of activities that begins at admission, continues through diagnosis and treatment, and ends with discharge or death. Patients may be exposed to errors through a variety of clinical and administrative activities over the course of their care. Lucian Leape and others categorized the most prevalent types of errors that patients are exposed to in an article titled "Preventing Medical Injury." *Diagnostic errors* were described as delays in diagnosis, failure to order appropriate tests, and failure to act on monitoring or test results. *Treatment errors* included, but are not limited to, errors in performance of a procedure, errors in administering a treatment, and errors in drug dose or administration. *Preventive errors* included failure to provide prophylactic treatment and inadequate monitoring or follow-up treatment. *Other errors* included communication failures, equipment failures, and other system failures. As the healthcare system has evolved to be more disaggregated and complex, it is helpful to study the causes of human error and preventable adverse events within the context of complex systems.

Human-System Error in Complex Systems

Medical specialization and advancements in medical knowledge and technology have created a complex healthcare delivery system. This complexity and disaggregation have greatly increased the opportunity for adverse events attributable to human error (i.e., preventable adverse events). Charles Perrow has discussed the propensity of human error in complex systems in other industries and attributes them in part to the nature of the systems themselves. Healthcare has been categorized as a complex and tightly coupled system. System complexity arises from the system's numerous specialized interdependent components (e.g., departments, providers, equipment). From a patient's perspective, complexity is reflected in the number of processes that must be accounted for and the dynamic nature of their relationships and response to medical treatment. Coupling refers to the degree of dependency between the system components and processes. Complex, tightly coupled systems are at greater risk for adverse events due to their inability to foresee the consequences of component interactions and the unique situations that are prone to error.

James Reason illustrates the human contribution to error in complex systems as either active or latent. Active errors occur only at the point of patient-provider interaction. The consequences of these errors are usually evident and experienced immediately. Latent errors are more suppressed, unknown, and await the appropriate initiating event or signal to trigger their effects on the system. In some cases, latent errors represent known or accepted conditions that either await correction or are not appreciated for the types of effects they may eventually unleash. Examples of latent errors include poor design, incorrect installation, poor management decisions, and poor communication processes. Latent errors are most dangerous in complex systems because of their ability to cause numerous active errors.

The study of human factors that contribute to error and preventable adverse events focuses on minimizing the risk of active and latent errors. Human factors methodologies assume that humans are imperfect and that errors are to be expected. However, there are specific factors within the work environment (i.e., system) that provoke the occurrence of errors. Just as the systems should be designed to minimize error-provoking circumstances, defense barriers and safeguards should be put in place to eliminate the ability of human error to create adverse medical events. Human factors analyses are meant to bridge the gap between system influences and human error and are especially relevant to complex systems such as healthcare. Human factors analyses may be used to study and understand human error within an environmental context and facilitate improvements, leading to better system performance and reliability.

Human Factors Analyses

Human factors analyses methods are concerned with the interaction of humans with the tools, systems, and other humans that make up their work environment. The design for these tools, systems, work, and communication processes are meant to strongly consider the characteristics, capabilities, and limitations of humans. Human factors analyses grew from work in aviation during World War II to improve cockpit design and aircrew performance. Cognitive psychology, engineering, computer science, sociology, anthropology, and artificial intelligence represent the roots of human factors methods. To date, human factors methods have been used extensively by other high-risk industries to improve human performance and organizational reliability. The application of human factors analyses in medicine has been less prevalent but is not necessarily new. Human factor analyses were first documented in medicine in the 1970s, and their application has steadily increased since then. Human factors analyses have been applied to a number of healthcare topics, including the following:

- Error reduction
- Hospital design
- Anesthesia delivery
- Patient safety
- Time and motion studies
- Workload management
- Communications and distractions
- Pharmacy operations and accuracy
- Team performance in operating rooms
- Curriculum development for medical schools
- The impact of information technologies on healthcare management and delivery

As is evident from this list, the field of human factors has always emphasized research and its application in work settings. As a result, a number of methods and tools are now available to analyze human performance. Although unique in purpose, these methods produce data that can be used within a workplace or system setting to assess requirements for safe, efficient, and economical operation in light of human capabilities and limitations. Three human factors methods that are among the most highly relevant to individuals and organizations involved in healthcare (i.e., task analyses, workload analyses, and situational awareness analyses) are described.

Task Analyses

The most common human factors method for studying human behavior in the work environment is task analysis. Task analysis involves the observation of people as they perform their work, structured interview techniques to elicit information from workers, and analysis of the resulting observations and data to describe an integrated set of activities that represent human performance in a work domain of interest. These methods employ a structured description or "decomposition" of work activities or decisions and classification of these activities as a series of tasks, processes, or classes. Each task is systematically described in terms of the mental and physical activities needed to perform the task successfully. The product of a task analysis is a description of tasks, the sequence and duration of their execution, conditions for task initiation and completion, the physical and mental dimensions of task performance, communications between work teams and between team members, and the tools and systems used to perform work. The results of task analysis are used to estimate the characteristics of predefined tasks, such as the frequency, complexity, time needed, equipment and tools needed, and communication performed.

Task analysis is an important analytical method for describing the way work is intended to be carried out. It works particularly well for sets of activities that occur in well-prescribed sequences. Results of task analysis are often ultimately used to develop or validate operational procedures, develop qualification requirements, develop training programs, and support the design of assistive tools employed in the workplace. The results may also be used to verify that the expectations for human activity are compatible with the capabilities and limitations of those expected to perform the work. This includes requirements for precision and accuracy, speed, strength, endurance, and other psychophysiological factors such as anthropometry (e.g., physical ability) and ergonomics.

Workload Analyses

Workload is a multidimensional, multifaceted concept that is difficult to define concisely. The elusiveness of a single satisfactory definition has challenged human factors researchers on many fronts and has fueled a lively and active debate among them. Even without consensus on a definition, human factors professionals agree that workload is a very valuable concept to understand and to measure in sociotechnical systems. Presently, the onset of technology and automation has greatly shifted the workload paradigm from the physical domain to the mental domain. Mental workload relates to the demands placed on a human's limited mental resources by a set of tasks being performed by an individual. The assumption behind this theory is that humans have a fixed amount of processing capacity. Tasks inherently demand processing resources, and the more difficult the task or tasks, the higher the processing capacity required for acceptable performance. If at any time the processing demands exceed the available processing capacity, performance quality may decrease. Thus, high levels of mental workload can lead to errors and poor system performance. Conversely, excessively low levels of mental workload can lead to complacency and errors, albeit for different reasons. As this implies, workload consists of an external, objective, observable referent as well as a subjective and mostly perceived referent. Both are important in understanding and predicting workload.

There are three primary methods for measuring workload: procedural, subjective, and physiological. Each of these methods can be applied in isolation, but generally, they are measured concurrently to obtain an integrated assessment of workload. *Procedural* measurement involves directly monitoring human behavior in the working environment. Task analysis, discussed above, is the most common method of procedural workload measurement.

Task analysis is used by observing a worker in an actual or simulated work setting and discerning changes in behavior as task loads vary. *Subjective* workload measures require a worker to rate or distinguish a level of workload required to perform a task. There are two major classes of subjective workload assessment techniques—unidimensional and multidimensional. Unidimensional techniques involve asking the subject for a scaled rating of overall workload for a given task condition. More comprehensive, multidimensional methods include various characteristics of perceived workload and are able to determine the nature of workload for a specific task or set of tasks. Validated workload measurement instruments include the NASA Task Load Index (NASA-TLX) and the Subjective Workload Assessment Technique (SWAT). *Physiological* techniques measure changes in subject physiology that correspond to different task demands. Most techniques emphasize cognitive task demands as opposed to actual physical demands. Studies have used physiological parameters such as heart rate, eye blink rate, perspiration, and brain activity to assess the state of workload of a human subject.

Situational Awareness Analyses

Mica Endsley defines situational awareness (SA) as a human's ability to perceive components within his or her environment (i.e., system), comprehend their meaning, and forecast their status in the future. This encompasses three distinct levels of SA. Level I SA refers to the perception of components within the environment. Level II SA involves comprehension of the current situation. Level III SA involves projecting the status of components and the situation picture in the near future. Progression through the levels of SA depends on the cognitive abilities and experience of an individual (and other team members) in performing mental operations on information from a dynamic process. SA is the product of cognitive activities and synthesis across the three levels of SA.

SA measurement can be used to evaluate system design and facilitate system improvements. Like workload measurement, SA measurement can be done using several methodologies. These methods include performance measures, subjective measures, questionnaires, and physiological measures. Again, these measurement techniques can be administered separately, but usually, they are used simultaneously to obtain a more global assessment of SA.

Performance measures are the most objective way to measure SA. These measures are divided into two major types—external task measures and embedded task measures. External task measures involve removing information from a subject's environment and then measuring the amount of time it takes the subject to notice this difference and react. Imbedded task measures involve studying subtasks of subjects and noting subtle deviations in expected performance versus actual performance. *Subjective* measures of SA continue to be popular because of their ease of use, low cost, and applicability to real-world environments. One of the most well-known and validated ways to subjectively measure SA is by the Situational Awareness Rating Technique (SART), developed by R. Taylor in 1990. SART is a measure based on subjects' opinions that is broken up into 14 component subscales. All these subscales are integrated to create an overall SART score for a system. SART measures have shown high correlation with SA performances measures. *Questionnaires* allow for an objective assessment of SA, eliminating the disadvantages of subjective measures. They evaluate SA on a component basis and compare a subject's assessment of a situation with actual reality. The most popular questionnaire method is Endsley's Situational Awareness Global Assessment Technique (SAGAT). SAGAT executes randomized time freezes in simulation scenarios. At the time of the freeze, questions are asked of the subject about the situation to accurately evaluate the subject's knowledge of the situation. *Physiological* measures of SA are similar to physiological workload measures, but they have been proven thus far to be much more difficult to interpret. Electroencephalograms (EEG) and eye-tracking devices have been used to measure SA.

Bruce P. Hallbert, Scott R. Levin, and Daniel J. France

See also Cognitive Psychology and Processes; Complications or Adverse Effects of Treatment; Human Cognitive Systems; Medical Errors and Errors in Healthcare Delivery; Unreliability of Memory

Further Readings

Bogner, S. (2003). *Misadventures in health care: Inside stories*. Hillsdale, NJ: Lawrence Erlbaum.

Brennan, A., Leape, L., Laird, M., Heber, L., Localio, A., Lawthers, A., et al. (1991). Incidence of adverse events and negligence in hospitalized patients: Results of the Harvard Medical Practice Study I. *New England Journal of Medicine, 324,* 370–376.

Cook, R., Woods, D., & Miller, C. (1998). *A tale of two stories: Contrasting views of patient safety*. Chicago: National Patient Safety Foundation.

Endsley, M. (1995). Toward a theory of situation awareness in dynamic systems. *Human Factors, 37*(1), 32–64.

Hart, S., & Staveland, L. (1988). Development of NASA-TLX (Task Load Index): Results of experimental and theoretical research. In P. Hancock & N. Meshkati (Eds.), *Human mental workload* (pp. 139–183). Amsterdam: North Holland.

Human Factors and Ergonomics Society: http://www.hfes.org

Institute of Medicine. (1999). *To err is human: Building a safer health system*. Washington, DC: National Academy Press.

Leape, L., Lawthers, A., Brennan, T., & Johnson, W. (1993). Preventing medical injury. *Quality Review Bulletin, 19*(5), 144–149.

Perrow, C. (1999). Organizing to reduce the vulnerabilities of complexity. *Journal of Contingencies and Crisis Management, 7*(3), 1450–1459.

Reason, J. (1990). *Human error*. Cambridge, UK: Cambridge University Press.

Salvendy, G. (1997). *Handbook of human factors and ergonomics* (2nd ed.). New York: Wiley.

Wickens, C. D. (2008). Multiple resources and mental workload. *Human Factors, 50*(3), 449–456.

Errors in Clinical Reasoning

Physicians make diagnostic and therapeutic decisions at every moment in their daily lives. Quality of care and patient outcomes, including sometimes distinction between life and death, come out of such decisions. In most cases, physicians' judgments are correct, but of course, they also fail. Errors, in fact, occur in medicine, and the Institute of Medicine's well-known report *To Err Is Human* recently called the public's and professionals' attention to this reality. Since then, the frequency and impact of adverse patient effects provoked by medical errors have been increasingly recognized. In the United States, it is estimated that medical errors result in 44,000 to 98,000 unnecessary deaths and around 1 million injuries each year. Even considering the lower estimate, deaths due to adverse events resulting from medical errors exceed the deaths attributable to motor vehicle accidents, breast cancer, or AIDS. Similar phenomena have been reported by studies in other countries. In Australia, for instance, medical errors are estimated to result in as many as 18,000 deaths, and more than 50,000 patients become disabled each year.

Medical errors occur in a variety of healthcare settings and in different stages of care. They may arise due to drug misuse or failures during the therapeutic phase, for instance, but due to their frequency and impact, diagnostic errors have received growing attention. Diagnostic error may be defined as a diagnosis that was unintentionally delayed (sufficient information for establishing the diagnosis was available earlier), incorrect (another diagnosis was made before the correct one), or missed (no diagnosis was ever made), as judged from the analysis of more definitive information. When a diagnosis is incorrect or does not entirely address the patient's problem, treatment can be delayed and/or wrong, sometimes with devastating consequences for patients and healthcare providers. Diagnostic mistakes represent a substantial and costly proportion of all medical errors. In the Harvard Medical Practice Study, the benchmark for estimating the amount of injuries occurring in hospitals, diagnostic errors represented the second largest cause of adverse events. In a recent study of autopsy, diagnostic discrepancies were found in 20% of the cases, and in half of them, knowing the correct diagnosis would have changed the case management. Indeed, postmortem studies indicate that the rates of diagnostic errors with negative impact on patient outcomes hover around 10%; this rate is stable across hospitals and countries and has not been affected by the introduction of new diagnostic technologies.

Undoubtedly, not all diagnostic errors can be attributed to faults in physicians' clinical reasoning. In a typology of medical errors that has been frequently used by Mark Graber and other authors, the so-called system-related errors come out from latent

flaws in the health system that affect physicians' performance. This type of error derives from external interference and inadequate policies that affect patient care; poor coordination between care providers; inadequate communication and supervision; and factors that deteriorate working conditions, such as sleep deprivation and excessive workload. In a second category of errors, referred to as *no-fault errors*, the correct diagnosis could hardly be expected due to, for example, a silent illness or a disease with atypical presentation. However, a third category of errors, namely, *cognitive errors*, occur when a diagnosis is missed due to incomplete knowledge, faulty data gathering or interpretation, flawed reasoning, or faulty verification. As arriving at a diagnosis depends largely on a physician's reasoning, cognitive faults play an important role, particularly in diagnostic errors. Indeed, a recent study in large academic hospitals in the United States found that cognitive factors contributed to 74% of the diagnostic errors in internal medicine.

This entry addresses this latter category of errors: diagnostic failures generated by errors in clinical reasoning. First, the mental processes underlying diagnostic decisions are briefly reviewed, and subsequently, origins of medical errors are discussed. Finally, the nature of reflective reasoning in clinical problem solving and its role in minimizing diagnostic errors are discussed.

The Nature of Clinical Reasoning

Throughout the past decades, research on clinical reasoning has generated substantial empirical evidence on how physicians make diagnoses. Two main modes of processing clinical cases—nonanalytical and analytical—have been shown to underlie diagnostic decisions. Experienced doctors diagnose common problems largely by recognizing similarities between the case at hand and examples of previously seen patients. As experience grows, this so-called pattern-recognition, nonanalytical mode of clinical reasoning tends to become largely automatic and unconscious. Complex or uncommon problems, however, may trigger an analytical mode of reasoning, in which clinicians arrive at a diagnosis by analyzing signs and symptoms, relying on biomedical knowledge when necessary.

Cognitive psychology research indicates that these two different types of reasoning result from diverse kinds of knowledge used for diagnosing cases. According to Henk Schmidt and Henny Boshuizen, medical expertise development entails a process of knowledge restructuring, and therefore, knowledge structures available to medical students and physicians change throughout training and practice. In the first years of their training, medical students develop rich networks of biomedical knowledge explaining causal mechanisms of diseases. This biomedical knowledge is gradually "encapsulated" under clinical knowledge, and with clinical experience, illness scripts (i.e., cognitive structures containing little biomedical knowledge but a wealth of clinically relevant information about a disease) and examples of patients encountered are stored in memory. Experienced physicians' diagnostic reasoning is characterized largely by nonanalytical processing that relies extensively on illness scripts, examples of patients, and encapsulated knowledge. In fact, not only have illness scripts been shown to play a crucial role in hypotheses generation, but they also organize a search for additional data and interpretation of evidence, thereby acting on hypotheses refinement and diagnosis verification. However, the diverse knowledge structures developed throughout training apparently do not decay but remain as layers in memory, and earlier acquired structures may be used to deal with problems when necessary. Physicians have been shown to make use of knowledge of pathophysiological processes, for example, to understand signs and symptoms in a patient when cases are unusual or complex and when immediate explanations do not come to mind. Indeed, expert clinicians' reasoning seems to be characterized by complexity and flexibility, and apparently, different mental strategies are adopted in response to different problems' demands.

Origins of Medical Errors

Studies of medical errors point to possible failures in the generation of hypotheses, in hypotheses refinement through data gathering and interpretation, and in diagnosis verification. These failures may come from multiple sources. First, it is to be acknowledged that uncertainty is inherent to clinical decision making. Despite the high value attributed to the rational use of objective, well-established

scientific knowledge within the medical domain and the growth of the medical knowledge base, this knowledge will always be insufficient to tell physicians what is to be done in a particular situation. Clinical judgment is a complex process that always involves perception and interpretation of findings within the context of a particular patient. The way diseases present themselves ranges from typical to very atypical manifestations, sometimes hardly recognizable. Physicians always have to interpret the scientific literature for making decisions in light of each patient's unique configuration of signs and symptoms, context, and needs. Second, traditional views of the physician as a neutral observer who objectively identifies and interprets a patient's signs and symptoms to make decisions have been increasingly questioned. Every physician always brings to a clinical encounter a body of medical knowledge that includes both theoretical knowledge from several disciplines and knowledge acquired through his or her own professional experience. From the interaction between this idiosyncratic body of knowledge and each unique patient, clinical knowledge required to solve the patient's problem is generated. Empirical studies have shown that physicians' experience, beliefs, and perspectives influence their perception and interpretation of features in a patient. Signs that corroborate a certain perspective may be recognized and emphasized, whereas another line of reasoning may not receive appropriate attention. Studies have shown that a suggested diagnosis influences identification and interpretation of clinical features in a patient. Misperceptions and misinterpretation of evidence, therefore, are not unusual in clinical problem solving and may compel physicians to make incorrect judgments.

Particular attention has been recently directed to the role of heuristics in medical errors. Heuristics are mental shortcuts or maxims that are used, largely unconsciously, by clinicians to expedite clinical decision making. Heuristics come out from professional experience or tradition, without being necessarily based on scientific evidence. They can be a very powerful instrument in the hands of experienced physicians, allowing them to take appropriate decisions, particularly within situations of time constraints. Nevertheless, heuristics can insert biases and distort reasoning throughout the diagnostic process, thereby generating cognitive errors. A set of biases have been frequently pointed as underlying diagnostic errors and exemplify the potential negative effects of the use of heuristics. *Availability bias*, for instance, occurs when the judgment of the probability of a disease is influenced by readily recalled similar events. Recent or frequent experiences with a disease may, therefore, unduly increase the likelihood that it is considered as a diagnostic hypothesis. *Confirmation bias*, another frequent distortion, compels physicians to gather and value evidence that confirms a hypothesis initially considered for the case rather than searching for and considering evidence that refutes it. Confirmation bias is frequently associated with another bias, namely, *anchoring*, which occurs when the clinician remains fixed on the initial impression of the case instead of adjusting hypotheses in light of new data. As a last example, *premature closure*, accounting for a high proportion of missed diagnoses, occurs when an initial diagnosis considered for the case is accepted before all data are considered and other alternatives are verified. These are only examples of a large set of biases, of different types, that may distort diagnostic reasoning. Some of them tend to affect the generation of hypotheses, whereas others influence processing of information or hypotheses verification.

A diversity of mechanisms may act, therefore, as underlying causes of diagnostic errors. These mechanisms may be favored by an excessive reliance on nonanalytical reasoning. Nonanalytical, pattern-recognition reasoning allows physicians to efficiently diagnose most of the routine problems but may introduce distortions in clinical reasoning, thereby leading to errors. This tends to happen particularly when physicians are faced with complex, unusual, or ambiguous problems, which would require them to adopt a more analytical reasoning mode. Studies have indicated that expert doctors may in fact shift from the usual automatic way of reasoning to an analytical, effortful diagnostic approach in some situations. This happened, for instance, when doctors diagnosed cases out of their own domain of expertise and adopted an elaborate biomedical processing approach for understanding signs and symptoms. More recent empirical studies have confirmed that doctors may engage in effortful reflection for diagnosing cases, which affects the quality of their diagnoses. These studies reflect a recent interest in the analytical

mode of diagnosing clinical cases. Research on clinical reasoning has traditionally focused on how physicians diagnose clinical problems through nonanalytical reasoning, and therefore, a substantial amount of empirical data about this mode of case processing are available. Not so much is known, however, about physicians' reasoning when they engage in reflection for solving clinical problems. Only recently, stimulated by concerns with avoidable medical errors, attention has been directed to the analytical diagnostic reasoning, and research conducted within the framework of reflective practice in medicine has contributed to shed some light on the nature and effects of reflection while solving clinical cases.

Reflective Reasoning and Diagnostic Errors

Reflective practice has been conceptualized as doctors' ability to critically reflect on their own reasoning and decisions while in professional activities. Critically reflecting on one's own practice has long been valued as a requirement for good clinical performance. Ronald Epstein suggested that by inserting "mindfulness" in their practice, physicians would become aware of their own reasoning processes during clinical problem solving. *Mindful practice*, as he called it, would compel physicians to observe themselves while observing the patient. It would then enable physicians to realize how their own body of knowledge, beliefs, values, and experiences influences their perception and interpretation of features encountered in a patient, thereby leading them to questioning and improving their own judgments.

Other authors, such as Pat Croskerry, have emphasized the potential role of metacognition, which means critically reflecting on one's own thinking processes as a crucial condition for good diagnostic performance. Metacognition consists of the ability to explore a broader range of possibilities than those initially considered for a case, the capacity to examine and critique one's own decisions, and the ability to select strategies to deal with decision-making demands.

Although they are easily encountered in the literature on medical errors, only recently did conceptualizations such as mindful practice, reflection, or reflective practice start to be investigated by empirical research. Recent studies provided empirical evidence of the nature of reflective practice in medicine. Reflective practice comprises at least five sets of behaviors, attitudes, and reasoning processes in response to complex problems encountered in professional practice: (1) an inclination to deliberately search for alternative hypotheses in addition to the ones initially generated when seeking explanations for a complex, unfamiliar problem; (2) an inclination to explore the consequences of these alternative explanations, resulting in predictions that might be tested against new data; (3) a willingness to test these predictions against new data gathered from the case and synthesize new understandings about the problem; (4) an attitude of openness toward reflection that leads reflective doctors to engage in thoughtful, effortful reasoning in response to a challenging problem; and (5) a willingness and ability to reflect about one's own thinking processes and to critically examine conclusions and assumptions about a particular problem, that is, metareasoning.

A physician who is open to reflection tends to recognize difficulties in solving a problem and to accept uncertainty while further exploring the problem instead of searching for a quick solution. By engaging in reflective practice, physicians would bring to consciousness and critically examine their own reasoning processes. Patients' problems would, therefore, be explored more thoroughly; alternative hypotheses would be more easily considered and more extensively verified. Clinical judgments would improve, and errors would be reduced. Although theoretically justified, these statements have only recently been supported by empirical studies. Experimental studies with internal medicine residents have explored the effects of the two main modes of reasoning—nonanalytical and reflective—on the quality of diagnoses. Residents were asked to diagnose simple and complex cases by following, in each experimental condition, instructions that led to either a nonanalytical or a reflective approach. Reflective reasoning was shown to improve the accuracy of diagnoses in complex clinical cases, whereas it made no difference in diagnoses of simple, routine cases. In a subsequent study with internal medical residents, this positive effect of reflective reasoning on the diagnosis of difficult, ambiguous clinical cases was reaffirmed.

These recent studies indicate that diagnostic decisions would improve by adjusting reasoning

approaches to situational demands. While nonanalytical reasoning seems to be highly effective for solving routine cases, complex, unusual, or unique clinical problems would require physicians to shift to a more analytical, reflective reasoning. This statement, however, is not so simple and obvious as it seems at first sight. As nonanalytical reasoning is inherently associated with expertise development, how would experienced physicians, who tend to reason highly automatically, recognize when a problem requires further reflection? It has been demonstrated that physicians in fact shift to analytical reasoning approaches, but conditions that break down automaticity are still under investigation. An experimental study with medical residents indicated that, as could be expected, the complexity of the case to be diagnosed seems to be one of these conditions. However, not only may the characteristics of the case itself trigger reflection, but apparently, contextual information may also play a role. In another study with residents, only information that other physicians had previously incorrectly diagnosed the case led participants to adopt a reflective approach. It is likely that factors related to the environment where the case is solved or to physicians' characteristics restrict or favor reflection. As an example, a study exploring correlates of reflective practice suggested that physicians with more years of practice and those working in primary-care settings in which high standards of performance are not so much valued tend to engage less frequently in reflection for diagnosing patients' problems.

These first studies shed some light on the conditions that trigger reflective reasoning and its effect on the quality of diagnoses, but much more remains to be explored. What seems clear now is that minimization of avoidable diagnostic errors depends on physicians' ability to adjust reasoning strategies to the problem at hand and appropriately, flexibly combine nonanalytical and reflective reasoning. While the usual pattern-recognition, nonanalytical approach allows physicians to efficiently solve familiar problems, diagnoses of complex or unusual problems would benefit from reflection. Much more, however, needs to be known about the knowledge structures and mental processes that constitute reflective reasoning, the conditions that lead physicians to effortful reflection while diagnosing cases, and the relative effectiveness of the different reasoning modes in various situations. By further investigating these issues, it would be possible to open perspectives for designing and testing educational interventions aimed at refining medical students' and practicing physicians' clinical reasoning.

Sílvia Mamede, Henk G. Schmidt, and Remy Rikers

See also Automatic Thinking; Bias; Cognitive Psychology and Processes; Heuristics; Medical Errors and Errors in Healthcare Delivery

Further Readings

Corrigan, J., Kohn, L. T., & Donaldson, M. S. (Eds.). (2000). *To err is human: Building a safer health system*. Washington, DC: Institute of Medicine/ National Academy Press.

Croskerry, P. (2003). The importance of cognitive errors in diagnosis and strategies to minimize them. *Academic Medicine, 78,* 775–780.

Epstein, R. M. (1999). Mindful practice. *Journal of the American Medical Association, 282,* 833–839.

Graber, M. L., Franklin, N., & Gordon, R. (2005). Diagnostic error in internal medicine. *Archives of Internal Medicine, 165,* 1493–1499.

Kassirer, J. P., & Kopelman, R. I. (1991). *Learning clinical reasoning*. Baltimore: Williams & Wilkins.

Kempainen, R. R., Migeon, M. B., & Wolf, F. M. (2003). Understanding our mistakes: A primer on errors in clinical reasoning. *Medical Teacher, 25*(2), 177–181.

Kuhn, G. J. (2002). Diagnostic errors. *Academic Emergency Medicine, 9,* 740–750.

Mamede, S., & Schmidt, H. G. (2004). The structure of reflective practice in medicine. *Medical Education, 38,* 1302–1308.

Mamede, S., Schmidt, H. G., & Penaforte, J. C. (2008). Effect of reflective practice on accuracy of medical diagnoses. *Medical Education, 42,* 468–475.

Rikers, R. M. J. P., Schmidt, H. G., & Boshuizen, H. P. A. (2002). On the constraints of encapsulated knowledge: Clinical case representations by medical experts and subexperts. *Cognition and Instruction, 20*(1), 27–45.

Schmidt, H. G., & Boshuizen, H. P. A. (1993). On acquiring expertise in medicine. *Educational Psychology Review, 5,* 1–17.

Schmidt, H. G., & Rikers, R. M. J. P. (2007). How expertise develops in medicine: Knowledge

encapsulation and illness script formation. *Medical Education, 41,* 1133–1139.
Weingart, S. N., Wilson, R. M., Gibberd, R. W., & Harrison, B. (2000). Epidemiology of medical error. *British Medical Journal, 320,* 774–777.

Ethnographic Methods

The term *ethnography* describes both a literary genre (writings that attempt to capture people's cultural beliefs/practices) and a qualitative research methodology (a way of collecting social scientific data based on long-term, face-to-face interactions). In the current era, ethnographic analysis seems to have lost some of its authority, especially since human genomics and the statistical analysis of massive data sets are privileged in the search for contemporary solutions to social problems. Even still, ethnography is alive and well and can be used to inform medical decision making.

Data Collection

Anthropology and sociology are the two academic disciplines that traditionally cornered the market on ethnographic methods, but other social sciences have become more interested in the kinds of nuanced information that is gathered during intimate and ongoing interactions between qualitative researchers and their research subjects, interactions euphemized as "deep hanging out." Ethnographers spend time drinking beers with the folks they study, eating meals at their dinner tables, and shadowing them on the job—all in an effort to figure out what people's everyday lives actually look like and to determine how people make sense of those lives.

When they first start conducting research in a particular community, ethnographers may stand out like sore thumbs, drawing attention to themselves and making their research subjects self-conscious, which means that they run the risk of witnessing things that probably wouldn't have taken place at all without the conspicuous seductions of an outside audience. But as ethnographers spend more and more time observing and participating in the same community, among the same community members, they eventually begin to lose some of their distracting influence on people's behaviors. They transform into proverbial flies on the wall. The ethnographer is still there, asking questions and watching people's daily reactions, but is hardly noticed any more, not in ways that might compromise the reliability of what the ethnographer sees or hears.

Ethnography's value is based on the kinds of intimate and unguarded data that researchers gain from extended contact with one particular social group. When the discipline first emerged, this meant relatively small-scale and remote societies. Bronislaw Malinowski's early-20th-century work with Trobrianders is taken as a powerful marker for the birth of full-fledged ethnographic research within anthropology. He crossed the seas, pitched his tent, and found a way to live among people whose cultural world seemed radically different from his own. Part of the point, of course, was about making it clear to the European audience back home that those foreign practices could be understood only with the fullest knowledge of how people's entire belief systems fit together—even and especially when those cultural systems seemed spectacularly exotic to the Western eye.

Ethnography in Anthropology and Sociology

Anthropology was traditionally about studying societies unsullied by the advances of modernity. From the attempts at *salvage ethnography* among Native American tribes in the early 19th century (archiving cultural practices before they disappeared forever) to the constructions of primitive societies as examples of the modern Western world's hypothetical pasts, anthropologists used ethnographic methods to study those populations most removed from the taint of modern living.

Sociologists also embraced ethnographic methods in the early 20th century, and people like Robert Park at the University of Chicago helped institutionalize the *ethnographic imagination* as a method for studying not just faraway villages but also modern urban life in a teeming American city. That dividing line (between the anthropological ethnographer who studies some distant community and the sociological ethnographer who focuses her eyes on the modern Western metropolis) still defines most people's assumptions about how

those two fields carve up the social landscape for qualitative examination (even though there are certainly sociologists who study small-scale societies and anthropologists who have been working in urban America for a very long time).

Both fields sometimes seem to place a premium on something close to the scientific equivalent of roughing it. They each have the highest regard for the "gonzo" ethnographer, the kind of heroic or mythical figure willing to put his or her very life at risk for the sake of ethnographic access. The more remote, removed, and potentially dangerous the location of the fieldwork experience, the more explicit and awestruck are the kudos offered up to any ethnographer bold enough to go where few have gone before. This search for dangerous exoticism can lead one halfway around the world or just to the other side of the tracks, the other end of town. But in either case, an added value is placed on access to the everyday lives of human beings and cultural perspectives that most middle-class Western readers know little about.

During the 1960s, anthropologists and sociologists in the United States wrote classic ethnographic offerings on the urban poor—specifically, the black poor, who were struggling to make ends meet in America's ghettos. Ethnographers were trying to explain the hidden realities of urban poverty, a tradition that continues today. Anthropologists and sociologists working in American cities still disproportionately study poor minority communities. That's because it may be harder to entice wealthier Americans to accept such scholarly intrusions. A $20 bill might suffice as an incentive for unemployed urbanites to answer some open-ended questions about their life history (and to allow an ethnographer to shadow them on an average afternoon), but it may not be enough to persuade middle-class citizens to expose their raw lives to an ethnographic gaze. Middle-class and wealthier Americans also sometimes live in gated communities or attend restricted social clubs, to which anthropologists may not have access. These same kinds of biases also tend to predetermine the kinds of communities ethnographers have access to abroad.

Traditionally, ethnographers have been taught that they must master the culture of the groups they study so completely that they should almost be able to see the world from that group's point of view, almost as if they were born into the community. Anthropologists call this an "emic" perspective, something that can only be acquired with long-term participant observation—many months, even years, of deep hanging out with the people being studied.

Medical Anthropology

The growing subfield of medical anthropology interrogates the often masked cultural assumptions that subtly inform medical decision making on the part of both doctors and their patients. Medical anthropologists deploy ethnographic methods (a) to uncover the hidden ethnocentricisms that might sometimes allow doctors trained in the West to underestimate the value of folk medicinal practices; (b) to describe how the roles of "doctor" and "patient" are constructed from social and cultural templates that usually go unexamined or unspoken; and (c) to emphasize how broader cultural expectations and interpretations configure the way medical practitioners conceptualize/operationalize diseases and translate medical theories for a lay audience. Ethnographers such as Rayna Rapp and Paul Farmer mobilize ethnographic methods (studying medicine as an inescapably cultural—not just biological or genetic—domain) to mount critiques of presuppositions that sometimes obstruct professional attempts to negotiate the political, moral, and biological dilemmas of medical treatment.

Future Directions

Ethnographers have started to retool this methodological intervention for the newness of the empirical present, calling for (a) multisitedness, (b) a specific focus on the culture of medical research, and (c) particular emphasis on the challenges that come with studying a media-saturated world. Even still, there remains something inescapably troubling to some ethnographers about ethnographic attempts to study several places at once, to engage phenomena spread out over large expanses of space in ways that outstrip any ethnographer's ability to experience them directly and holistically. Does it mean sacrificing depth for breadth, and how much does that compromise ethnography's specific contribution to the

constellation of methodological options open to social scientific researchers?

John L. Jackson Jr.

See also Cultural Issues; Qualitative Methods

Further Readings

Cerwonka, A., & Malkki, L. H. (2007). *Improvising theory: Process and temporality in ethnographic fieldwork*. Chicago: University of Chicago Press.
Clifford, J., & Marcus, G. E. (Eds.). (1986). *Writing culture: The poetics and politics of ethnography*. Berkeley: University of California Press.
Farmer, P. (1993). *AIDS and accusation: Haiti and the geography of blame*. Berkeley: University of California Press.
Madison, D. S. (2005). *Critical ethnography: Method, ethics, and performance*. Thousand Oaks, CA: Sage.
Marcus, G. E. (1998). *Ethnography through thick and thin*. Princeton, NJ: Princeton University Press.
Rabinow, P. (2003). *Anthropos today: Reflections on modern equipment*. Princeton, NJ: Princeton University Press.
Rapp, R. (1999). *Testing women, testing the fetus: The social impact of amniocentesis in America*. New York: Routledge.
Willis, P. (2000). *The ethnographic imagination*. Cambridge, UK: Polity Press.

EuroQoL (EQ-5D)

EuroQol, also referred to as EQ-5D, is one of the multi-attribute health status classification systems. It is a generic instrument for measuring the health-related quality of life. Along with other multi-attribute health status classification systems, such as Health Utilities Index (HUI) and Quality of Well-Being (QWB), EuroQol is used as an alternative to measure the utility or health preference. Measuring utilities or preferences can be a complex and time-consuming task. EuroQol is attractive due to its simplicity. Thus, it has been widely used throughout the world in both clinical investigations and health policy determinations.

The EuroQol questionnaire was developed by the EuroQol group, original members of which came from various research teams in Europe. The name EuroQol comes from European Quality of Life. There are three components within the EuroQol questionnaire. The first component, the most important one, comprises five dimensions (5D): mobility, self-care, usual activities, pain/discomfort, and anxiety/depression.

EuroQol Questionnaire

The EuroQol group was established in 1987, with investigators coming from various countries in western Europe. The group has expanded into an organization with members from all over the world in 1994. The EuroQol questionnaire is designed for self-completion by the respondents, and it was initially developed to complement other health-related quality-of-life measures. The primary component of the EuroQol questionnaire originally had six dimensions: mobility, self-care, main activity, social relationships, pain, and mood. EuroQol has become a stand-alone questionnaire subsequently, and the primary component was revised to five dimensions, including mobility, self-care, usual activities, pain/discomfort, and anxiety/depression. It has been publicly available since 1990.

The EuroQol questionnaire has three components. The first component is the primary one that includes five dimensions. Each dimension is measured by a question that has three possible responses: no problem, some problem, or severe problem. A preference-based index score can be created based on the answers to these five dimensions. The second component of the EuroQol questionnaire is a visual analog scale, where respondents can indicate their current health status on a "thermometer" scaled from 0, *the worst imaginable health state*, to 100, *the best imaginable health state*. The third component of the EuroQol questionnaire is for respondents to answer their background information, including disease experience, age, gender, smoking status, education, and others. The first two components are the instruments to be used if the researchers are only interested in knowing the health-related quality of life from the respondents.

Preference-Based Scoring Algorithm

Since the first component of the EuroQol questionnaire has five dimensions, with each having

three levels of answers, the combination of responses results in 243 (3^5) possible health states. Methods were developed to assign preference scores to each of the 243 health states that represent an average preference for one state versus another. By adding two additional health states, "unconscious" and "dead" for a total of 245 health states, this method was initially developed based on a random sample of about 3,000 adults in the United Kingdom. The scoring function was developed using econometric modeling based on the time trade-off technique. The final preference-based index scores were assessed on a scale where 0 represents *a health state of being dead* and 1 represents *perfect health*.

The first component of the EuroQol has also been weighted according to the social preferences of the U.S. population. Similarly, the U.S.-based EuroQol preference-based scoring algorithm was developed using econometric modeling through the time trade-off technique. A representative sample of the U.S. general adult population with approximately 4,000 participants completed the interview. The interview was carried out in the United Kingdom and the United States in 1993 and 2002, respectively.

Application

As a quick and well-validated instrument, the EuroQol has been widely used in clinical and economic evaluations of healthcare as well as in population health surveys. The EuroQol is available in many languages. Most researchers use the first two components of the EuroQol questionnaire for respondents to rate their current health states. Both the preference-based index and the visual analog scale have been used in various ways, including establishing national and local population health status, comparing patients' health status at different times, and evaluating the seriousness of disease at different times. The EuroQol has also been used in a number of clinical areas to provide effectiveness outcomes during the drug approval process. Recent work has furnished a national catalog of the EuroQol preference-based index for all chronic conditions in the United States.

The score provided by the EuroQol is important in cost-effectiveness analysis, where quality-adjusted life year (QALY) has become increasingly used to assess the treatment outcomes in clinical trials and health economic evaluations. However, EuroQol is designed to measure generic and global health-related quality of life. It is not sensitive or comprehensive enough to measure disease-specific quality of life.

Analyzing the EuroQol Preference-Based Index in Regressions

Previous research has noted that the EuroQol preference-based index score is, similar to other utility scores, far from being normally distributed. Methods designed for continuous data, such as the ordinary least squares (OLS) regression, are often inappropriate for such data. The OLS models the conditional mean as a linear function of the covariates. The idiosyncrasies of the EuroQol index distribution demand that the residuals should not be assumed to be normal or have constant variance. Although versions of OLS exist that are valid without any distributional assumptions on the residuals, the special features of the EuroQol index are neglected by only modeling the conditional mean.

Several other methods have been proposed, including the Tobit model and the censored least absolute deviations estimator (CLAD). One important feature of the EuroQol index score distribution is that many individuals reported perfect health with their EuroQol index at 1.0, thus forming a spike. The Tobit model and the CLAD model are extensions of the OLS that treat the health status of these patients as being *censored* at 1.0; that is, their health status, if it can be mapped onto the scale of EuroQol index, would be larger than 1.0. In other words, these methods assume that there is an underlying latent health status variable. When it is less than 1.0, it is observed as the EuroQol index; when it is larger than 1.0, we only observe EuroQol = 1 as an indicator of censoring. These methods then model the conditional mean of the latent variable, instead of EuroQol itself, as a linear function of the covariates. The difference between the Tobit model and the CLAD model is that Tobit model assumes that the latent variable has a normal distribution, while in the CLAD model, the latent variable can have any continuous distribution. Therefore, the Tobit

model can be viewed as a special case of the CLAD model.

A two-part model approach has also been proposed to model the special features of the EuroQol index, particularly a large proportion of subjects having the score at 1.0. The first part is a logistic model for the probability of reaching the maximum score. The second part is a model for the rest of the scores that are less than 1.0, which can be either a least squares regression with robust standard errors for the conditional mean or a quantile regression for conditional quantiles such as the median. It has been shown that the two-part model has some desirable features that are not available in the aforementioned regression methods for the EuroQol preference-based index score.

Alex Z. Fu

See also Expected Utility Theory; Health Utilities Index Mark 2 and 3 (HUI2, HUI3); Quality of Well-Being Scale; Utility Assessment Techniques

Further Readings

Dolan, P. (1997). Modeling variations for EuroQol health states. *Medical Care, 35,* 1095–1108.

Drummond, M. F., Sculpher, M. J., Torrance, G. W., O'Brien, B. J., & Stoddart, G. L. (2005). Cost-utility analysis. In *Methods for the economic evaluation of health care programmes* (3rd ed., pp. 137–209). Oxford, UK: Oxford University Press.

Fu, A. Z., & Kattan, M. W. (2006). Racial and ethnic differences in preference-based health status measure. *Current Medical Research and Opinion, 22,* 2439–2448.

Li, L., & Fu, A. Z. (2008). Methodological issues with the analysis of preference-based EQ-5D index score. *Value in Health, 11,* A181.

Shaw, J. W., Johnson, J. A., & Coons, J. S. (2005). US valuation of the EQ-5D health states: Development and testing of the D1 valuation model. *Medical Care, 43,* 203–220.

Sullivan, P. W., & Ghushchyan, V. (2006). Preference-based EQ-5D index scores for chronic conditions in the United States. *Medical Decision Making, 26,* 410–420.

Sullivan, P. W., Lawrence, W. F., & Ghushchyan, V. (2005). A national catalog of preference-based scores for chronic conditions in the United States. *Medical Care, 43,* 736–749.

Evaluating and Integrating Research Into Clinical Practice

The impetus for evidence-based medicine (EBM), or its younger brother, evidence-based practice, has been that it takes too long for efficacious and effective treatments to be brought to bear in routine clinical practice. The usual time given is a 17-year delay between demonstration of efficacy and routine practice, although the evidence for this specific time frame is sparse. However, as a social value in medicine, most believe that it is better for patients to receive effective care than none, so regardless of the true time delay, researchers, healthcare administrators, policy makers, clinicians, and patients all now recognize as crucial the systemic issues that delay the integration of research into practice. Like medical care, addressing this issue requires diagnosis of the systemic issues that prevent the translation of research into practice (TRIP) and requires treatment based on those diagnoses.

Diagnosis

A number of different approaches have been used to diagnose the systemic barriers. One is the diffusion of innovation formalism. Rogers identified five components of diffusion: (1) relative advantage, (2) compatibility, (3) complexity, (4) trialability, and (5) observability. Berwick and Greenhalgh provide a general framework for applying these to medical care. Early studies documented the slow uptake of basic innovations and documented, for instance, from the physician's point of view, the need for observability—the need for a local champion. Later studies showed that apparently not much had changed; from the patient's perspective, only about 55% received recommendation-based care for preventive, acute, or chronic care. Cabana showed the application of a barrier-based framework to the (non)use of clinical practice guidelines (CPGs), touted as one solution to the TRIP problem. He discerned that barriers ranged from issues of physician self-efficacy to systemic difficulties in getting access to the guidelines as well as traditional concerns such as disagreement over applicability.

Treatment

There are two basic approaches to the incorporation of research-based evidence into practice: active and passive. *Active* means that the clinical practitioner must make the explicit effort of finding the evidence and evaluating it. *Passive* means that the environment has been architected to bring the evidence to bear automatically.

Active Approaches

The primary active approach has been to teach clinicians the process of EBM in the hope that they would use those methods at the bedside. Supporting this agenda has required several components. First, EBM resources have been needed. The primary one has been PubMed, which references several thousand journals and several million articles. Almost all EBM searches end up at PubMed (in English-speaking countries), because the latest, authoritative results are available there. Searches there depend on skillful use of the PubMed-controlled vocabulary—MeSH (Medical Subject Headings)—as well as free text and other specifics of the indexing system. The Cochrane Database of Systematic Reviews houses systematic reviews of studies (primarily randomized controlled trials) that, themselves, are often indexed on PubMed. However, these reviews are extensive, reproducible, and go beyond PubMed, to include unpublished articles or novel data provided by published authors. Perforce, these reviews are not as current as PubMed. CPGs go beyond Cochrane reviews in authority, because they include the definition of standard of practice, as defined by professional societies. Because of this added layer of vetting, CPGs are the least up-to-date but the most authoritative. Thus, a reasonable search strategy is to start with CPGs (as indexed or contained at the National Guideline Clearinghouse), then move on to Cochrane to see if there is anything newer, and then move on to PubMed to look for anything newer still.

Evidence searching goes beyond the searching of reference or full-text databases to include evaluation or appraisal of the report found. Tools that support this process include the *JAMA* reading guides and worksheets; both are available via the University of Alberta.

There are many sites on the Web that cater to clinicians. Each of them requires clinicians to go on their own through the cycle of searching and evaluating the retrieved evidence. Some sites such as the TRIP site in the United Kingdom search many sites for the user and bring them together. Choosing, appraising, and using any specific source is left to the user.

There are also a number of commercial tools that supply resources and levels of evidence, are kept up-to-date, and are available on handheld devices.

Finally, there is a small industry in teaching EBM methods to clinicians, whether in medical school, through journals, on the Web, or in continuing medical education (CME) classes.

Evidence shows that medical students can learn the methods, that physicians do not have time to use them, and that CME lecturing is the least effective way of learning a skill.

Passive Approaches

In passive approaches, barriers are broached by others. Pharmaceutical companies invest more than any others in educating clinicians about available evidence, but they are generally not thought to have the clinician's EBM process as their primary goal. On the other hand, pharmaceutical methods have been tried through *academic detailing*, where trained staff attempt to teach clinicians about the best evidence and most effective therapies with one-on-one encounters. While there have been successes, they have been frustratingly Pyrrhic and not clearly worth the investment required.

Decision support systems, embedded in clinicians' workflow, have been thought to offer the best possibility of getting the best evidence and practice before a clinician's eyes, with the system's blessing. There are several degrees of decision support relevant to TRIP.

The first type is generic access to relevant material. This access is usually to the CPG and leaves the user to read the text (if guided to it) and apply it as seen fit. The next level of access is more tailored, using an "Infobutton," where the computer system uses generic patient information to find generic information about that patient. So a patient's diagnosis of sickle-cell disease will be used by the system to provide the user with instant access to

definitions, normal values, textbook entries, and CPGs on sickle-cell disease. The next level of access is customized, where the system takes several pieces of information about the patient and provides access to yet more specialized information, say, a relevant PubMed reference. Systems providing such access are rare and, because of the difficulty of automating the EBM process, end up leaving it to the user to judge the applicability or evidential quality of the linked article, since there can be no ante vetting by the system builders.

The next class of decision support is guided choice, where evidence can be put into the workflow by making it difficult to act on the basis of bad evidence. Thus, the generic guided choice may be a calculator for total parenteral solution ordering that would prevent generic conditions such as high osmolarity or simultaneous inclusion of calcium and bicarbonate in the solution. The next level of decision support is tailored guided choice, such as order sets. Here, guideline-based care can be instituted by the healthcare organization by specifying, for example, that any discharge of a patient with a myocardial infarction will include a prescription for a beta blocker. Thus, rather than rely on the physician having read the CPG, the systematic review, and the most recent articles confirming the effectiveness of such prescribing; rather than rely on the physician remembering this effectiveness at the time of discharge; and rather than rely on the physician ordering it, the system provides a checkbox for the physician; checking that box represents the entire evidence search and evaluation cycle. The challenge is for the system to know that the specific patient had a myocardial infarction at the time of discharge and to make sure that the physician is using the computer system to enter discharge orders.

Customized guided choice is possible as well but is generally not available. Here, the system composes a checklist, say, for the specific patient. While composing a checklist from a union of other checklists is clearly easily done, checking for interferences, dependencies, and other interactions is much less so.

The third class of decision support is knowledge-based prompts; these are the classic alerts, where the physician has ordered something and the machine responds that that order is in error, or the physician has not ordered something and the computer recommends an action. The knowledge behind these alerts is generally framed as rules, and these rules are usually referenced to the literature. While the knowledge of effectiveness would seem to be the same across institutions, the ideal of sharing rules has not been borne out by the realities of system implementation because of the variety of ways in which different systems store the information needed by the rules. Thus, each institution is left to vet its own rules on its own. In addition, commercial entities that sell knowledge bases, such as those containing evidence-based drug-drug interactions, are concerned with their own risk profile and so include a wide range of interactions that make the systems generally unusable, leaving institutions, again, to face the decisions themselves over what alerts to keep and what not even to show the physician.

The evidence on all such systems is mixed. Kawamoto and colleagues' systematic review showed the systems to have a positive impact 68% of the time and confirmed the factors most likely to lead to success: providing decision support within the context of the workflow at the place and time the action was needed, providing action items (not just assessments), and using a computer-based system. The harms that provider-order entry systems have demonstrated recently have not been related to evidence-based decision support. However, too much experience shows that the low specificity and high sensitivity of the alerts leads to "alert fatigue" and inattention when the system cries wolf.

The Future

Interventions for evidence-based practice are based on the experience of EBM but with application to different domains. Evidence-based nursing has led to specific resources for nurses but not the depth of computer-based support that clinicians have available to them. Evidence-based public health has focused on clinical issues and not on the more systemic interventions that public health practitioners must effect nor on the more global concerns that affect their work. There are some generic and guided-choice-based tools for decision support, but outside of biosurveillance, there is little decision support based on knowledge-based prompts, and in biosurveillance, the alerts are not necessarily based on research evidence. Each of these areas will likely grow in the future.

The National Institutes of Health's initiative regarding Clinical and Translational Science Awards will push innovation to the "left" side of the translation and evidence-generation process. The new innovations may aggravate matters by generating too many technologies to accommodate—or may induce a new attention to the entire translation process on the "right"-hand side. Such attention jibes well with the new attention given to the care provider system itself. Computer-based decision support systems seem to be the best bet for bringing evidence into practice. A further source of evidence will be the electronic patient record itself, as the data from operational use are stored in clinical data warehouses for mining and local research. Such research will overcome several of Roger's barriers:

1. Relative advantage could be assessed directly or modeled, based on local data.
2. Compatibility could be assessed by reviewing the number and types of patients to whom the new evidence (technology) applies.
3. Complexity could be assessed through an environmental scan of clinic, unit, and staff capabilities.
4. Trialability could be assessed through pilot projects whose data are made available in a regular manner.
5. Observability could be achieved by review of the data warehouse data of patients treated with the new technology.

It may just require linking the workaday, sloppy observational data of routine care with the pristine results of carefully constructed studies to achieve the long-wished-for goal that patients receive the best care that science says they should receive. This possibility provides researchers with further challenges in providing healthcare institutions and clinicians the new tools they need to achieve this synthesis.

Harold Lehmann

See also Clinical Algorithms and Practice Guidelines; Computational Limitations; Evidence-Based Medicine; Randomized Clinical Trials

Further Readings

Avorn, J., & Soumerai, S. B. (1983). Improving drug-therapy decisions through educational outreach: A randomized controlled trial of academically based "detailing." *New England Journal of Medicine, 308*(24), 1457–1463.

Berwick, D. M. (1975). Disseminating innovations in health care. *Journal of the American Medical Association, 289*(15), 1969–1975.

Cabana, M. D., Rand, C. S., Powe, N. R., Wu, A. W., Wilson, M. H., Abboud, P.-A. C., et al. (1999). Why don't physicians follow clinical practice guidelines? A framework for improvement. *Journal of the American Medical Association, 282*(15), 1458–1465.

Cimino, J. J., Aguirre, A., Johnson, S. B., & Peng, P. (1993). Generic queries for meeting clinical information needs. *Bulletin of the Medical Library Association, 81*(2), 195–206.

Cochrane Database of Systematic Reviews: http://www.cochrane.org

Greenhalgh, T., Robert, G., Macfarlane, F., Bate, P., Kyriakidou, O., & Peacock, R. (2005). Storylines of research in diffusion of innovation: A meta-narrative approach to systematic review. *Social Science & Medicine, 61*(2), 417–430.

Grimshaw, J. M., Thomas, R. E., MacLennan, G., Fraser, C., Ramsay, C. R., Vale, L., et al. (2004). Effectiveness and efficiency of guideline dissemination and implementation strategies. *Health Technology Assessment, 8*(6), iii–iv, 1–72.

Kawamoto, K., Houlihan, C. A., Balas, E. A., & Lobach, D. F. (2005). Improving clinical practice using clinical decision support systems: A systematic review of trials to identify features critical to success. *British Medical Journal, 330*(7494), 765.

Institute of Medicine. (2001). *Crossing the quality chasm: A new health system for the 21st century*. Washington, DC: National Academy Press.

McGlynn, E. A., Asch, S. M., & Adams, J. (2003). The quality of health care delivered to adults in the United States. *New England Journal of Medicine, 348,* 2635–2545.

National Guideline Clearinghouse: http://www.guideline.gov

Perreault, L. E., & Metzger, J. (1999). A pragmatic framework for understanding clinical decision support. *Journal of Healthcare Information Management, 13*(2), 5–21.

PubMed: http://www.ncbi.nlm.nih.gov/sites/entrez?db=pubmed

Rogers, E. M. (1995). *Diffusion of innovations* (4th ed.). New York: The Free Press.

University of Alberta, Evidence Based Medicine Toolkit: http://www.med.ualberta.ca/ebm

Williamson, J. W., German, P. S., Weiss, R., Skinner, E. A., & Bowes, F. (1989). Health science information management and continuing education of physicians: A survey of U.S. primary care practitioners and their opinion leaders. *Annals of Internal Medicine, 110,* 151–160.

Evaluating Consequences

Decisions in medical contexts have immediate and obvious consequences in terms of health and sometimes death or survival. Medical decisions also have less obvious and less immediate consequences, including effects on the long-term physical and mental well-being of patients, their families, and caregivers, as well as on the distribution of scarce medical resources. Some of these consequences are hard to measure or estimate. Even harder, perhaps, is the determination of the relative value of different consequences. How should consequences be evaluated? How do uncertainties and biases affect our evaluations? What influence should our evaluations of consequences have on our actions? These questions are all philosophical in nature.

Consequences and Value

To evaluate something is most basically to determine its value or to determine its effect on that which has value. The positive value of health may be taken as a given in medical decision making. Sometimes, however, it is not clear what concrete outcomes contain more health. Will a patient in chronic pain be more healthy taking opiates that reduce her mental abilities and may create dependency, or will she be more healthy without opiates but with more pain? Will an elderly patient with myeloma enjoy better health after treatment with cytostatics that pacify the disease but weaken the immune system, or will his health be better without the treatment? Depending on the details of the case, the answers to these questions are far from obvious, showing that the concept of health is complex and will sometimes stand in need of specification.

Health may be defined biomedically as the absence of disease and infirmity. This is the common definition in medical practice, though seldom explicitly stated. Alternatively, health may be defined biopsychosocially, which is common in theoretical contexts. The 1946 constitution of the World Health Organization (WHO) states that health is "a state of complete physical, mental and social well-being." Several recent definitions aim to avoid the somewhat utopian character of the WHO definition and to shift focus from outcome to opportunity, by defining health in terms of potential or ability rather than well-being.

Quantitative measurements of health have increasingly been made in terms of quality-adjusted life years (QALYs), that is, the number of person life years adjusted by a factor representing the quality of the person's life. Like health, quality of life may be defined biomedically or biopsychosocially, and more or less broadly. What will be said in the following about values in general and health in particular holds equally for quality of life. Regardless of how exactly quality is defined, evaluating consequences in terms of QALYs incorporates a richer understanding of why we value life, as opposed to measuring only years of life of whatever quality or only death or survival. A strategy of QALY *maximization* has the further advantage of allowing quantitative comparisons of different alternatives, such as treatment programs, but has the disadvantage that other values may be disregarded, such as equity and autonomy.

Like any value, the value of health may be final and/or instrumental. Health is obviously instrumental to other values such as happiness and achievement. In other words, we need health to promote or protect these other values. In addition, however, health may also be of final value—of value in itself, independently of its impact on other values. Whether or not health has final value becomes important in conflict cases, where it must be balanced against other values. If, for example, health, defined biomedically, is important only because of its instrumental contribution to the higher value of happiness, a healthy life without happiness has no value. This conclusion may have direct relevance for important medical decisions concerning life and death, including the issue of euthanasia.

Values may be subjective or objective. That the value of health is subjective would mean that health is of value only to the extent that the individual patient considers it to be of value or to the extent that she desires it. That the value is objective, on the other hand, would mean that health may be of value despite the fact that the patient does not subjectively value it. That a value is objective does not mean that it is insensitive to individual preferences, since objective values depend on individual preferences indirectly. Even if happiness, for example, is objectively valuable, what makes people happy depends on their preferences. Similarly, even if health is objectively valuable, what makes people healthy will depend on their physical constitution and individual character, including preferences. Whether values are subjective or objective naturally affects how we should treat each other in medical and other contexts.

Beyond the somewhat related values of health, quality of life, well-being, and happiness, autonomy is arguably the main value relevant for medical decision making. This value is institutionalized through the practice of informed consent, but it may be affected also in other ways. For example, addictions may be considered to decrease autonomy, and so treatment of addiction may promote autonomy. Further values of possible relevance include dignity, equity, personal relationships, and perfection or excellence. Dignity may be relevant to hospice care and other care of dying patients, equity to any decision affecting the distribution of scarce medical resources, relationships to how families are treated and to decisions affecting the patients' potential to uphold personal relationships after treatment, and perfection to neonatal screening and genetic and medical enhancement.

Which things have objective value, if any, is a fundamental philosophical question, and opinions and theories diverge. Lacking agreement, we may look to social value as determined by willingness to pay or stated preference; to politically, ideally democratically, determined values; to expert judgment; or to our own judgment. Again, opinions and theories diverge. The consequences of decisions should be evaluated in terms of those things that are determined to have value.

If more than one value is affected by a decision, as seems likely for most medical decisions, we must determine how these values relate to each other. Most fundamentally, values may or may not be commensurable. If the value of health and the value of autonomy are incommensurable, we cannot weigh one against the other and so must make decisions that affect both values without guidance from such weighing. If the values are commensurable, they may be more or less open to comparison. At one end of the spectrum, we may know only that a little health is less important than a lot of autonomy, but we may not know how to compare much of each or little of each. At the other end of the spectrum, any amount of each value may be represented by a number and the values aggregated in multi-attribute utility analysis. The very different character of some values may make them seem incommensurable, while the need to make decisions that affect more than one value forces us to compare them, or at least to act as if we had compared them.

Uncertainties and Biases

In evaluating consequences, we are inescapably faced with a number of uncertainties and biases. It is widely recognized that we do not even know if established medical practice on the whole efficiently promotes best outcomes (though the growing field of outcomes research aims to address that question). The uncertainty is naturally greatest for consequences of decisions not yet made. We often do not know what consequences will follow from alternative courses of action. In evaluating possible future consequences, these uncertainties can to some extent be handled by decision theoretical methods. If we are uncertain about what consequences will follow, we may at least know, or be able to estimate approximately, the probabilities of different possible outcomes, each with a set of consequences. Given these probabilities, we may estimate the expected value of different alternatives. To a large extent, however, uncertainty about the future must simply be accepted as a fact of life.

Uncertainty does not pertain only to future consequences but also to the value of consequences, future as well as past and present. Even if we know that we value health and we know the consequences of a certain decision, we might not know to what extent those consequences further our values. This may be because we are not certain how exactly our values should be specified or because

we are not certain how much the concrete consequences contribute to our values, however thoroughly specified. For example, if health is defined in terms of ability, we may not know to what extent successful treatment of radical mastectomy will contribute to this value. A person's overall ability depends partly on her attitudes, and patients may react differently to this medical procedure even when the physical outcome is the same.

Uncertainty about the value of consequences is increased by different sorts of biases. We tend to exaggerate the impact of certain things and belittle the impact of others. Some biases concerning our own well-being have been rather straightforwardly proven by psychological research. For example, we tend to overvalue variation in our consumption in the sense that we opt beforehand for variation but regret this once we get it. Other biases are harder to prove. For example, we value good things in the near future higher than similarly good things in the more distant future, and the reverse for bad things. This means, for example, that the social value of QALYs in the distant future is much lower than the social value of the same number of life years and QALYs in the near future. Whether this is an irrational bias that should be compensated for or an indication of our true values is a matter of controversy.

Uncertainties about consequences introduce another level of value—it requires us to determine how much we value certainty. A program of maximization of expected QALYs presumes that 1 QALY for sure is as good as a one-in-two chance of 2 QALYs. This is not so if we are risk-averse, that is, if we value goods that we are certain to get higher than goods we may or may not get, even when the expected value is the same. In fact, people tend to be risk-averse. However, this may be considered an irrational bias.

Consequences and Principles

In bioethics, principles are often understood as nonrigid rules and recommendations that must be interpreted in concrete cases with a large dose of moral judgment. Such principles are essentially statements of what has value, with the add-on that we have a duty to promote or protect that value. The question of which bioethical principles there are and how they should be understood corresponds to the question of what values there are and how they should be understood. Whether one prefers duty talk or value talk depends on whether one finds duty or value to be the more fundamental moral category. This is another matter on which opinions or sentiments diverge.

There are other kinds of principles, however, that do not as closely resemble values but that rather regulate the evaluation of consequences. Some of these principles are rules of thumb, stating that for practical reasons such as time constraint and limited information and information processing capacity, we should restrict our evaluation of consequences in different ways. A rule that the most severely injured should be treated first may be such a rule. It is not a deep moral truth that the most severely injured deserves the first treatment, but in most cases, the rule is fair and efficient and reasonably easy to follow without time-consuming judgment. That this is a rule of thumb rather than a fundamental principle is shown by our reactions to the hypothetical case where there are obvious reasons to diverge from the rule, for example, when it is clear that the most severely injured will not benefit from quick treatment while others will. If diverging from the rule in such circumstances is morally unproblematic, then the rule is one of thumb. In contrast, while a moral principle may be overridden, this is not unproblematic but normally gives cause for regret and may give rise to residual obligations.

Rules of thumb replace or restrict evaluations of consequences for practical reasons. Moral principles do so for moral reasons. There are essentially two sorts of moral principles. Action-focused principles, or side constraints, state that certain things must or may not be done, regardless of other considerations. Examples include general principles such as "never lie" as well as specific medical principles such as "never force medical care on a patient against her explicit wish." Reason-focused or value-focused principles, in contrast, state that certain reasons or values should be disregarded in the molding of various considerations into an all-things-considered judgment of what should be done. An example is the principle that a patient's estimated future contribution to society should not influence our medical treatment of the patient.

Many principles are tied to our social and legal roles, for example, as medical practitioners.

These roles come with social expectations, rules, and laws, which regulate how and to what extent we may consider certain consequences of our actions. If such role principles are motivated only by expedience, they may be seen as rules of thumb. However, if they become ingrained in the culture of a society, they acquire the status of moral principles. Even as rules of thumb, role principles are unusually rigid, because they are motivated by practical reasons on a collective or system level. While individual practitioners may on occasion have the time and capacity to judge a case on its own merits, they may be obliged to follow rules nonetheless, because this makes for stability and transparency in the medical system as a whole. The rigidity of role principles should not be exaggerated, however. The social and legal frameworks rarely, if ever, determine in detail how we should act and think. Even in applying well-defined rules, we need value judgments to guide our application of those rules to particular circumstances. Furthermore, as rational and moral beings, we can always question the social and legal framework within which we live and work.

A Model for Evaluating Consequences

The different aspects of evaluating consequences covered above may be captured in the following model. This somewhat novel model incorporates a series of not-so-novel considerations. The model does not describe how evaluations are performed in practice but rather proscribes what steps should be taken in order that all the aspects of evaluation discussed above be considered. In other words, the model is not psychological but philosophical. If implemented in practice, the steps of the model should not necessarily be taken in strict order. In particular, Steps 2, 3, and 4 may all require glancing ahead to subsequent steps.

1. Determine which things have value—that is, which values there are. This includes deciding whether values are subjective or objective, and final or instrumental.
2. Determine the available alternatives.
3. Decide whether an alternative is demanded by principle. If so, act.
4. Decide whether some alternatives are forbidden by principle. If so, exclude them from further consideration. If only one alternative is not forbidden, act.
5. Estimate for each alternative the possible outcomes and the (approximate) probability of each outcome.
6. Estimate the consequences of each outcome in terms of each value; adjust for bias.
7. Decide whether the consideration of some values is forbidden by principle, and if so, disregard these values.
8. Estimate the expected consequence of each alternative in terms of each value.

9a. If values are commensurable, estimate or decide the overall value of each alternative and act on the best alternative.

9b. If values are incommensurable, act on the alternative with the most appealing or most acceptable mix of expected consequences.

Kalle Grill

See also Bioethics; Construction of Values; Decision Rules; Expected Utility Theory; Health Outcomes Assessment; Moral Factors; Multi-Attribute Utility Theory; Outcomes Research; Protected Values; Quality-Adjusted Life Years (QALYs); Quality of Well-Being Scale; Risk Aversion; Values; Willingness to Pay

Further Readings

Beauchamp, T. L., & Childress, J. F. (2001). *Principles of biomedical ethics* (5th ed.). Oxford, UK: Oxford University Press.

Bircher, J. (2005). Towards a dynamic definition of health and disease. *Medicine, Health Care and Philosophy, 8,* 335–341.

Griffin, J. (1986). *Well-being: Its meaning, measurement and moral importance*. Oxford, UK: Oxford University Press.

Hastie, R., & Dawes, R. M. (2001). *Rational choice in an uncertain world*. Thousand Oaks, CA: Sage.

Kane, R. L. (2006). *Understanding health care outcomes research*. Sudbury, MA: Jones & Bartlett.

O'Neill, O. (2001). Practical principles and practical judgment. *Hastings Center Report, 31*(4), 15–23.

Raz, J. (2003). *The practice of value*. Oxford, UK: Oxford University Press.

Savulescu, J., Gillon, R., Beauchamp, T. L., Macklin, R., Sommerville, A., Callahan, D., et al. (2003). Festschrift edition in honour of Raanan Gillon. *Journal of Medical Ethics, 29,* 265–312.

Schroeder, M. (2008). Value theory. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Spring 2008 ed.). Retrieved February 2, 2009, from http://plato.stanford.edu/archives/spr2008/entries/value-theory

World Health Organization. (1946). Constitution of the World Health Organization. In *Basic documents*. Geneva, Switzerland: Author.

Evidence-Based Medicine

Evidence-based medicine (EBM) is the judicious application of the best, relevant clinical study results to patient care. EBM is not a new form of medical practice. It neither replaces medical expertise nor ignores patient preferences. EBM is a tool to enhance medical practice. While it is axiomatic that clinicians are interested in using the results of clinical studies for their patients' benefit, until recently, lack of access, limited critical analysis skills, and overreliance on expert opinion, personal experience, and clinical habit have hampered the rapid integration of high-quality clinical study evidence into clinical practice. This entry discusses how EBM and the EBM process address this issue and reviews the origins of EBM and its scope, resources, and role in modern clinical practice.

Origins

The term *evidence-based medicine* was coined in 1990 by Gordon Guyatt. While there have been many contributors to the development of EBM, Guyatt and his colleagues at McMaster University—principal among them David Sackett and Brian Haynes—have played major roles in developing the principles of EBM and have been instrumental in popularizing it throughout the world. In 1985, Sackett, Haynes, and Guyatt, together with Peter Tugwell, published the book *Clinical Epidemiology: A Basic Science for Clinical Medicine*. In this book, the authors explained, simplified, and organized the basic EBM principles (though not yet referred to as EBM) for the practicing clinician. In essence, this was the first EBM book, which served as the basis for their later books and articles that generated and developed the EBM approach to clinical practice.

Scope

EBM was developed for practicing physicians. However, over the past decade, it became increasingly clear that many other professions participating in patient care would benefit equally from the EBM approach. In recent years, dentistry, nursing, pharmacy, physical therapy, occupational therapy, public health, library sciences, and other disciplines have developed a strong interest in EBM. With this broadened focus, the term EBM is slowly being replaced with EBP, evidence-based practice.

The past decade also has seen the rapid spread of EBM learning in all aspects of medical training. It is routine now in medical schools and residency programs in North America, Europe, and elsewhere to include EBM as a standard part of their curriculum. The acquisition of skills of critical judgment is a requirement of the Liaison Committee on Medical Education (accreditation committee for medical schools) in the United States. EBM learning is explicitly mentioned in the standards of the Accreditation Council for Graduate Medical Education and is included among the subcategories of their six core competencies for residency programs in the United States.

One remarkable corollary of EBM's increasing popularity has been the encouragement and expectation of scientific rigor in clinical research. This has led to the ubiquitous use of statistical methods to evaluate results, the rise of the randomized controlled trial as the standard for determining therapeutic benefit, and greater attention generally to methodological validity across all areas of clinical investigation.

Finally, the emergence of EBM has brought about a changed relationship of the practicing clinician to the medical literature. Previously, the busy clinician was forever attempting to catch up on journal reading—most poignantly represented by unread stacks of journals lying forlorn in the corner of one's office. The EBM process has encouraged decreasing catch-up journal reading and increasing patient-focused journal reading. The patient encounter has become the catalyst of learning about new treatments, diagnostic tests,

and prognostic indicators and the focus of a personal program of continuing medical education.

Process

The EBM process begins and ends with the patient. The patient is the impetus for and genesis of a four-step approach to knowledge acquisition, ending in application of that new knowledge to the care of the patient. The four steps include (1) formulating the clinical question, (2) searching for and acquiring evidence from the medical literature, (3) assessing that evidence for methodological validity and analyzing the study results for statistical significance and clinical importance, and (4) applying, where appropriate, the valid important study results to the patient.

Formulating the Clinical Question

The first step in the process is recognizing that one has a knowledge gap in some aspect of a specific patient's management and to craft this knowledge gap into a focused question. One of the more popular approaches is the PICO (patient/problem, intervention, comparison, outcome) format. An example of this type of format is as follows. P: In otherwise healthy infants with presumed herpetic gingivostomatitis, I: what is the therapeutic efficacy of acyclovir, C: compared with placebo, O: in reducing the time to resolution of symptoms. The PICO format helps clarify and focus the clinician's specific evidence needs as well as to suggest an evidence search strategy, employing the terms in the question as online literature search terms. Carl Heneghan and Douglas Badenoch reviewed the mechanics of using the PICO approach in their book *Evidence-Based Medicine Toolkit*. A useful tool is the PICOmaker from the University of Alberta (www.library.ualberta.ca/pdazone/pico), which provides a palm-based platform for developing and saving clinical questions.

Searching for the Evidence

Primary Evidence

The next step in the EBM process is to look for an answer. Brian Haynes has suggested an *information hierarchy* to assist clinicians in their search for evidence. There are five levels in the hierarchy (from lowest to highest): primary studies, syntheses, synopses, summaries, and systems—the assumption being that the higher one ascends the hierarchy, the more reliable and applicable the evidence. It follows that one would begin an evidence search at the highest level. However, in practice, the systems level of evidence is rarely encountered. Summaries are more commonly found, synopses even more common, and so on. At the lowest level are primary studies. These are the individual clinical investigations that form the corpus of the clinical scientific literature. In many instances, one's search for evidence will end here as the other levels of the hierarchy are not available. The two most commonly employed search engines for identifying primary studies include PubMed and Ovid, the former being free. PubMed searches the U.S. National Library of Medicine online medical database MEDLINE. Ovid can search other databases as well (such as the European-based EMBASE). Methodological quality filters, designed to identify only the highest-quality studies, have been incorporated into PubMed, under the "Clinical Queries" section, and are available for Ovid as well.

Syntheses

The next level up in the information hierarchy is syntheses, including systematic reviews and meta-analyses. A systematic review is a study that answers a focused clinical question using all relevant primary research studies. When appropriate, these results may be mathematically combined, resulting in a new, combined estimate of the outcome. This last step is termed *meta-analysis*. The assumption underlying placing systematic review/meta-analysis at this level of the hierarchy is that consideration of results from multiple studies is more reliable than consideration of the results of any one individual study.

Systematic reviews and meta-analyses are increasingly popular in the medical literature. They are indexed in the clinical databases, and there are specific search engines to help locate them, including one in PubMed. A key source for these types of studies is the Cochrane Collaboration, named in honor of Archie Cochrane, a British physician and researcher who, in 1979, challenged the medical community to critically

summarize—by specialty and with periodic updates—all relevant scientific literature. The Cochrane Collaboration is a fellowship of volunteers from around the world who produce systematic reviews following the exacting Cochrane guidelines. In addition to the collection of systematic reviews, the Cochrane Collaboration Web site is home to a listing of hundreds of thousands of controlled therapeutic trials. Cochrane reviews are listed in PubMed, but the full review is available only by subscription.

Synopses

The next level up on the information hierarchy is the synopsis. A synopsis is a brief review of a primary study or systematic review, summarizing the key methodological issues and results, as well as pointing out areas of concern and caution. The purpose of this type of resource is to free the clinician from the bother of critically appraising the methodology and results of a study. Synopses come in many shapes and sizes, ranging from high-quality, peer-reviewed sources, such as ACP Journal Club, Bandolier, DARE, and Evidence-Based Medicine, to private online collections of study reviews of questionable quality.

Summaries

Summaries are reviews of all the methodologically sound evidence on a medical topic. An example of a topic would be "the treatment of asthma with bronchodilators," in contrast to a question relating to a specific outcome and a specific bronchodilator. Summaries allow for comparison of various competing therapeutic options for a specific problem, all evaluated using the analytical tools of EBM. Sources for summaries include clinical evidence and PIER (Physicians' Information and Education Resource; http://pier.acponline.org/index.html).

Systems

The highest level in the information hierarchy is systems. The most sophisticated of these would be a computerized decision support system, linking a patient's specific medical characteristics and preferences to the best available evidence and then recommending a specific management approach. These are not widely available. Included in this category, though much less sophisticated, are clinical guidelines. However, a note of caution is in order. Similar to the situation mentioned above with regard to synopses, the quality of guidelines varies. Guyatt and others have developed a grading system (the GRADE approach) for assessing the quality of a guideline based on methodological validity and the results of the guideline's evidence base.

Analyzing the Evidence

This is the EBM step that has been called "critical appraisal." If one has found a piece of evidence in the synopsis, summary, or systems categories above, this step should have been completed by the authors. If not, the methodology and results require critical analysis. Medical schools, residencies, and other clinically oriented programs are now providing instruction in EBM, much of which focuses on critical appraisal. There are many online and print resources that can aid in critical appraisal of a study and in critical appraisal skill learning. This is the crucial step in the EBM process determining whether the results of the study are reliable and important and therefore worthy of consideration in the management of one's patient. This is also the point at which one may decide that the study validity is insufficient and therefore one should look for other evidence.

Applying the Results

This is the final step of the EBM process—bringing the results of the EBM analysis back to the patient. At times, this step may be very straightforward—when benefits clearly outweigh costs, such as vaccination against pneumococcal infection and IVIG treatment in Kawasaki disease. At other times, the decision is not straightforward, even when the results of a study are highly valid, such as surgery versus medical treatment for prostate cancer. In all cases, patient peculiarities and preferences need to be factored into the medical decision. There is an entire discipline, medical decision making, which is dedicated to aiding clinicians and their patients in making medical management choices. The field is still relatively new, and results from research are only beginning to be applied in the clinical setting.

Future Directions

The rise of EBM has changed the way clinicians approach patient management problem solving, medical education at all levels, and the scientific basis for clinical investigations. There is a significant amount of work currently under way to simplify the EBM process, automate it, and generally make it more user-friendly for the clinical consumer. On the horizon are the decision systems discussed above that will link high-quality, current research with specific patient characteristics and preferences. Given the remarkable ascendancy of EBM in a relatively short period of time and its broad acceptance, the demand among clinicians for such systems will likely spur their rapid development.

Jordan Hupert and Jerry Niederman

See also Diagnostic Tests; Number Needed to Treat

Further Readings

ACP Journal Club: http://www.acpjc.org
Bandolier: http://www.jr2.0x.ac.uk/Bandolier
Clinical Evidence: http://clinicalevidence.bmj.com/ceweb/index.jsp
Cochrane Collaboration: http://www.cochrane.org
CRD database—DARE: http://www.crd.york.ac.uk/crdweb
Evidence-Based Medicine: http://ebm.bmj.com
GRADE Working Group. (2004). Grading quality of evidence and strength of recommendations. *British Medical Journal, 328,* 1490–1497.
Guyatt, G., & Drummond, R. (2002). *Users' guides to the medical literature.* Chicago: AMA Press.
Heneghan, C., & Badenoch, D. (2006). *Evidence-based medicine toolkit* (2nd ed.). Malden, MA: Blackwell.
PICOmaker, University of Alberta Libraries: http://www.library.ualberta.ca/pdazone/pico
PIER: http://pier.acponline.org/index.html
Schwartz, A. (2006). *Evidence-based medicine (EMB) decision tools.* Retrieved February 2, 2009, from http://araw.mede.uic.edu/~alansz/tools.html
Straus, S. E., Richardson, W. S., Glasziou, P., & Haynes, R. B. (2005). *Evidence-based medicine. How to practice and teach EBM* (3rd ed.). Edinburgh, UK: Elsevier.
University of Illinois, Library of the Health Sciences, Peoria. (2008). *Evidence-based medicine.* Retrieved February 2, 2009, from http://www.uic.edu/depts/lib/lhsp/resources/ebm.shtml

Evidence Synthesis

There is a plethora of information in almost every area of healthcare. For example, a search of MEDLINE (the U.S. National Library of Medicine's bibliographic database) using only the terms *depressed, depressive*, or *depression* yields more than 3,000 hits of articles published since 1980—and MEDLINE is only one of many health-related electronic bibliographic databases. The same terms on the search engine Google on the World Wide Web yield upward of 84,000,000 hits. Yet informed health-related decision making is dependent on having access to current knowledge. Without some help in assembling, organizing, and summarizing this information, the patient, healthcare practitioner, or policy maker would be at a loss to navigate through this mass of information. The vast amount of information available gives rise to the need for literature reviews that synthesize the available evidence to provide an overall reflection of the current knowledge base. Yet evidence synthesis itself is not a simple or straightforward task. There are many different factors that should be considered and many different views on how evidence synthesis should be conducted.

Types and Sources of Evidence

One of the important factors to be considered in both carrying out and making use of a synthesis is the tremendous diversity in the types and sources of evidence that a synthesis might potentially consider. Most current evidence syntheses restrict themselves to research studies published in peer-reviewed journals. Yet even this restriction can yield an overwhelming amount of evidence, given the numerous electronic biographic databases that can be searched (of which MEDLINE, PubMed, PsycINFO, and EMBASE are only a few examples) and the number of languages in which health research studies are published. Other evidence synthesis methodologies also strive to include "fugitive" literature, that is, the search for evidence is expanded to include unpublished studies through searching conference proceedings and the Internet, by including relevant government reports and unpublished studies conducted by pharmaceutical companies, and by using personal networking to identify other studies that

may not have been submitted to or published by peer-reviewed journals. The main advantage of this strategy is that there is still a bias on the part of journals to publish studies that report positive findings. However, an important limitation of including these reports is that they have not undergone (or passed) a peer review process. Plus, attempting to include all published and unpublished studies can become an overwhelming task because of the sheer volume of information.

Still other evidence syntheses draw on professional expertise and opinion. Although professional expertise must surely be considered a form of evidence, good research studies, where they are available, should take precedence in an evidence synthesis. However, professional expertise and expert consensus can be especially useful when there are few studies available, and they are invaluable in helping interpret the meaning of and importance of study findings where they do exist. Issues of practicality and evidence of public acceptance of a particular practice (e.g., a particular prevention or intervention strategy) should also be important considerations, although not all evidence syntheses attend to these issues. However, in using an evidence synthesis to aid in clinical or policy decision making, these are important issues to consider.

Currently, most evidence syntheses combine evidence from quantitative studies only, and most synthesis methodologies do not consider the role of theoretical frameworks or the context in which a particular study (with its particular findings) was conducted. This may be of more relevance in reviews of some health issues (such as the prognosis in whiplash-associated disorders or most effective strategies for prevention of teen pregnancies) and of less relevance in reviews of other health issues (such as the efficacy of a particular medication for treating a particular disease). A recently emerging area, requiring a very different synthesis methodology, is the synthesis of evidence from qualitative studies. The combination of evidence from both quantitative and qualitative studies is even less well developed but would conceivably yield a rich and more complete understanding of many health concerns.

Considerations for Users of Syntheses

Even where the primary source of information is published quantitative studies, there is a great diversity in the particular questions addressed in those studies. For example, when synthesizing information from published intervention studies, there are a number of different questions that the study might address. Among others, these questions might include the following: Can an intervention be effective? What happens when a particular intervention is used on a widespread basis with the "usual" patient in the "usual" clinical setting? What interventions do patients prefer? Is that intervention *cost-effective* when used on a large-scale basis? In studies of a single intervention strategy, these questions might yield very different answers. An evidence synthesis is useful to a particular reader only when it addresses the same question that concerns the reader and where the study participants and clinical care providers are similar to those of the reader's interest. For example, a particular procedure carried out on the "ideal" patient by a highly specialized healthcare provider may yield very different results when carried out on "usual" patients (who may have comorbid conditions that are different or more serious than those in the studies included in the evidence synthesis). Similarly, some procedures have a higher success rate and lower complication rate when carried out at highly specialized and selected facilities than at facilities providing more varied and generalized care. The reader of an evidence synthesis needs to ensure that the synthesis is relevant to the particular decisions that he or she needs to make.

The reader of an evidence synthesis should also keep in mind that different research questions require different research designs. For example, wherever possible, questions about the effectiveness of a particular intervention strategy are best answered using a randomized controlled trial (RCT). Most Cochrane Collaboration reviews and evidence syntheses are of this sort, and many such reviews exclude observational studies. However, designs other than RCTs are better suited to answering other important clinical questions. In particular, RCTs are rarely the design of choice to assess diagnostic techniques and have little to offer us when studying the usual course and prognosis of a health condition. The types of evidence needed to develop an evidence synthesis and the methods employed to synthesize the information once the evidence is assembled are highly dependent on the particular question being asked.

Another important source of diversity in evidence syntheses relates to the methodological quality of that evidence. Some synthesis procedures, but not all, evaluate the methodological quality of the studies included. Moreover, some methodological quality assessments use rating scales (and may or may not use these ratings for establishing weights for each study in developing their conclusions). Others use dichotomous criteria by which only those studies judged to have adequate methodological soundness are included in the synthesis. Still others use combinations of these strategies. An important consideration is that the conclusions in the synthesis have been shown to differ depending on which method is used to appraise the quality of the evidence. Even in peer-reviewed publications, the methodological quality of studies varies widely. Generally, an evidence synthesis that does not consider the methodological quality of the evidence it combines should be viewed with caution. The evidence synthesis should alert the reader to the strength of the findings so that the reader can assess how much confidence he or she should place on the synthesis conclusions and what populations these findings might apply to. As stated previously, a crucial consideration in interpreting evidence from studies is whether the findings from a series of carefully selected participants and settings can be generalized to the wider clinical setting or population.

Different stakeholders also have different needs from a synthesis of evidence. For example, a policy maker and a clinician may have different requirements in an evidence synthesis, since the types of decisions to be made are quite different. The healthcare provider is responsible for considering his or her patient's individual needs and ensuring that his or her patient receives the most appropriate diagnostic and treatment options. A policy maker must consider the needs not only of healthcare patients but also of the community in general. Decision making in this situation considers not only efficacy of clinical practice but also allocation of resources; the general values, standards, and beliefs of the community; and the practicality of incorporating, on a larger scale, what the evidence tells us. This has implications for the type of evidence required for the synthesis, how the synthesis is interpreted, and even the makeup of the group doing the evidence synthesis. The creators of the evidence synthesis should alert the reader as to the target audience of their synthesis; in addition, readers should consider for themselves whether the evidence synthesis is appropriate for their own particular needs.

Combining the Evidence

Once a researcher has addressed these important questions and has a well-formulated question, has clarified the target audience, has adequately searched for evidence, and has appraised the quality of the study (or has decided against such appraisal), he or she comes to the issue of how to perform the actual synthesis of the evidence. What then? How does one combine the evidence? In an ideal world, the researcher would have a substantial number of studies of excellent methodological quality, all with consistent findings. These would provide sufficient variability in the populations studied to assure him or her that the findings can confidently be generalized to the broader population with that particular health problem. Outcomes considered would cover the range of relevant outcomes and would provide the researcher with confidence that the findings are robust. Where this is the case, the evidence is clear, unambiguous, and strong. The researcher knows what the evidence shows and how strong the associations are: The particular strategy used for combining the evidence becomes largely irrelevant.

However, that scenario is the exception for most areas of medical research. More frequently, the researcher has many studies of (at best) moderate-quality evidence, which may or may not contradict each other; one or two studies of adequate methodological quality, which may or may not contradict other studies of poor methodological quality; or many studies of poor methodological quality, which may or may not contradict each other. Even studies examining the same research question, using the same research design, and of similar methodological quality may have widely diverse findings and come to different conclusions. And where the available studies are all methodologically strong, they may lack relevance in real-world settings because all include only highly selected patients, who may not reflect the usual clinical practice.

This is one of the main challenges in developing a useful and valid evidence synthesis. There are

two main ways of combining evidence. These are meta-analysis, whereby findings from the relevant studies are combined statistically to yield an overall direction and size of an effect, and a more narrative, descriptive approach to synthesizing the information from the relevant studies.

A meta-analytic approach is most often seen in evidence syntheses that address issues of intervention effectiveness—for example, many Cochrane Collaboration reviews. To be valid and useful, this approach requires that the studies be reasonably homogeneous, that is, similar with respect to the particular intervention assessed, the population being studied, the context or setting in which the intervention is being assessed, the nature of the outcomes of interest, the measures used to assess those outcomes, and the follow-up time—when the outcomes are assessed. This, of course, requires that the studies to be included report this information in sufficient detail and that the findings (the estimate and variability of the effect) are reported clearly and in a manner that permits statistical combination of these effects. Where these conditions are present, this strategy can overcome some of the limitations of multiple small, underpowered studies that fail to achieve statistical significance because of low sample size. It should be remembered, however, that small RCTs may not only lack statistical power, but they are also at greater risk that random group allocation may have failed to equalize groups, thus introducing confounding. This problem is not necessarily eliminated by pooling studies in a meta-analysis. In addition, where an "overall effect size" is reported, some important explanations for differences among the individual studies might be neglected. For example, differences in patient characteristics among study settings might be responsible for differences in the effectiveness of a particular treatment. If these issues are not explored, the user of the synthesis may miss some important information necessary for successful implementation of the intervention with individual patients.

However, in many cases, there is too much heterogeneity in the studies to justify statistical pooling of study findings. In this case, a qualitative (rather than a quantitative) synthesis of the available evidence must be employed. It is important to distinguish this from a traditional, narrative review. In a traditional, narrative review, the search for studies is neither comprehensive nor systematic, nor is there a systematic critical appraisal performed. Although traditional narrative reviews can be useful sources of information, they are often based on a selected number of studies, which may reflect the biases of the author. Even when a meta-analysis cannot occur because of heterogeneity in the studies, a qualitative synthesis (such as a best-evidence synthesis) based on studies ascertained using a comprehensive and systematic search and a thorough critical review of the studies' methodological soundness can be an important strategy for summarizing the available literature. A sound, informative qualitative analysis of the literature can be a complex and challenging task, since not only similarities but also dissimilarities in studies (e.g., study populations, context and setting, exact nature of the intervention, type and timing of outcomes measured) need to be described and explored as they relate to similarities and differences among study findings.

Where meta-analytic techniques can be employed, they can provide very important information to the clinician and policy maker. Meta-analytic techniques are relatively well standardized and codified. However, meta-analyses rarely explore theoretical or conceptual issues and generally do not address the mechanisms through which the intervention has its effect. A qualitative analysis of the evidence produced by a systematic search and critical review of the literature requires more judgment, and the procedures are less codified. Such approaches lend themselves more easily to an exploration of theoretical issues and of the mechanisms of the intervention, although not all qualitative analyses address these issues.

Whichever approach is used in the evidence synthesis, the reader should have access to a description of each study included in the synthesis, preferably in tabular form so that the studies can be easily compared. At a minimum, this should include a description of the research design used (if more than one research design is included in the synthesis), the study setting, the study sample (source of sample, sample characteristics, number in each group at inception and at follow-up), a summary of the intervention (if more than one is included in the synthesis), the outcomes assessed, their timing and measures used, and the findings (estimates and the variability around those estimates, e.g., confidence intervals). This allows readers to determine for

themselves how closely the study samples and settings relate to their own patient populations and healthcare settings, whether the outcomes being assessed are of relevance in their own particular circumstances, and how much variability there is in the literature.

Finally, it should always be remembered that no matter how the evidence is combined in an evidence synthesis, both the individual studies and the synthesis of these studies report average, not individual, risks and benefits. Whether the decisions being made are policy decisions or clinical decisions, the quality of the decision depends on having access to both good evidence and good judgment.

Linda Jean Carroll

See also Meta-Analysis and Literature Review; Qualitative Methods; Randomized Clinical Trials

Further Readings

Altman, D. G., Schulz, K. F., Moher, D., Egger, M., Davidoff, F., Elbourne, D., et al. (2001). The revised CONSORT statement for reporting randomized trials: Explanation and elaboration. *Journal of the American Medical Association, 134,* 663–694.

Forbes, A., & Griffiths, P. (2002). Methodological strategies for the identification and synthesis of "evidence" to support decision-making in relation to complex healthcare systems and practices. *Nursing Inquiry, 9,* 141–155.

Slavin, R. E. (1995). Best evidence synthesis: An intelligent alternative to meta-analysis. *Journal of Clinical Epidemiology, 48,* 9–18.

Starr, M., & Chalmers, I. (2003). *The evolution of the Cochrane library, 1988–2003.* Oxford, UK: Update Software. Retrieved February 2, 2009, from http://www.update-software.com/history/clibhist.htm

Van der Velde, G., van Tulder, M., Côté, P., Hogg-Johnson, S., Aker, P., Cassidy, J. D., et al. The sensitivity of review results to methods used to appraise and incorporate trial quality into data synthesis. *Spine, 32,* 796–806.

Expected Utility Theory

Expected utility theory (EUT) states how an EUT decision maker makes choices among options that have specified characteristics. Each option is in some sense viewed by the EUT decision maker as beneficial to the EUT decision maker, but the option also has risks associated with the benefits and the EUT decision maker must bear the adverse outcomes associated with these risks should they occur. In addition, both the benefits and the risks of the options are uncertain, hence EUT decision makers must consider a set of uncertainties of benefits and risks among options (or alternatives) as they make their way through the decisions they face. Compared with other types of decision makers who pursue different routes in coming to a choice among alternatives with trade-offs, the EUT decision maker makes his or her choice in one way: by comparing the weighted sums of the options that are open to him or her. The weighted sums of options are obtained by adding the utility values of each of the outcomes multiplied by each outcome's respective probability of occurrence across the set of outcomes open to the EUT decision maker.

The origins of the EUT can be traced back to 1738, when Daniel Bernoulli wrote what he described as a new theory of the measurement of risk. But what assumptions was Bernoulli coming up against that required a "new" formulation?

Floris Heukelom traces the history of the mathematics of rational behavior to 1654, when Chevalier de Méré instigated Blaise Pascal, and therewith Pierre Fermat, to consider gambling problems. Heukelom notes that from an examination of a large body of literature on Enlightenment mathematicians who were interested in probability, it seemed as if these mathematicians were not making a real distinction between the determination of what they considered to be an answer to the question "What should the rational solution to the problem be in situations of uncertainty?" and the question "What would a rational person actually do (or how would a rational person act) in those same situations of uncertainty?" For these mathematicians, the two questions were one and the same.

One such construction of a gamble is the St. Petersburg game that came under the scrutiny of Bernoulli. Chris Starmer notes the following about EUT as it was first proposed by Bernoulli to the St. Petersburg game. Starmer notes that Bernoulli proposed EUT in response to an apparent puzzle surrounding what price a reasonable

person should be prepared to pay to enter a gamble. It was the conventional wisdom at the time that it would be reasonable to pay anything up to the expected value of a gamble. But Bernoulli proposed making a game out of flipping a coin repeatedly until a tail is produced, and let us make a game of this situation. The game rules are as follows: If one is willing to participate in the game, one will receive a payoff of, say, $\$2^n$, where n is the number of the throw producing the first tail (T). If one goes about looking for players for this game, one finds that people do not want to get involved in this game, where, in fact, the expected monetary payoff is infinite. In fact, and to the surprise of theoretical mathematicians, people are only prepared to pay a relatively small amount to even enter the game. Bernoulli argued that the "value" of such a gamble to an individual is not, in general, equal to its expected monetary value as theoretical mathematicians believe. Rather, Bernoulli argued and proposed a theory in which individuals place subjective values, or *utilities*, on monetary outcomes. Here, for Bernoulli, the value of a gamble is the expectation of these utilities.

Heukelom notes that instead of the "objective value of the monetary gain" being taken as the expectation in people, the "subjective value of the utility" should be taken as the mathematical expectation of a game or gamble. Here, when considering the subjective value of the utility, the St. Petersburg paradox does not go to infinity but, depending on the exact parameters of Bernoulli's equation, will asymptotically go to a number that is in fact quite reasonable.

Bernoulli's Theory of the Measurement of Risk

Bernoulli's first paragraph of his formulation of the St. Petersburg game, translated from Latin into English by Louise Sommer, notes that ever since mathematicians first began to study the measurement of risk, there has been general agreement on the proposition that "expected values" are computed by multiplying each possible gain by the number of ways in which it can occur and then dividing the sum of these products by the total number of possible cases where, on this theory, the consideration of cases that are all of the same probability is insisted on.

Bernoulli then notes that the proper examination of the "numerous demonstrations of this proposition" all rest on one hypothesis: Since there is no reason to assume that of two persons encountering identical risks, either should expect to have his or her desires more closely fulfilled, the risks anticipated by each must be deemed equal in value.

Bernoulli then focuses in on the term *value* above and argues that the determination of the *value* of an item must not be based on its *price* but rather on the *utility* it yields. The price of the item depends only on the thing itself and is equal for everyone; the utility, however, depends on the particular circumstances of the person making the estimate. Bernoulli concluded by making explicit the point that there is no doubt that a gain of 1,000 ducats is more significant to a poor person than to a rich person, although both the poor person and the rich person gain the same amount.

For Bernoulli, what becomes evident is that no valid measurement of the value of a risk can be obtained without consideration being first given to its utility, that is, the utility of whatever gain accrues to the individual or how much profit is required to yield a given utility.

Exceedingly Rare Exceptions

Bernoulli, however, was quick to recognize that he needed to consider the case of what usually happens and not place his focus solely on the case of exceedingly rare exceptions. The exceedingly rare exception referred to was the case of a prisoner. For although a poor person generally obtains more utility than does a rich person from an equal gain, it is nevertheless conceivable, for example, that a rich prisoner who possesses 2,000 ducats but needs 2,000 ducats more to repurchase his freedom will place a higher value on a gain of 2,000 ducats than does another person who has less money than he.

Bernoulli's Risk Aversion

Excluding these rare exceptions, Bernoulli argued that we should consider what usually happens and assume that there is an imperceptibly small growth in the individual's wealth, which proceeds continuously by infinitesimal increments. For Bernoulli, it is highly probable that *any increase in*

wealth, no matter how insignificant, will always result in an increase in utility, which is inversely proportionate to the quantity of goods already possessed.

Daniel Kahneman notes that Bernoulli suggested that people do not evaluate prospects by the expectation of their monetary outcomes but rather by the expectation of the subjective value of these outcomes. This subjective value of a gamble is again a weighted average, but now it is the subjective value of each outcome that is weighted by its probability. Kahneman then argues that to explain "risk aversion" within this framework, Bernoulli had to propose that subjective value, or utility, is a concave function of money. Hence, the concavity of the utility function entails a risk-averse preference for the sure gain over a gamble of the same expected value, although the two prospects have the same monetary expectation.

Commentary on Bernoulli's Work

Heukelom gives Bernoulli credit for successfully introducing a theory of maximizing expected utility (EUT) as the basis for the study of rational decision behavior under uncertainty and adds that—as anachronistic as it may seem—what is being seen in these discussions is the beginnings of today's decision theory.

Writing on the early history of experimental economics, Alvin E. Roth considers Bernoulli's work on the St. Petersburg paradox as perhaps the best candidate for the *first* economic experiment. Roth is referring here to Bernoulli's paper "Exposition of a New Theory on the Measurement of Risk." For Roth, Bernoulli did not simply rely on and attempt to publish only his own intuitions but rather adopted the practice of asking other famous scholars for their opinions on difficult choice problems. Here, Bernoulli is argued by Roth to be using a similar information report methodology to what is now being used in the hypothetical choice problems that generate hypotheses about individual choice behaviors today, and furthermore, it can be argued that this can be seen as a continuum from Bernoulli's work to the work of theorists of individual choice behavior in cognitive psychology today.

In this history of experimental economics, Roth gives credit to L. L. Thurstone's 1931 experiment on individual choice and the problem of experimentally determining an individual's indifference curves. Here, Roth argues that Thurstone was concerned with testing the indifference curve representation of preferences and with the practicality of obtaining consistent choice data of the sort needed to estimate these indifference curves.

Kahneman also traces the psychophysical approach to decision making to this essay by Bernoulli on risk measurement.

Starmer considers Bernoulli's theory the first statement of EUT with his solution to the St. Petersburg puzzle but asserts that modern economists in the 1950s only discovered and built on Bernoulli's insight. Here, Starmer argues that a possible explanation for this time delay in theory development is at least partly explained by the fact that the form of Bernoulli's theory presupposes the existence of a cardinal utility scale. And this assumption about cardinal utilities did not sit well with the more modern theorists' drive toward ordinalization in the first half of the 20th century. John von Neumann and Oskar Morgenstern revived interest in Bernoulli's approach and showed that the expected utility hypothesis could be derived from a set of apparently appealing axioms on preference.

Paul J. H. Schoemaker notes that $U(x)$, the utility function proposed by Bernoulli, was logarithmic and thus exhibited diminishing increases in utility for equal increments in wealth. However, Schoemaker notes that Bernoulli did not explicitly address the issue of how to measure utility, nor did Bernoulli address why his expectation principle should be considered as rational. Without such further exploration, Shoemaker argues that Bernoulli's theory may only be interpreted as a "descriptive model" by some commentators, even though the expectation principle at the time may have enjoyed face validity as a "normative model."

Medical Decision Making

Today, in the area of medical decision making, the following questions regarding work like Bernoulli's on risk are still being asked: Is expected utility theory supposed to describe individuals' choices? Is it supposed to be prescriptive for medical decision making? Is it supposed to be

normative for medical decision making? Is it normative or just simply practical? These challenges to expected value theory's strengths and further defining of expected values theory's weaknesses have followed expected value theory since its early formulation by Bernoulli, and these perspectives continue to challenge ethicists, researchers, and theorists in medical decision making, economics, and EUT today.

Dennis J. Mazur

See also Nonexpected Utility Theories; Risk Aversion; Subjective Expected Utility Theory

Further Readings

Bernoulli, D. (1954). Exposition of a new theory on the measurement of risk (Trans. L. Sommer). *Econometrica, 22,* 23–26. (Original work published 1738)

Cohen, B. J. (1996). Is expected utility theory normative for medical decision making? *Medical Decision Making, 16,* 1–6.

Heukelom, F. (2007). *Kahneman and Tversky and the origin of behavioral economics* (Tinbergen Institute Discussion Paper No. 07-003/1). Retrieved February 2, 2009, from the SSRN Web site: http://ssrn.com/abstract=956887

Kahneman, D. (2007). Preface. In D. Kahneman & A. Tversky (Eds.), *Choices, values, and frames* (pp. ix–xvii). New York: Cambridge University Press.

Roth, A. E. (1993). On the early history of experimental economics. *Journal of the History of Economic Thought, 15,* 184–209.

Samuelson, P. A. (1977). St. Petersburg paradoxes: Defanged, dissected, and historically described. *Journal of Economic Literature, 15,* 24–55.

Schoemaker, P. J. H. (1982). The expected utility model: Its variants, purposes, evidence, and limitations. *Journal of Economic Literature, 20,* 529–563.

Starmer, C. (2000). Developments in non-expected utility theory: The hunt for a descriptive theory of choice under risk. *Journal of Economic Literature, 38,* 332–382.

Thurstone, L. L. (1931). The indifference function. *Journal of Social Psychology, 2,* 139–167.

von Neumann, J., & Morgenstern, O. (1944). *Theory of games and economic behavior.* Princeton, NJ: Princeton University Press.

Expected Value of Perfect Information

Simply basing decisions on expected cost-effectiveness or, equivalently, net health or monetary benefit will ignore the question of whether the current evidence is a sufficient basis for adopting or reimbursing a health technology. It would fail to address the question of whether further research is needed to support such a decision in the future. The value of evidence or the health costs of uncertainty can be illustrated using a simple example as shown in Table 1. Each row represents a realization of uncertainty, that is, the net health benefit (commonly measured in quality-adjusted life years, or QALYs) that results when all the parameters that determine expected costs and effects each take one of their many possible values. These realizations may be generated by probabilistic sensitivity analysis, which commonly randomly samples (Monte Carlo simulation) from each of the distributions assigned to parameters. Therefore, each row can be thought of as representing one of the ways things could turn out given our current uncertainty. The expected net benefit for Treatments A and B is the average over all these possibilities (in this example, the range of potential values is simplified to only five possibilities).

On the basis of current evidence, we would conclude that Treatment B was cost-effective, and on average we expect to gain an additional 1 QALY per patient treated compared with Treatment A. However, this decision is uncertain, and Treatment B is not always the best choice (only 3 times out of 5), so the probability that B is cost effective is .6. For some realizations (2 out of 5), Treatment A would have been the better choice. Therefore, a decision to adopt B based on current evidence is associated with an error probability of .4. This is substantially greater than the traditional benchmarks of statistical significance, such as .05. But whether or not this level of uncertainty "matters" depends on the consequences, that is, what improvement in net benefit (or avoidance of harm) could have been achieved if this uncertainty had been resolved.

The decision maker is faced with three choices: (1) adopt Technology B based on current evidence, (2) adopt the technology now but conduct further

Table 1 Expected value of perfect information

How Things Could Turn Out	*Net Health Benefit*			*Best We Could Do if We Knew*
	Treatment A	*Treatment B*	*Best Choice*	
Possibility 1	9	12	B	12
Possibility 2	12	10	A	12
Possibility 3	14	17	B	17
Possibility 4	11	10	A	11
Possibility 5	14	16	B	16
Average	12	13		13.6

research so that this initial decision can be reconsidered once the new evidence is available, or (3) withhold approval until further research resolves some of the uncertainty. Therefore, some assessment of whether uncertainty matters and of the value of additional evidence is required.

For example, if uncertainty could be completely resolved, that is, through complete evidence or perfect information about effect and cost, then we would know the true value of net health benefit before choosing between A and B. Therefore, with perfect information, we should be able to adopt whichever technology provided the maximum net benefit for each realization of uncertainty (the fifth column in Table 1). Of course, we can't know in advance which of these values will be realized, but on average (over the fifth column) we would achieve 13.6 rather than 13 QALYs—a gain of .6 QALYs. It should be clear that the cost of uncertainty or the value of evidence is just as "real" as access to a cost-effective treatment, as both are measured in terms of improved health outcomes for patients. In principle, evidence can be just as, or even more important than, access to a cost-effective technology. In this case, the expected value of perfect information is .6 QALYs, which is more than half the value of the technology itself, that is, 1 QALY gained by adopting B.

Additional evidence can be used to guide the treatment of all other current and future patients. Therefore, the maximum value of evidence to the healthcare system as a whole requires estimates of this current and future patient population (where the population expected value of perfect information [EVPI] is the discounted sum). This requires a judgment to be made about the time over which additional evidence that can be acquired in the near future is likely to be useful and relevant. Generally, fixed time horizons of 10, 15, and 20 years have commonly been used in the health literature as well as the environmental risk and engineering literature. There is some empirical evidence that suggests that clinical information may be valuable for much longer (a half-life of 45 years). However, any fixed time horizon is really a proxy for a complex and uncertain process of future changes, all of which affect cost-effectiveness and the future value of evidence. In health, some future changes can be anticipated (a new technology will be launched, a trial that is recruiting will report, or a branded drug will go generic), and differing judgments about time horizons in different contexts might be appropriate.

As well as a simple metric of the relative importance of uncertainty across different clinical decision problems, the population EVPI can be expressed as net monetary benefit and compared with the expected cost of additional research, which includes the net benefit forgone if conducting research requires delaying approval of a technology that appears to be cost-effective based on current evidence. If these expected opportunity costs of research exceed the population EVPI (maximum benefits), then the research is not

worthwhile—the resources could generate more health improvement by being used elsewhere in the healthcare system, and coverage should be based on current estimates of expected cost-effectiveness. Therefore, EVPI provides a necessary condition for conducting further research and a means to start to prioritize the allocation of research and development resources within the healthcare system.

Expected Value of Perfect Partial (Parameter) Information

If further research is potentially worthwhile (EVPI exceeds the expected cost of research), it would be useful to have an indication of what type of additional evidence might be most valuable. This can inform the decision of whether approval should be withheld until the addition research is conducted or whether a "coverage with evidence development" would be appropriate.

The analysis of the value of information associated with different (groups of) parameters is, in principle, conducted in a very similar way to the EVPI for the decision as a whole. The expected value of perfect information for a parameter or group of parameters (EVPPI) is simply the difference between the expected net benefit when their uncertainty is resolved (and a different decision can be made) and the expected net benefit given the existing uncertainty.

EVPPIs can be used as a simple metric of the relative importance (sensitivity) of different types of parameters and sources of uncertainty in contributing to the overall EVPI. As a simple measure of sensitivity, it has a number of advantages: (1) It combines both the importance of the parameter (how strongly it is related to differences in net benefit) and its uncertainty; (2) it is directly related to whether the uncertainty matters (whether the decision changes for different possible values); and (3) it does not require a linear relationship between inputs and outputs. In addition, it can be expressed in health or money values and either per patient or for the population of current and future patients.

When population EVPPI is expressed in monetary terms, it can be directly compared with the expected opportunity costs of the type of research that might be needed to provide the evidence. This is important as some uncertainties are relatively cheap to resolve (in terms of time and resource) compared with others (e.g., an observational study to link a clinical end point to quality of life compared with a randomized clinical trial of long-term relative treatment effect). Which source of uncertainty is most important requires a comparison of these benefits and opportunity costs.

Evaluating EVPPI often comes at a computational cost. For linear models, each estimate of EVPPI requires some additional computation (the manipulation of the simulated values rather than repeated simulations). When the model itself is not computationally expensive, it is a generally manageable expense. However, if the model is nonlinear, EVPPI may require many repeated runs of the same probabilistic model, which can become prohibitively expensive. Therefore, the computational expense of EVPPI needs to be justified, that is, if the analysis of population EVPI suggests that additional evidence might be required. It is also more efficient and more informative to first consider a limited number of groups of parameters, informed by the types of research required, for example, randomized clinical trial, survey of QALYs, or an observational epidemiological study. If there is substantial EVPPI associated with a particular group, only then conduct additional analysis to explore which particular source of uncertainty within the group matters the most.

Expected Value of Sample Information

The EVPI and EVPPI place an upper bound on the returns to further research so can only provide a necessary condition for conducting further research. To establish a sufficient condition, to decide if further research will be worthwhile and identify efficient research design, estimates of the expected benefits and the cost of sample information are required.

The same framework of EVPI analysis can be extended to establish the expected value of the sample rather than perfect information. For example, a sample from a particular type of study that provides information about some or all parameters will generate a sample result that can be used to update the parameter estimates and recalculate net benefits of the alternative

treatments. Once the result of this sample is known, then the decision maker would choose the alternative with the maximum expected net benefit when those expected net benefits are averaged over the posterior distribution (the combination of sample result and prior information). Of course, there are many possible results that might occur, so the range of possible sample results from the sample must be evaluated, that is, similar to the realizations in Table 1 but now realization of the sample results rather than uncertainty itself. Which particular sample result will occur should the sample be taken is unknown, so the expected value of a decision taken with the sample information is the average over all the possible predicted results and predicted posteriors, that is, similar to averaging over Column 5 in Table 1. The difference between the expected net benefits with sample information and expect net benefit with current information is the expected value of sample information (EVSI).

This type of calculation would provide the EVSI for a single study design and only one sample size. To establish the optimal sample size for this particular type of study, these calculations would need to be repeated for a range of possible sample sizes. The difference between the EVSI for the population of current and future patients and the costs of acquiring the sample information (Cs), which should include both resource and opportunity costs, is the expected net benefit of sample information (ENBS) or the societal payoff to proposed research. The optimal sample size for a particular type of study is simply the sample size that generates the maximum ENBS. This is illustrated in Figure 1 and shows how the EVSI will increase with sample size but at a declining rate (it will approach the relevant EVPI or EVPPI in the limit). In this case, the costs of sampling increase at a constant rate and the ENBS reached a maximum at $n = 1{,}100$.

There are usually a number of different ways in which a particular type of study could be designed. For example, a randomized clinical trial can be designed to collect information on limited clinical end points or include quality of life and costs. A range of follow-up periods is also possible, providing information on either short- or long-term effects. Patients recruited to the trial can also be allocated in different ways to the different arms. The efficient design of a

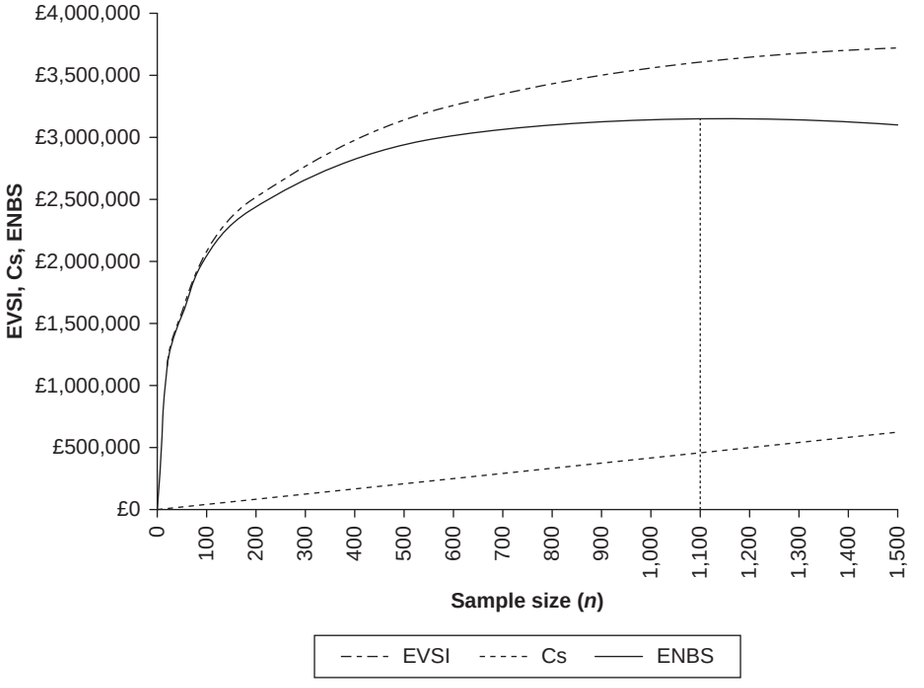


Figure 1 Expected value of sample information (EVSI), costs of acquiring the sample information (Cs), and expected net benefit of sample information (ENBS)

particular type of study will be one that provides the maximum ENBS. However, in most decision problems, a range of different types of study can be conducted at the same time to provide information about different types of parameters, for example, a randomized clinical trial to inform relative effect, a survey of the quality of life associated with a clinical end point, and an epidemiological study to inform other events. The problem is now to evaluate each possible portfolio of research, including the optimal allocation of sample (patients) to these different types of study. Of course, these dimensions of design space become even larger once the sequence in which studies might be conducted is considered. In principle, a measure of societal payoff to research provides a means to explore this design space and identify efficient research design and optimal portfolios of research.

Karl Claxton

See also Decision Trees: Sensitivity Analysis, Basic and Probabilistic; Expected Value of Sample Information, Net Benefit of Sampling; Managing Variability and Uncertainty; Net Benefit Regression

Further Readings

Ades, A. E., Lu, G., & Claxton, K. (2004). Expected value of sample information in medical decision modelling. *Medical Decision Making, 24*(2), 207–227.

Briggs, A., Claxton, K., & Sculpher, M. J. (2006). *Decision analytic modelling for the evaluation of health technologies*. Oxford, UK: Oxford University Press.

Claxton, K. (1999). The irrelevance of inference: A decision making approach to the stochastic evaluation of health care technologies. *Journal of Health Economics, 18,* 341–364.

Claxton, K., & Sculpher, M. J. (2006). Using value of information analysis to prioritise health research: Some lessons from recent UK experience. *Pharmacoeconomics, 24,* 1055–1068.

Pratt, J., Raiffa, H., & Schlaifer, R. (1995). *Statistical decision theory*. Cambridge: MIT Press.

Yokota, F., & Thompson, K. M. (2004). Value of information literature analysis: A review of applications in health risk management. *Medical Decision Making, 24,* 287–298.

Expected Value of Sample Information, Net Benefit of Sampling

Information has a value in utility terms. Consider a diagnostic test. It provides *data*, which, *when duly interpreted*, become *information* that may allow treatment to be individualized and expected outcome utility to increase. Three qualifications include the following:

First, a test may be too uninformative to influence treatment: VOI (*value of information*) = 0. However, expected utility cannot decrease: VOI is never negative.

Second, these are average statements. New diagnostic tests, even when correctly interpreted, will often cause outcome utility to decrease for *some* patients. When population screening is introduced, false positives pay a price.

Third, *misinformation* does carry negative value. Utility may suffer when decisions rest on biased research or when diagnostic test results are wrongly interpreted, for example, due to overly optimistic ideas concerning sensitivity or specificity.

In this clinical, single-case illustration, the *expected value of (perfect) test information* is the expected utility gained by a (perfect) diagnostic or therapy-guiding test. Analogous concepts find application in the collection of data to inform clinical policies. Complete elimination of uncertainty or biased opinions by means of properly conducted research offers a benefit, called the *expected value of perfect information* (EVPI), which ideally should outweigh research costs. Once again, however, *some* may pay a price. Suppose a vigorously promoted new drug is proven dangerous. However welcome this result may be—misinformation carries a negative value!—the result may deprive an unrecognized minority of the only drug that would save their lives.

A Pared-Down Example

Consider patients with a complaint that may signal a special endocrine disorder. The composition of the case stream is known (Table 1a), except

that the sensitivity (Se) of a relevant imaging test is uncertain: It may be .60 or .80, giving rise to the question-marked numbers. The specificity (Sp) is .90.

Decisions have good and bad consequences. Here, we focus on human costs and, more specifically, on regret, that is, the "cost" of not treating the patient as one would were his or her condition fully known. As PVneg is high anyhow, the test negatives will always be treated by the wait-and-see policy, and either 16 or only 8 false negatives (out of 1,000 patients) will incur a regret *B* associated with a delayed clarification of their condition. There are 96 false positives that pay *C* units; this is the human cost of the invasive tests they must undergo. Obviously, this leaves two promising policies: W = wait and see (no need to test) or F = follow the test's advice.

Experts and ex-patients reach a consensus on *C* and *B*: *C* = 1 month (= 1/12 quality-adjusted life year, or QALY), *B* = 3.5 months. As an unfortunate result, the optimal policy depends on the unknown Se (see Table 1b): If Se is only .60, PVpos is too low to have any consequences, and W is optimal (as its cost of 140 is less than 152). If Se = .80, F is optimal (as 124 < 140).

Now assume that the endocrinologists after studying the literature decide that the two values for Se are equally likely: This gives rise to an a priori mean number of (16 + 8)/2 = 12 false negatives (see the third row of Table 1b marked "F (average)"), so F beats W with a narrow margin of 2 months (= 140 – 138).

This was an assessment based on a priori hunches. What is the expected value of perfect information about Se? With prior probability .5, the Se proves to be .60, causing a change of policy from F to W for a mean regret of 140; with probability .5, the Se proves to be .80, in which case one sticks to F for a mean regret of 124. This *pre-posterior assessment* (we are guessing the situation that will prevail after obtaining some hoped-for information) leads to an average of (140 + 124)/2 = 132. That is, the EVPI = 138 – 132 = 6 months per 1,000 cases.

The situation, one may say, involves 6 units of uncertainty due to "cross-case" uncertainty concerning the performance of a clinical tool—and 132 (either 140 or 124, or the average of the two)

Table 1a Composition of case stream (1,000 cases with prompting complaint)

Image Test	*If Se Equals*	*Diseased*	*Healthy*	*Total*
Positive	.60	(Se) × 40 = 24?	96	120? PVpos = .20
	.80	32?		128? PVpos = .25
Negative	.60	(1 – Se) × 40 = 16?	864 (as Sp = .90)	880? PVneg > .98
	.80	8?		872? PVneg > .98
Total	Total	40	960	1,000

Table 1b Choosing a management policy (Wait and see vs. Follow test result)

Policy	*Se*	*Regret (per 1,000 Cases)*	*Regret (per 1,000 Cases) With Consensus Values for B and C (Low Value Desirable)*
W	.60 or .80	40*B*	140 months
F	.60	16*B* + 96*C*?	56 + 96 = 152 months?
	.80	8*B* + 96*C*?	28 + 96 = 124 months?
F	(Average)	12*B* + 96*C*	42 + 96 = 138 months

units of uncertainty that can only be eradicated by introducing a perfect diagnostic test, not by learning more about the present diagnostic tool ("intracase" uncertainty due to an imperfect tool). The 132 units constitute an immutable dead weight in the calculations.

One starts out with either $C = 1$ or $B = 3.5$ months at stake per patient and ends up concluding that improved knowledge of the situation can only save 6/1,000 month, or 4 hours, per patient. Clearly, this is because everything is known apart from a minor uncertainty as to the Se. Ceteris paribus, Se will therefore have low research priority. Perhaps one should rather try to eat into the 132 months load by perfecting the imaging procedure.

Expected Value of (Research) Sample Information (EVSI)

Suppose a patient presents with verified disease. Hoping to benefit future patients, we seize the opportunity to obtain an improved Se estimate by testing this sample of $n = 1$ case.

If the patient proves test positive, it only reinforces the high-Se alternative and hence the prior decision in favor of F. Actually, the expected regret drops from its prior value of 138 to 136; this happens with prior probability .7 (the calculations in Table 1c). With probability .3, the patient will be test negative. This is the interesting case because it favors the low-Se alternative and hence W. Will it prompt a policy change to W? Yes. If the negative result materializes, the expected regret on Policy F increases (deteriorates) from its prior value of 138 to 142.67 (Table 1c), which is >140, so W now beats F.

Overall, what does the prior distribution predict the world to look like once the patient has been tested (pre-posterior assessment)? Foreseen is an expected regret of .7 × (updated consequences of F) + .3 × (consequences of W (which happen to be unchanged)) = .7 × 136 + .3 × 140 = 137.2, and EVSI is the gain from the prior level of 138 to 137.2, that is, .8 month.

The EVSI/EVPI fraction is here (.8 month)/(6 months) = .13, so the very first patient with the

Table 1c Updated consequences of Policy F in light of data from a single verified case

Unknown Parameter	*Se = .60*	*Se = .80*	*Probability Sum*
Parameter-dependent mean regret on Policy F	*152*[a]	*124*[a]	. .
Prior probabilities	*.5*	*.5*	1
Prior mean regret	*152 × .5 + 124 × .5 = 138*		
Data = "Test is positive" with probability	.60 × .5 = .3	.80 × .5 = .4	.7[b]
Mean regret if positive	(*152* × .3 + *124* × .4)/.7 = *136*[c]		
Data = "Test is negative" with probability	(1 − .60) × .5 = .2	(1 − .80) × .5 = .1	.3[b]
Mean regret if negative	(*152* × .2 + *124* × .1)/.3 = *142.67*		

a. From Table 1b.

b. Bayes's denominator.

c. Quick route to this result: The odds of high versus low Se change from prior odds {1:1} to posterior odds = (prior odds) × (likelihood ratio) = {1:1} × {.80:.60} = {4:3}; that is, the updated chance that Se is .80 is 4/7. The expected regret associated with Policy F is the correspondingly weighted average of *152* and *124* (months per 1,000 cases).

disease that one gets a chance to study will eliminate 13% of the overall uncertainty.

Theoretical Formulation

Proper management of a prospectively delimited class of cases depends on some unknown parameter(s) θ (for notation, see Table 2), such as Se in the example. Had the true value of θ, θ°, been known, there would be complete elucidation of the decision task; expected utility would attain the best possible level attainable with θ°, and there would be 0 expected regret, because regret, by definition, is utility deficit relative to optimal handling of whatever happens to be the latent truth. The clinician's job is to minimize expected regret.

Before sampling, only the policy maker's prior knowledge is available, and the resulting expected regret is the deficit that perfect knowledge of θ would eliminate and thus constitutes the EVPI. It is 0 only if there is no residual uncertainty as to how to act. As defined in the table, $R(a|\theta)$ is the (expected) regret of action a when $\theta^\circ = \theta$, and $A(\theta)$ is the prior probability that $\theta^\circ = \theta$, so the ensuing optimal policy, $a^*(A)$, is the a that minimizes $\sum_\theta A(\theta)R(a|\theta)$,, and the minimum expected regret thus attained, symbolically $R^*(A)$, is also the EVPI, as just explained. In sum,

$$\begin{aligned}\text{EVPI} = R^*(A) &= \sum_\theta A(\theta)R(a^*(A)|\theta) \\ &= \min_a \left\{\sum_\theta A(\theta)R(a|\theta)\right\},\end{aligned}$$

which is ≥0 because all terms are.

In the example above, policy a may be W or F; and θ, that is, Se, has two equally likely values: $A(\text{Se} = .60) = A(\text{Se} = .80) = .5$. So

$$\begin{aligned}\text{EVPI} = \min\{&.5 \times R(\text{W}|.60) + .5 \times R(\text{W}|.80), \\ &5 \times R(\text{F}|.60) + .5 \times R(\text{F}|.80)\}.\end{aligned}$$

Now, $R(\text{W}|.60)$ is 0 because, if we knew that Se = .60, we could do nothing better than adopt Policy W, whereas F would be an inferior choice. Key figures from Table 1b tell us that its use would entail an unnecessary loss of 152 – 140 months per 1,000

Table 2 Notation and probability model

θ	parameter, or vector of parameters, describing a clinical population
θ°	the true value of θ
$R(\ldots)$	the expected regret associated with . . .
a	available policy options
$R(a\|\theta)$	expected regret, given θ, were option a to be chosen
$a^*(\ldots)$	optimum policy choice based on . . .
$R^*(\ldots)$	the expected regret when . . . is optimally responded to
Standard Bayesian data model	
A	the policy maker's prior distribution of θ
x	observed study data
$Q(x\|\theta)$	probability of observing x given θ, when
	$\text{Prob}_{AQ}(\theta, x) = A(\theta)Q(x\|\theta)$,
	$\text{Prob}_{AQ}(x) = \sum_\theta A(\theta)Q(x\|\theta)$.
B	posterior distribution based on prior A, observation x, and model Q:
	$B(\theta) = \text{Prob}_{AQ}(\theta\|x) = \text{Prob}_{AQ}(\theta, x)/\text{Prob}_{AQ}(x)$

cases. That is, $R(\mathrm{F}|.60) = 152 - 140 = 12$. By analogous arguments, $R(\mathrm{F}|.80) = 0$, while $R(\mathrm{W}|.80) = 140 - 124 = 16$. Substituting these figures, we have

$$\begin{aligned}\mathrm{EVPI} &= \min\{.5 \times 0 + .5 \times 16,\ .5 \times 12 + .5 \times 0\} \\ &= \min\{8, 6\} = 6 \text{ months},\end{aligned}$$

as previously calculated in a more transparent way; and $a^*(A) = \mathrm{F}$ as F beats W with a margin of $8 - 6 = 2$ months, again confirming the original analysis.

Once a data set x is available, the Bayesian policy maker's updated θ distribution $B(\theta)$ can be calculated the standard way (Table 2); the letter B is short for "Based on A, Q, and x." Proceeding as before, the data-conditional best action, $a^*(B)$, and associated expected regret are given by

$$R^*(B) = \sum_{\theta} B(\theta)R(a^*(B)|\theta) = \min_a \left\{\sum_{\theta} B(\theta)R(a|\theta)\right\},$$

This quantity may be larger than $\mathrm{EVPI} = R^*(A)$ because an outlying x may discredit θ° vis-à-vis other θs, but on average, sample information will hold a regret reduction, alias the EVSI:

$$\begin{aligned}\mathrm{EVSI} &= \mathrm{EVPI} - \mathrm{E}_{AQ}\{R^*(B)\} = R^*(A) \\ &\quad - \sum_x \mathrm{Prob}_{AQ}(x)R^*(B).\end{aligned}$$

The right-hand term is ≥0 and reflects the mean uncertainty left after observing the sample, proving EVSI ≤ EVPI.

Note 1. In the example, x took two values (the only patient studied was positive or negative). It was natural to calculate EVSI = .8 as $138 - (.7 \times 136 + .3 \times 140) = 138 - 137.2$, but both terms contain the deadweight of 132 units, so a strict application of the formula above would pass via: $\mathrm{EVSI} = (138 - 132) - (137.2 - 132) = 6 - 5.2 = .8$ units. The deadweight term is innocuous because it involves the "intracase" burden of diagnostic imperfection, represented by the figures 140 and 124 from Table 1b only, which the policy maker cannot change (though he or she may gradually learn which of them applies). Formally, a term $f(\theta)$ that only depends on θ may be added to each $R(a|\theta)$ without affecting optimal actions or EVSI-type regret differences (as both terms change by $\mathrm{E}_A\{f(\theta)\}$).

Note 2. One may dissect the EVSI to prove that it, too, is ≥0:

$$\begin{aligned}\mathrm{EVSI} &= \min_a \left\{\sum_{\theta} A(\theta)R(a|\theta)\right\} - \sum_x \mathrm{Prob}_{AQ}(x)R^*(B). \\ &= \min_a \left\{\sum_x \left\{\sum_{\theta} A(\theta)Q(x|\theta)R(a|\theta)\right\}\right\} \\ &\quad - \sum_x \left\{\min_a \left\{\sum_{\theta} A(\theta)Q(x|\theta)R(a|\theta)\right\}\right\}.\end{aligned}$$

That this difference is ≥0 follows from "the fundamental trick of utility analysis," namely, that a sum of minima is smaller than, or equal to, the minimum of a sum: You save something by being allowed to minimize each term separately.

Note 3. Like the EVPI, the EVSI is *subjective*, as both depend on the point of departure, namely, the policy maker's prior for θ. The EVSI also depends on the design of the empirical study and on sample size(s).

Note 4. With its focus on Bayesian prediction of the situation that may prevail, or *on average will prevail*, once a planned data collection is completed, this is an instance of *pre-posterior analysis*.

Expected Value of (Partial, Alias) Parameter Information

When several clinical parameters are unknown, separate calculations can be made for each parameter or group of parameters, the uncertainty concerning the other parameters being handled as before. The expected value of perfect parameter information (EVPPI) for any one parameter is the EVSI of an imaginary study that reveals the true value of that parameter (without providing further empirical information). It is therefore ≤ the overall EVPI but ≥ the information afforded by real studies (*partial, parameter, EVSI*).

Clearly, a parameter that is inexpensive to investigate and also has a high EVPPI should receive high research priority.

Sample Planning: Expected Net Benefit of Sampling

Given an exchange rate between utilities and research expenses, the design and dimensions of the planned sample can be optimized. When sample

size, n, is the issue, the *expected net benefit of sampling* (ENBS) becomes

$$\text{ENBS}(n) = \text{EVSI}(n) - \text{Cost}(n).$$

A Standard Example

If no research is undertaken, everything is zero. Otherwise, one faces an initial cost, C, and a cost per observation, c. Regrets may be roughly proportional to the squared standard error of the θ estimation and therefore inversely proportional to n, at least for large n. The regret expectation that remains after n observations is then Z/n, where the constant Z subsumes some variance and regret factors. So one gains $\text{EVSI}(n) = \text{EVPI} - Z/n$ by sampling. Combining the elements, one gets

$$\text{ENBS}(n) = [\text{EVPI} - Z/n] - [C + cn] \quad \text{for reasonably large } n.$$

A small study is never profitable because of the initial cost. As n grows beyond limits, costs also become prohibitive. The optimal sample size is $n* = \sqrt{Z/c}$, and sampling is profitable if the resulting $\text{ENBS}(n = n^*)$ is positive, thus beating the no-research option.

Qualitative Policy Selection

When it suffices to document that θ lies right or left of a clinical decision boundary, $\text{EVSI}(n)$ usually approaches EVPI exponentially fast, and the required sample becomes small and less cost dependent than when the actual value of θ matters.

Interpersonal Aspects

Multiple Decision Makers

EVSI (and similar) calculations based on "an average policy maker's prior" may not match a sophisticated analysis that acknowledges differences of prior belief. However, even if rational experts start out with different prior beliefs, sound data collection will eventually bring about numerical agreement on parameters; and prior to that, it will induce a qualitative consensus about patient management policies. Lack of consensus implies regret (when two camps recommend different interventions, they cannot both be right), but a Bayesian formalization of the notion of *value of professional consensus* is difficult.

Ethics

Cool calculi face ethical obstacles. Informed consent is problematic toward the end of a randomized trial, when strict equipoise is impossible to maintain. What kinds of appeal to altruism are justifiable? Can skewed randomization be used in the trade-off between the interests of current and future patients? To benefit the former, "play the winner"; to benefit the latter, maximize VOI, which typically means playing the least explored alternative.

Jørgen Hilden

See also Economics, Health Economics; Expected Value of Perfect Information, Net Benefit Regression; Regret; Subjective Expected Utility Theory

Further Readings

Brennan, A., Kharroubi, S., O'Hagan, A., & Chilcott, J. (2007). Calculating partial expected value of perfect information via Monte Carlo sampling algorithms. *Medical Decision Making, 27,* 448–470.

Claxton, K., Ginnelly, L., Sculpher, M., Philips, Z., & Palmer, S. (2004). A pilot study on the use of decision theory and value of information analysis as part of the NHS Health Technology Assessment programme. *Health Technology Assessment, 8,* 1–103.

Claxton, K., & Posnett, J. (1996). An economic approach to clinical trial design and research priority-setting. *Health Economics, 5,* 513–524.

Eckermann, S., & Willan, A. R. (2008). Time and expected value of sample information wait for no patient. *Value in Health, 11,* 522–526.

Hilden, J., & Habbema, J. D. (1990). The marriage of clinical trials and clinical decision science. *Statistics in Medicine, 9,* 1243–1257.

Philips, Z., Claxton, K., & Palmer, S. (2008). The half-life of truth: What are appropriate time horizons for research decisions? *Medical Decision Making, 28,* 287–299.

Welton, N. J., White, I. R., Lu, G., Higgins, J. P., Hilden, J., & Ades, A. E. (2007). Correction: Interpretation of random effects meta-analysis in decision models. *Medical Decision Making, 27,* 212–214.

Experience and Evaluations

The manner by which individuals evaluate how good or bad it is to be in a health state is central to reaching an informed medical decision. Evidence has shown that personal experience with illness, such as being diagnosed with cancer, leads to a more positive evaluation of that health state than the general public's perception. This disparity has been attributed to a focusing bias on the part of the general public—the tendency to focus too narrowly on a single event, for example, cancer, while forgetting all the other aspects of life that will remain unaffected. One potential means for overcoming such a bias is to ask the public to imagine standing in the shoes of the patient. This perspective-taking exercise might be achieved through exposure to a vicarious illness experience, though further research is needed to test this hypothesis.

Personal Illness Experience

Researchers have consistently found that the general public gives lower evaluations of a particular health state, such as having chemotherapy to treat cancer, compared with individuals who have had personal experience with that health state. This has been described as the distinction between predicted utility, people's predictions about what they think chemotherapy *would* be like (i.e., unimaginably horrible), versus experienced utility, how the experience of chemotherapy actually *is* like for cancer patients (i.e., not as bad as they expected).

Discrepancy Between Patients' and Public's Evaluations

In trying to understand how health state evaluations are affected by personal experience (or the lack thereof), researchers seem to have converged on a single explanation: focusing bias. This is the tendency for the general public to focus too much on a particular event (i.e., the cancer diagnosis) and not enough on the consequences of other new and ongoing future events that will compete for one's attention. For example, the general public may evaluate health states as worse than patients do because the general public focuses too narrowly on the (a) illness, forgetting that other facets of life will be unaffected; (b) immediate loss of health, forgetting patients' ability to adapt; (c) intense negative emotions aroused by the diagnosis, forgetting that extreme emotions tend to dissipate over time; and so on.

If the general public's inability to predict the effect of illness is due to focusing too narrowly, the question then becomes "What can broaden this narrow perspective individuals bring to the medical decision-making process when they have no personal experience?"

Vicarious Illness Experience

To broaden the general public's perceptions, they could be asked to imagine what it is like to live with a long-term, chronic illness. One means for achieving this perspective-taking task could be through exposure to a second type of illness experience: the vicarious experience (VE) of illness. For clarity, it is necessary to define the terminology used here. Firsthand personal experience is when A has been diagnosed with cancer; secondhand experience is when A tells B about his cancer diagnosis; and thirdhand experience is when B tells a third party, C, about A's cancer. Of course, one may have multiple types of experiences simultaneously, as when a man's father is diagnosed with cancer. The son has his own experience of being with his father while he is treated (firsthand) and also hears from his father what the experience of being diagnosed with and undergoing treatment for cancer was like for him (secondhand). Here, VE is defined as secondhand, being directly told about another's experience.

Why VE? When patients are newly diagnosed with cancer, they are faced with decisions about health states they typically have no real understanding of. Therefore, many actively seek out others with expertise, particularly former cancer patients. When former patients vicariously share their experiences, they may help newly diagnosed patients (a) broaden their focus by stepping back from their immediate, narrow fears and, consequently, (b) develop more informed expectations of how treatment will (and will not) change their lives, but this proposition has not yet been tested.

Theoretically, VE could have a positive impact because it provides information typically unavailable to individuals for two reasons. First, from an

information-processing perspective, learning from VE is rational and adaptive for events that are rare but of high consequence, such as cancer, because direct experience may be fatal. It is not adaptive to have to wait until one has a cancer scare to learn the importance of, and be motivated to undergo, screening for cancer.

Second, real-world personal experiences are idiosyncratic and asymmetric in nature. Individuals only learn about the outcomes of the particular choices they make. They get no information, and therefore learn nothing, from the alternatives they did not choose. If they develop false beliefs based on these experiences, such as the belief that cancer treatment is useless, these false beliefs cannot be disconfirmed if they do not change their behavior and experience different outcomes. However, they *can* learn from the experience of others that treatment may increase the chances of survival.

This is not to imply VE is always beneficial. As with anything, if implemented poorly or if inaccurate information is conveyed, it can have suboptimal results. Accordingly, one may also learn from others that cancer treatment does not lead to survival. A poignant example exists in African American communities, where many believe that cancer is a death sentence. Because of the fear and stigma surrounding cancer, neither cancer patients nor survivors feel free to discuss their experiences. Therefore, the VE most individuals have is attending the funerals of those who have died from cancer.

Real-World Example

There is a real-world experiment that provided evidence that being exposed to a positive VE could both (a) improve noncompliant individuals' evaluations of an invasive and uncomfortable cancer-screening test and (b) motivate them to undergo screening. One of the most efficacious and least used cancer-screening tests is colonoscopy to detect and treat colorectal cancer. In March 2000, the NBC anchor Katie Couric underwent a live, on-air colonoscopy on the *Today* show to screen for colon cancer, a cancer that had led to the death of her husband. Researchers compared colonoscopy utilization rates before and after this powerful VE. They found that colonoscopy rates significantly increased after Couric's program, whereas there was no concomitant increase in other cancer-screening tests.

Vicarious Illness Experience Remains Poorly Understood

To an extent, the gap in our knowledge about VE reflects the fact that much experimental research in psychology has focused on intraindividual factors. Therefore, it has been necessary to experimentally control potentially confounding factors, such as the influence of others' experiences. Further research is needed to draw a more complete picture of the role that personal experience and VE play in the evaluation of health states in medical decision making.

Julie Goldberg

See also Biases in Human Prediction; Cognitive Psychology and Processes; Construction of Values; Context Effects; Decision Making in Advanced Disease; Decision Psychology; Expected Utility Theory; Health Outcomes Assessment; Hedonic Prediction and Relativism; Judgment; Managing Variability and Uncertainty; Subjective Expected Utility Theory

Further Readings

Cram, P., Fendrick, A. M., Inadomi, J., Cowen, M. E., Carpenter, D., & Vijan, S. (2003). The impact of a celebrity promotional campaign on the use of colon cancer screening: The Katie Couric effect. *Archives of Internal Medicine, 163*(13), 1601–1605.

Hagen, K., Gutkin, T., Wilson, C., & Oats, R. (1998). Using vicarious experience and verbal persuasion to enhance self-efficacy in pre-service teachers: "Priming the pump" for consultation. *School Psychology Quarterly, 13*(2), 169–178.

Llewellyn-Thomas, H. A., Sutherland, H. J., & Thiel, E. C. (1993). Do patients' evaluations of a future health state change when they actually enter that state? *Medical Care, 31*(11), 1002–1012.

Ubel, P. A., Loewenstein, G., Hershey, J., Baron, J., Mohr, T., Asch, D. A., et al. (2001). Do nonpatients underestimate the quality of life associated with chronic health conditions because of a focusing illusion? *Medical Decision Making, 21*(3), 190–199.

Wilson, T. D., Wheatley, T., Meyers, J. M., Gilbert, D. T., & Axsom, D. (2000). Focalism: A source of durability bias in affective forecasting. *Journal of Personality and Social Psychology, 78*(5), 821–836.

Experimental Designs

As in other branches of science, the time-honored method of research in the realm of medicine is one factor at a time. This practice of minimizing or eliminating changes in all factors of interest and then, one by one, changing the levels of each factor and recording the responses to those changes has been and continues to be used for the simple reason that it works and because most researchers do not realize that better methods exist. From the standpoint of efficiency with respect to time, money, effort, and quality of results, one-factor-at-a-time research is a failure.

A factor is any variable whose changes might result in responses of interest to an investigator. Factors include, but are not limited to, things such as dosage levels of one or more medicines, exercise regimens, types of sutures, mechanical properties of prosthetic devices, and material compositions of any medically implanted item or device.

The most efficient method for investigating the effects of variables over which an investigator has a degree of control is that of experimental design. The first work on experimental designs was done by R. A. Fisher at the Rothamsted Experimental Station in Hertfordshire, England, in the early 1920s. Work on the development of new designs and methods for their analysis continues to the present day.

To provide a basic understanding of the concepts of experimental designs, the discussion will be limited to the most elementary types of design, where the factors are limited to two levels, and the discussion will focus only on the assessment of single-factor effects.

For the purposes of this discussion, the levels of the factors in the design will be referred to as "absent" or "present"; however, designs are not limited to this simple dichotomy. In most designs, "absent" and "present" are usually a "low" and a "high" level of some property of a given factor.

Experimental Designs

An experimental design is, at the most basic level, nothing more than carefully organized one-factor-at-a-time experimentation. For example, let us assume we have two factors that we need to test on a sample population. The simplest basic *ideal* set of one-at-a-time experiments in this case would be that of Table 1.

For purposes of illustration, assume that we are interested in studying the effects of medicine and exercise on the speed of recovery following a surgical procedure. If Factor 1 was a dose level of a given medicine and Factor 2 was the number of minutes of treadmill walking at a given speed, then the experimental design from Table 1 would look like that of Table 1a. Thus, a patient assigned to receive the treatment of Experiment 1 would be the control—no medicine or exercise, and the patient assigned to the treatment of Experiment 4 would be given medicine and assigned 15 minutes of exercise on the treadmill.

Table 1 Matrix of experiments for ideal one-factor-at-a-time experimental design for two factors

Experiment	*Factor A*	*Factor B*
1 (Control)	Absent	Absent
2	Present	Absent
3	Absent	Present
4	Present	Present

Table 1a Illustrated examples of actual factor names and levels

Experiment	*Dose Level*	*Minutes of Walking*
1 (Control)	None	None
2	10 mg	None
3	None	15
4	10 mg	15

This set of experiments is identical to an experimental design of two factors at two values (levels). In this case, the values are the simple presence or absence of the factor of interest.

For three factors, the basic ideal one-factor-at-a-time list of experiments would be those in Table 2.

Table 2 Matrix of experiments for ideal one-factor-at-a-time experimental design for three factors

Experiment	*Factor A*	*Factor B*	*Factor C*
1 (Control)	Absent	Absent	Absent
2	Present	Absent	Absent
3	Absent	Present	Absent
4	Present	Present	Absent
5	Absent	Absent	Present
6	Present	Absent	Present
7	Absent	Present	Present
8	Present	Present	Present

This list of experiments is identical to a three-factor design, where each factor has two values (levels). This kind of a design is called a *full factorial*. Thus, for one to truly adhere to the principle of one factor at a time, an investigator would need to run eight experiments to properly identify the effects of three factors.

If the only concern is the ability to assess the effects of the three factors and assess them independently of one another, then it is possible to use the methods of experimental design and fractionate the above design so that the three factors can be assessed using only four experiments:

Table 3 Fractionated design

Experiment	*Factor A*	*Factor B*	*Factor C*
1 (Control)	Absent	Absent	Absent
4	Present	Present	Absent
6	Present	Absent	Present
7	Absent	Present	Present

If the third factor was the application of heat for 10 minutes to the area of repair, then the final fractionated design for three variables would be that of Table 3a.

The methods used to fractionate a design will not be discussed here. However, the interested reader is referred to the Further Readings at the end of this entry.

Table 3a Fractionated design

Experiment	*Dose Level*	*Minutes of Walking*	*Minutes of Heat*
1 (Control)	None	None	None
4	10 mg	15	None
6	10 mg	None	10
7	None	15	10

One-Factor-at-a-Time Design Matrix

The design matrices in Tables 1 through 3 are, as mentioned, the ideal one-factor-at-a-time design matrices. In reality, the *typical* one-factor-at-a-time design matrix for three factors is that of Table 4.

Table 4 Typical design of a one-factor-at-a-time matrix

Experiment	*Factor A*	*Factor B*	*Factor C*
1 (Control)	Absent	Absent	Absent
2	Present	Absent	Absent
3	Absent	Present	Absent
4	Absent	Absent	Present

Or we could express the matrix in terms of our three hypothetical factors (see Table 4 [Modified]).

Table 4 (Modified) Typical design of a one-factor-at-a-time matrix

Experiment	*Dose Level*	*Minutes of Walking*	*Minutes of Heat*
1 (Control)	None	None	None
2	10 mg	None	None
3	None	15	None
4	None	None	10

At first glance, a simple count of experiments in the design tables would seem to suggest that the design of Table 4 is superior to that of Table 2 and

equal to the design of Table 3. However, Table 4 only lists the *basic combinations* an experimenter would need to run in a typical one-factor-at-a-time experiment involving three factors, whereas Tables 2 and 3 list the *total number* of experiments needed for a single run of an experimental design.

For a typical one-factor-at-a-time experiment to have the same precision of estimate of the effects of the factors that would be achieved by a single run of the experiments in Table 2, the investigator would need to run each of the low and high settings of each of the three variables in Table 4 four times for a total of 8 runs per factor and a total experimental effort of 24 runs. Thus, the true matrix of experiments for a typical three-factor one-factor-at-a-time experiment would be that of Table 4a.

In some cases, where all the experimentation was performed during a short period of time (a day or two) and the factors were all biological in nature, it might be possible to run a single control group of four animals. This would result in some decrease in the precision of the estimates of the effects, and it would reduce the above matrix from 24 to 16 runs. However, this would still be twice as many experiments as Table 2, and it would have the additional assumption that over the ranges of the factors of interest, the effect of any given factor would be the same regardless of the settings of the other variables—that is, over the ranges of the factors of interest, the effect of the factors on the response is that of simple addition. If this is not the case, then in addition to better precision with fewer experiments, the design in Table 2 will also provide the means to detect and estimate the interactions (synergistic effects) that measure this nonadditive behavior.

The reason for the differences in the number of experimental runs needed for a one-factor-at-a-time versus a factorial design is due to the way in which the two methods compute the mean estimates of the factor effects.

For the one-factor-at-a-time matrix in Table 4, the effect of Factor A is computed by taking the sum of the responses to Experiments 1, 3, 5, and 7 and subtracting this from the sum of the response values to Experiments 2, 4, 6, and 8. This result is then divided by 4, the number of measurements at each of the two values of Factor A (absent and present). The result is the average effect of Factor A. This same procedure must then be carried out for the eight experiments for Factor B and the eight experiments for Factor C.

Table 4a Typical design of a one-factor-at-a-time matrix: Three factors

Experiment	*Factor A*	*Factor B*	*Factor C*
1 (Control)	Absent	Absent	Absent
2	Present	Absent	Absent
3 (Control)	Absent	Absent	Absent
4	Present	Absent	Absent
5 (Control)	Absent	Absent	Absent
6	Present	Absent	Absent
7 (Control)	Absent	Absent	Absent
8	Present	Absent	Absent
9 (Control)	Absent	Absent	Absent
10	Absent	Present	Absent
11 (Control)	Absent	Absent	Absent
12	Absent	Present	Absent
13 (Control)	Absent	Absent	Absent
14	Absent	Present	Absent
15 (Control)	Absent	Absent	Absent
16	Absent	Present	Absent
17 (Control)	Absent	Absent	Absent
18	Absent	Absent	Present
19 (Control)	Absent	Absent	Absent
20	Absent	Absent	Present
21 (Control)	Absent	Absent	Absent
22	Absent	Absent	Present
23 (Control)	Absent	Absent	Absent
24	Absent	Absent	Present

In the full-factorial experimental design in Table 2, the effect for Factor A is computed by taking the sum of the responses of the experiments where Factor A was absent and subtracting them from the sum of the responses of the experiments where Factor A was present and dividing this difference by 4. To compute

the effect for Factor B, the same strategy is followed, only now we are adding and subtracting the responses based on the values of Factor B.

If you are accustomed to thinking in terms of one-factor-at-a-time experimentation and analysis, the explanation of the computation of the effect of Factors A and B in the above paragraph would appear to be complete and utter rubbish. How is it possible to take the same eight experiments where both Factors A and B (and C!) are changing at the same time and independently identify the effects of these three factors?

The key to understanding this is to visualize the sentence "The effect for Factor A is computed by taking the sum of the responses of the experiments where Factor A was absent and subtracting them from the sum of the responses of the experiments where Factor A was present and dividing this difference by 4" in tabular form. If this is done, then for Factor A, Table 2 will be modified as shown in Table 5.

If we add up Experiments 1 to 8 according to the coefficients in the column for Factor A, we will have the following:

$$\begin{aligned} &1 \times \text{Present} + 1 \times \text{Present} + 1 \times \text{Present} \\ &+ 1 \times \text{Present} - 1 \times \text{Absent} - 1 \times \text{Absent} \\ &- 1 \times \text{Absent} - 1 \times \text{Absent}, \end{aligned}$$

which reduces to

$$4 \times \text{Present} - 4 \times \text{Absent}.$$

This divided by 4 will give us the average effect of Factor A.

If we apply this same pattern of ± values to the column for Factor B, we will have Table 6.

If we add up Experiments 1 to 8 according to the coefficients in the column for Factor B, we have the following:

$$\begin{aligned} &1 \times \text{Present} - 1 \times \text{Present} + 1 \times \text{Present} \\ &- 1 \times \text{Present} + 1 \times \text{Absent} - 1 \times \text{Absent} \\ &+ 1 \times \text{Absent} - 1 \times \text{Absent}, \end{aligned}$$

which reduces to

$$\begin{aligned} &2 \times \text{Present} - 2 \times \text{Present} + 2 \times \text{Absent} \\ &\qquad - 2 \times \text{Absent} = 0. \end{aligned}$$

Table 5 Illustration of Factor A level coding for purposes of computing Factor A effects

Experiment	*Factor A*	*Factor B*	*Factor C*
1 (Control)	−1 × Absent	Absent	Absent
2	1 × Present	Absent	Absent
3	−1 × Absent	Present	Absent
4	1 × Present	Present	Absent
5	−1 × Absent	Absent	Present
6	1 × Present	Absent	Present
7	−1 × Absent	Present	Present
8	1 × Present	Present	Present

Table 6 Illustration of Factor A and Factor B level coding for purposes of computing Factor A and Factor B effects

Experiment	*Factor A*	*Factor B*	*Factor C*
1 (Control)	−1 × Absent	−1 × Absent	Absent
2	1 × Present	1 × Absent	Absent
3	−1 × Absent	−1 × Present	Absent
4	1 × Present	1 × Present	Absent
5	−1 × Absent	−1 × Absent	Present
6	1 × Present	1 × Absent	Present
7	−1 × Absent	−1 × Present	Present
8	1 × Present	1 × Present	Present

In other words, the computation of the average effect of Factor A results in the simultaneous elimination of the effect of Factor B (Factor B's average effect when computed in this manner is 0). If the same set of coefficients is applied to the column for Factor C, it too will disappear. The same thing occurs when you compute the effect of Factor B—Factors A and C disappear, and similarly for the computation of the effect of Factor C.

The computation of factor effects outlined above is the key to understanding the power and utility of experimental designs. All experimental designs, regardless of the name, are based on this method of determining factor effects.

Table 7 Number of experiments needed to investigate a given number of factors using fractional-factorial, full-factorial, and one-factor-at-a-time methods of experimental design

Number of Factors	*Fractional Design*	*Ideal One-Factor-at-a-Time (Full Factorial)*	*Typical One-Factor-at-a-Time*
2	4	4	12
3	4	8	24
4	8	16	64
5	8	32	160
6	8	64	384
7	8	128	896
8	16	256	2,048
9	16	512	4,608
10	16	1,024	10,240

Reduction of the Experimental Effort

This ability to fractionate a design means that it is possible for an investigator to independently examine the effects of large numbers of factors on one or more measured responses. Table 7 illustrates the savings in experimental effort that can be achieved with this method.

Since biological units (patients and lab animals) typically exhibit more natural unit-to-unit variation than units in an engineering setting (e.g., machines, processes), an investigator will want to run more than one unit with each experimental condition. If one runs as few as four animals per experimental condition and is interested in the effects of just three factors, the total number of animals required to measure the effects of those factors using one-at-a-time methods versus that of fractional-factorial experimental designs is 16 versus 96—a sixfold difference. It is easy to see that the differences in total number of animals and total number of experiments translate into large differences in time, effort, and cost.

Benefits

Experimental designs are the most efficient methods available for identifying significant relationships between factors and responses. They avoid the serious methodological problems of one-factor-at-a-time experimental efforts, and they allow the investigator to independently assess the significance of the effects of multiple factors on any measured response.

Robert S. Butler

See also Equivalence Testing; Hypothesis Testing; Statistical Testing: Overview

Further Readings

Box, G. E. P., Hunter, W. G., & Hunter, J. S. (1978). *Statistics for experimenters: An introduction to design, data analysis, and model building.* New York: Wiley.

Expert Opinion

Expert opinion is a judgment that applies knowledge to a domain-specific problem by a person with superior knowledge in that domain. The term therefore involves two concepts, domain specificity and superiority of knowledge—called expertise. Both are necessary for one to be in a position to offer expert opinion.

Expert opinion is based on judgment. Judgment is an integration task, integrating relevant available cues while excluding irrelevant cues and inferring unavailable information. Judgment becomes opinion with the inference of the unavailable information.

Expertise

Domain specificity means that expertise in one domain does not necessarily transfer to another. An expert in medicine does not likely possess expertise in law. Although there are a few individuals who have training and experience in both domains, whether or not they maintain expertise in both is open to question. Furthermore, within a broad domain such as medicine, expertise is generally limited to subsets of domain knowledge. Thus, an expert in orthopedic surgery would not likely possess expertise in vascular surgery, nor would the expert be likely to have expertise in internal medicine. That does not mean that an individual with expertise in a specific domain would not have useful knowledge of other domains. It merely means that, generally, an individual possesses expertise in only a narrow subset of domain-specific knowledge.

Superior knowledge entails a number of prerequisites. Experience is a necessary, but not sufficient, prerequisite for expertise. Experience can allow an individual to develop schema for domain-specific problems. Schemata are mental representations of a situation. For instance, an internist specializing in infectious tropical disease would likely have a schema for schistosomiasis. A general practitioner practicing in the rural United States would not be likely to have such a schema.

Experience may further elaborate schemata through feedback and allow for the development of ability to discriminate between similar schemata. For instance, a specialist with extensive experience in tropical infectious disease should be able to differentiate between schistosomiasis, Chagas disease, and malaria. Other physicians likely would not. Experience and the feedback that is gained through experience allows for the development of scripts to match specific schema. Scripts are behavioral protocols that are appropriate for specific schemata.

With experience, discrimination of the script that accompanies a schema becomes increasingly automatic. This is why experts often have difficulty articulating their thoughts; the schema and scripts have become so automatic that they are processed rapidly without conscious awareness. Thus, experts may be able to offer an expert opinion more easily than they can explain how they reached that opinion. However, if one is not organizing experience into schemata, attending to feedback, developing scripts to accompany specific schema, and continually updating these memory structures, one may have experience without expertise.

As is implied by the need to update memory structures, expertise must be continuously updated. Domain knowledge in many fields, medicine being a prime example, is not static. An individual who is an expert in orthopedic surgery at one point in time, but who does not continually update and expand his or her knowledge, loses expertise. This is why expertise is often found in academic arenas. To teach, one must continually update knowledge to maintain and further develop schemata and scripts.

Level of expertise in making judgments in any specific domain is related to how much knowledge is available about how that domain operates and is structured and how much feedback is available from decisions previously made in that domain. People are more likely to become expert if they operate in fields where much is known and feedback from previous decisions is consistent and relevant. Those who practice without these environmental elements are handicapped in their ability to develop expertise.

Experts need not and often don't agree. Although on the surface, this seems like an oxymoron, it follows from two facts. First, if two people provide opinions that disagree, one may later be found to be correct, and the other by elimination would be incorrect. However, which is correct may not be known at the time a decision must be made. Sometimes, the correct opinion is not known until after the decision is made. Second, since judgment is an integration of known information to infer otherwise unavailable information, agreement between judges does not imply that an agreed-on opinion is correct. At one time, experts agreed that the sun revolved around the earth.

Types of Expertise

Within a domain, information and performance can be separated into three kinds of mental models. This

delimitation of expertise is the work of Jens Rasmussen. Expertise can be described as skill based, rule based, and knowledge based. Skill-based mental models allow for the ability to physically manipulate the environment within a spatial and temporal frame of reference, based on superior sensory motor skill. Skill is useful for many domains and necessary for some, for example, surgery. Skill qualifies one as an expert in a domain of physical practice, such as surgery. Skill-based expertise allows one to physically intervene in a situation where skill is required. However, superior skill does not qualify one to offer an expert opinion on the domain.

Rule-based mental models involve knowledge of relationships between cues that activate familiar schemata and scripts. Superior skill and superior rule-based knowledge may be found in the same individual. Recognition-primed decisions are rule based. An expert in rule-based decision making can quickly identify the schema and scripts that are appropriate to a familiar situation. A person who has superior rule-based expertise is in a position to offer an opinion about which rule should be applied to situations for which there are established rules, but this person does not necessarily have the ability to offer an expert opinion about novel situations.

Mental models based on knowledge involve understanding of the organization and operation of domain phenomena and of relationships between structures and concepts within the domain. Knowledge-based mental models allow novel situations to be understood and appropriate responses to be developed. It is possible to have all three levels of expertise, but this is not always the case. A unique trait of knowledge-based decision makers is the ability to know when a rule does not cover a situation and to develop novel alternatives.

Well-developed knowledge-based mental models allow one to offer an opinion about how to respond to a novel situation for which the rules are unclear or for which rules do not exist. Knowledge-based decision making is not restricted to the ability to diagnose but rather includes the ability to recognize what information is demanded by the situation and what tests and procedures will clarify that information. Knowledge-based decision making includes the ability to select the best treatment and to know how to monitor that treatment so that it can be evaluated and adjustments made.

Measurement of Superior Knowledge

One measure of expertise is to survey those in the domain for which one requires expert opinion and choose the person whom most peers judge to be the most expert. This approach is likely to confuse skill-, rule- and knowledge-based expertise. It has the added limitation of a halo effect: Those who are most likeable are often judged as more expert. Still another method of establishing expert knowledge is to develop a panel of people with domain experience and assume that the points on which the panel agrees can be considered expert opinion. Guidelines for clinical practice often encapsulate a consensus view from professionals designated as experts. This approach is based on two assumptions: first, that experience, and often hierarchical position, captures expertise and second, that consensus captures truth. Both assumptions have been shown to be invalid, as noted above.

There is no way to measure superior knowledge directly. Expert opinion involves making a judgment rather than acting on that judgment. Since judgment is necessary prior to decision and action, expert opinion involves the knowledge from which to make a judgment but does not necessarily involve decision making. However, performance implies knowledge and can be objectively measured.

The best way to identify expert performance is to identify those who exhibit the ability to discriminate relevant cues in a domain of practice and do so consistently. The focus on the ability to discriminate relevant cues from irrelevant ones taps into cognitive elements underlying performance. The focus on consistency in this ability eliminates performance that is effective only part of the time because the individual does not have a thorough grasp of the knowledge necessary to make a consistent decision. Expert opinion might be available from those who are able to consistently discriminate what is important to decisions in a particular practice domain.

Application

Expert opinion is often used to provide guidance when more objective guidance, such as testing, is unavailable or equivocal or in decisions for which the rules are unclear. Therefore, expert opinion usually refers to knowledge-based expertise. It is

not surprising to see much of the literature on expert opinion aimed at forensic decisions, such as likelihood to reoffend. However, some use of expert opinion may involve providing information on the correct rules that should be applied, such as when an expert is asked to state the standard of practice for a given situation.

Selection Criteria

When selecting someone from whom to obtain an expert opinion, the selection criteria should include the following: (a) The experts must be knowledgeable, not just skilled, in the specific domain about which they are to express an opinion; (b) they should understand the rules of that domain as well as possess an in-depth understanding of the mechanisms that underlie the operation of that system; (c) they should have shown that they are able to make accurate judgments within that domain on the task for which they will offer an opinion; and (d) they should have done this with a high level of consistency. Nowhere in these criteria is there a direct requirement for experience; however, most, if not all, of the criteria imply experience as a prerequisite.

James Shanteau and
Alleene M. Ferguson Pingenot

See also Expert Systems; Judgment

Further Readings

Ericsson, K. A., Charness, N., Feltovich, P. J., & Hoffman, R. R. (Eds.). (2007). *The Cambridge handbook of expertise and expert performance*. New York: Cambridge University Press.

Rasmussen, J., Pejtersen, A. M., & Goodstein, L. P. (1994). *Cognitive systems engineering* (Wiley Series in Systems Engineering). New York: Wiley.

Weiss, D. J., & Shanteau, J. (2003). Empirical assessment of expertise. *Human Factors, 45*(Spring), 104–114.

Expert Systems

The concept of expert medical systems has changed over several decades from that of a system that would replace human decision making with machines modeled on the behavior of experts to that of software systems that provide information and support to human decision makers. Expert medical systems are computer systems that facilitate the work of clinical decision makers, increasing their efficiency and accuracy while remaining responsive to changes in knowledge and flexible in response to clinical needs. Despite progress in design, recent systems still experience failure more often than is acceptable, and performance is suboptimal in many cases.

It was easy, in the rush to capitalize on the ability to store information in computers, to design what some considered expert medical systems without first gaining a thorough understanding of the concepts integral to expertise in medicine. Furthermore, knowledge in medicine is always expanding. Any system designed without a mechanism for continual review and updating of information quickly becomes out of date and is hence worse than useless. A system that does not consider the needs of all users is an error-prone system.

Researchers have gained insights into how effective human decision makers think as well as knowledge of what machines do best and what humans do best. They also are learning about how the two, man and machine, interact. These are tools necessary to accomplish the goal of designing functional and reliable expert medical systems.

Human Decision Behavior

It is not necessary to have an exact model from which to design a functional system. Rather, it is important to identify information that is critical to effective decisions in the targeted situations. The ability to design a system that includes critical information, but is inexact otherwise, allows for the development of adaptive systems.

Research on human experts can identify information needed for effective decision making. Human experts do not use all available information. Rather, they use information relevant to the decision at hand. They know what information is missing and look for disconfirming as well as confirming evidence. Experts use feedback from each small, incremental decision to adjust their understanding of the situation before making the next

decision. This approach allows for both flexibility and recovery from error.

Interaction of Human and Machine

The interaction of human and machine (computer) can be conceptualized as similar to two individuals working on a joint project. That is, the human and the machine are part of a decision-making dyad. Machines can only do what they are programmed to do. The human part of the dyad must be able to perceive whether and how some action might be best accomplished using the machine. This means that the interface must be designed to be intuitive to the user.

The work of humans interacting with machines is supervisory. Machines do some things well, such as retrieving stored information, conducting complex operations, performing repetitive or routine tasks, and maintaining archives. The human decides what information to provide to the machine, uses information retrieved from storage, directs computations to be performed, and updates evidence for practice.

Safety

Keeping the human decision maker in charge is especially important when exceptions arise or the situation changes—circumstances that machines are not designed to accommodate. Experts have knowledge of the situation and the goals to be accomplished, and can devise novel approaches to solve unusual problems that arise.

Change in system state must be collaborative between the human user and the machine. Without feedback, the behavior of the human user can be irrational, even dangerous. Accident investigations often reveal that lethal errors occured when a human user misunderstood what a machine was doing, for example, when the interface did not provide information in a way that was intuitive to the user. As a result, the human made erroneous decisions. Feedback from the machine is important even when a program does most of the work because when the human takes over, he or she needs to know what has happened in the system prior to making a decision and taking action.

Constraints should be built into the system. Constraints are identified by a thorough understanding of the work as a whole, including the specific goals, tasks, and options of the operator. Constraints identify behaviors that *can't* be done; for example, one should not order incompatible drugs. A well-designed expert system would notify the user of drug incompatibility rather than blindly documenting administration, as is the case with many existing systems. An expert system designed with constraints in mind, but that includes flexibility for situations when the built-in rules do not apply, is critical for success.

When expert decision systems are programmed with default settings, the default should be a safe setting. Fatalities have resulted from machines that were programmed for default settings that turned out to be lethal. To properly select a default setting, research should identify the typical or "normal" setting of the system and program that as the default, requiring the user to actively change the default settings if using other than typical values. Expert systems must never be programmed to perform outside the safe limits of operation. Any change to a setting outside safe operational range should require verification for the change.

The issue of locking out behavior that shouldn't be performed has generated lively debate. Decision aiding and warning flags should be viewed as information exchange between the machine and the human decision maker. Experienced clinicians can think of examples where exceptions must be made to the general rule or where a decision support simply does not have the relevant information. The ability to know when rules don't apply or when critical information is missing is a trait identifying human experts.

Flexibility in expert systems allows for human experts to modify the system's behavior based on experience or information not available to the machine. Clinicians should be able to override warnings by documenting their clinical reasoning and take responsibility for the decision. In addition, it would be useful for clinicians to supply a plan for identifying and responding to adverse outcomes to their decision. This approach preserves flexibility while demanding accountability. Design of warnings within decision aids is an area ripe for research.

Decision Support

The usefulness of decision support systems depends on how well they are designed. There are areas

about which researchers have a great deal of prior knowledge and can therefore build "expert systems" using if-then rules. These systems are particularly useful for nonexperts, who must sometimes make decisions for which they lack the knowledge or skills. Also, such systems can be useful for training students.

Some machine behavior enhances human performance by accomplishing things that humans are physically or cognitively unable to perform, such as the precise serial radiography of a CT machine. However, these machine behaviors serve to enhance the behavioral response of the decision maker rather than replace the human. Although a CT machine incorporates an expert system, it must be programmed by humans using specific parameters to accomplish the task to be done.

There are a number of ways in which expert advice can be designed into medical systems to assist in accessing relevant information. Rather than simply adapting existing machine structures from other applications, such as business, it is imperative that clinical decision making be examined in terms of goals, the needs of the clinician, information flow, and a deep understanding of the clinical situation. When appropriately designed systems meet the needs of users, they will be used. However, evidence shows that decision support systems do not necessarily lead to better clinical decision making, nor do they necessarily increase patient safety or reduce costs. It is well known that well-intentioned decision support systems are often overridden by users.

There are areas where the knowledge needed to build an expert system simply doesn't exist (e.g., some complex treatment problems). In these decision contexts, it is more useful to produce probabilistic advice based on linear modeling of what is known, rather than outputting a single decision per se. The question to the decision support tool in some of these situations might be posed as "the probability of x happening if treatment regimen y is pursued, given the known facts of the patient situation." Such linear models have been shown to outperform human decision makers, particularly in situations where information is ill-defined and incomplete.

One area where decision supports are being developed is for aiding patients in their own healthcare decision making. Decision aids designed for the lay public are necessarily different in focus from those designed for clinicians. Research on these decision tools focuses on issues such as how best to display information, which information is most relevant on specific topics, and designs for ease of access and use. Interestingly, it appears that more research may be dedicated to the design of patient decision aids than to the design of clinical decision support for clinicians.

Innovative Uses of Machines to Manage Information

The availability of large clinical data sets led to research that identifies and categorizes information for the study of specific clinical problems. For such work (collection and organization of information), machine systems are invaluable. Research using large clinical data sets includes studies of adverse drug reactions and analysis of the relationship of cancer stages to other clinical information. In addition, computer systems assist with quality-of-care assessments by informing clinical decisions that improve delivery of care.

The complexity of medical data is at the root of many of the problems encountered in developing effective expert tools for supporting clinical decision making. Several research programs studying design of expert medical systems have explored the use of fuzzy logic systems as a way to model the complex flow of information required in medicine. This approach seems compatible with the fact that human experts use information in an incomplete but highly functional way, as was discussed above.

It is especially encouraging to find that expert system design innovations are now being more carefully evaluated than were early systems. However, many of these evaluations are based primarily on qualitative feedback from users. As research on the design of clinical systems matures, it is hoped that more objective measures, such as clinical outcomes and efficiency, will become standards of design excellence.

Future Directions

The outlook for expert medical systems is bright. However, the future belongs to systems that augment human decision making by performing simple

repetitive activities and calculations that humans do poorly and providing critical information in a timely way. Once these systems become functional, they will likely be well accepted. It might be useful, however, to recognize that interacting with machines changes our behavior. It seems likely that the integration of expert medical systems has already and will continue to change the social environment in which medicine is practiced, perhaps in ways we can't imagine at present.

James Shanteau and
Alleene M. Ferguson Pingenot

See also Computer-Assisted Decision Making

Further Readings

Nelson, W. L., Han, P. K. U., Fagerlin, A., Stefanek, M., & Ubel, P. A. (2007). Rethinking the objective of decision aids: A call for conceptual clarity. *Medical Decision Making, 27*(5), 609–618.

Pingenot, A., Shanteau, J., & Sengstache, D. (2008). Cognitive work analysis of an inpatient medication system. In *Computers, informatics, nursing*. Hagerstown, MD: Wolters Kluwer/Lippincott Williams & Wilkins. Manuscript submitted for publication.

Shanteau, J. (1992). The psychology of experts: An alternative view. In G. Wright & F. Bolger (Eds.), *Expertise and decision support* (pp. 11–23). New York: Plenum Press.

Vicente, K. J. (1999). *Cognitive work analysis: Toward safe, productive and healthy computer-based work*. Mahwah, NJ: Lawrence Erlbaum.

Weir, C. R., Nebeker, J. J. R., Hicken, B. L., Campo, R., Drews, F., & LeBar, B. (2007). A cognitive work analysis of information management strategies in a computerized provider order entry environment. *Journal of the American Medical Informatics Association, 14*(1), 65–75.

Wright, G., & Bolger, F. (Eds.). (1992). *Expertise and decision support*. New York: Plenum Press.

Extended Dominance

The term *dominance* in the context of cost-effectiveness analysis refers to the situation in which two clinical strategies are being compared. One strategy, Strategy *X*, is said to dominate another, Strategy *Y*, if either (a) the expected costs of Strategy *X* are less than the expected costs of Strategy *Y* and the expected benefits of Strategy *X* are at least as great as the expected benefits of Strategy *Y* or (b) the expected benefits of Strategy *X* are greater than the expected benefits of Strategy *Y* and the expected costs of Strategy *X* are not greater than the expected costs of Strategy *Y*. Usually, the dominant strategy is both more effective and less costly than the alternative. This concept of dominance is also referred to as *strong dominance* or *simple dominance*.

The *extended dominance principle* (also known as *weak dominance*) is applied in cost-effectiveness studies that compare mutually exclusive interventions. This is the situation where only one of the strategies is available to each participant.

The concept of extended dominance is applied in incremental cost-effectiveness analysis to eliminate from consideration strategies whose costs and benefits are improved by a mixed strategy of two other alternatives. That is, two strategies may be used together as a "blended" strategy, instead of assigning a single treatment strategy to all members of a population. Blending strategies only becomes relevant when the most effective strategy is too costly to recommend to all.

The concept may have been first suggested when a particular clinical strategy was "dominated in an extended sense," thus leading to the term *extended dominance*. Extended dominance rules out any strategy with a higher incremental cost-effectiveness ratio (ICER), which is greater than that of a more effective strategy. That is, extended dominance applies to strategies that are not cost-effective because another available strategy provides more units of benefit at a lower cost per unit of benefit.

Among competing choices, an alternative is said to be excluded by extended dominance if its ICER relative to the next less costly undominated alternative is greater than that of a more costly alternative.

Here is a simple example of a competing choice problem that can be evaluated for strong dominance and extended dominance. Table 1 shows costs and outcomes for standard of care and five hypothetical interventions.

From the comparison of costs and outcomes, we can rule out Intervention E because it is strongly dominated by Intervention D. Intervention D costs

Table 1 Costs and outcomes for standard of care and five hypothetical interventions

Strategy	*Cost ($)*	*Effectiveness (QALYs)*
Standard of care	5,000	1
E	12,000	1.5
D	10,000	2
C	25,000	3
B	35,000	4
A	55,000	5

less and gives better outcomes than E. Having ruled out Intervention E, we can compare the remaining strategies based on their ICERs. This is where the principle of extended dominance comes in. Table 2 shows the remaining interventions listed in order of effectiveness. The ICER of each intervention is found by comparing it with the next most effective option.

We can now use the principle of extended dominance to rule out Intervention C. Intervention C has an ICER of $15,000 per quality-adjusted life year (QALY). To agree to use Intervention C, the deciding body would have to agree to adopt all interventions with ICERs up to $15,000 per QALY. If so, they would be much better off choosing Intervention B over Intervention C, since a greater number of QALYs can be obtained with this intervention at a lower cost per QALY. The logic goes thus: If one is willing to pay a smaller amount to gain a life year (or QALY or whatever unit of effectiveness) with the more expensive strategy, then one should not choose the strategy with the higher ICER.

Table 3 shows the interventions and their ICERs after the extended dominance principle has been applied. It is now up to the decision maker to choose among the interventions based on how much they are willing to pay for a QALY.

If willingness to pay (WTP) is not even $5,000 per QALY, then none of the interventions generates sufficient worth to be adopted. If however, WTP is greater than $20,000 per QALY, then Intervention A would be adopted.

As mentioned above, when extended dominance exists, it is possible to create a mixed strategy of two alternatives (i.e., when one portion of the population receives one strategy and the remainder receives an alternative strategy) that can yield greater or equal benefits at an equal or cheaper cost than would a third alternative, if applied to all members of the population. For those strategies that were eliminated from consideration by extended dominance, a range of plausible mixed strategies that would dominate the eliminated alternatives can be computed.

The coefficient of inequity is defined as the minimum proportion of people receiving the worst strategy within a mixture of two strategies when invoking extended dominance. The coefficient of inequity represents a level of unfairness if a mixed strategy were ever to be implemented. Since in extended dominance, a linear combination of two strategies can be shown to dominate a third strategy, from a practical perspective, this may have ethical ramifications. It implies that a strategy is dominated because a given fraction of the population may be receiving an inferior strategy for the overall health of the population to be improved.

Table 2 Strategies after considering simple (strong) dominance

Strategy	*Cost ($)*	*Effectiveness (QALYs)*	*ICER ($)*
Standard of care	5,000	1	—
D	10,000	2	5,000
C	25,000	3	15,000
B	35,000	4	10,000
A	55,000	5	20,000

Table 3 Strategies after considering extended (weak) dominance

Strategy	*Cost ($)*	*Effectiveness (QALYs)*	*ICER ($)*
Standard of care	5,000	1	—
D	10,000	2	5,000
B	35,000	4	12,500
A	55,000	5	20,000

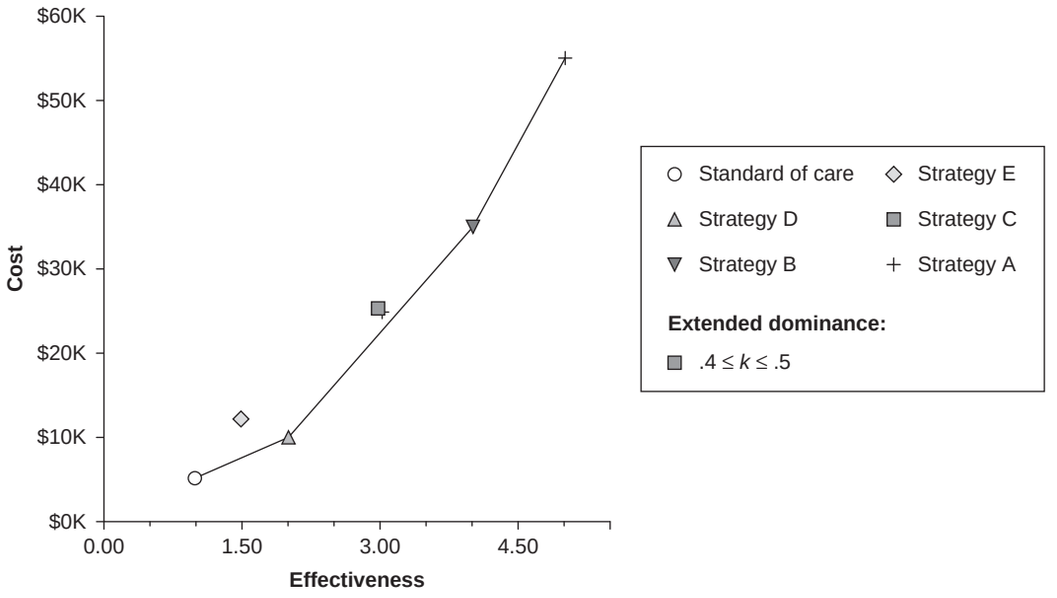


Figure 1 Example of extended dominance

The above example is graphically represented in Figure 1.

In the example, Strategy E is dominated by Strategy D (dominance). Strategy C is dominated by a blend of Strategy D and Strategy B (extended dominance), with a coefficient of inequity equal to .4. The coefficient of inequity is calculated as the difference of the cost of the more expensive strategy and the cost of the weakly dominated strategy divided by the difference of the cost of the more expensive strategy and the cost of the cheaper strategy; in this case, this is

$$(35{,}000 - 25{,}000)/(35{,}000 - 10{,}000) = 10{,}000/25{,}000 = .4.$$

Lesley-Ann N. Miller and Scott B. Cantor

See also Cost-Benefit Analysis; Cost-Effectiveness Analysis; Cost-Utility Analysis; Dominance; Efficacy Versus Effectiveness

Further Readings

Cantor, S. B. (1994). Cost-effectiveness analysis, extended dominance and ethics: A quantitative assessment. *Medical Decision Making, 14,* 259–265.

Gold, M. R., Siegel, J. E., Russell, L. B., & Weinstein, M. C. (Eds.). (1996). *Cost-effectiveness in health and medicine*. New York: Oxford University Press.

Hunink, M. G. M., Glasziou, P. P., Siegel, J. E., Weeks, J. C., Pliskin, J. S., Elstein, A. S., et al. (2001). *Decision making in health and medicine: Integrating evidence and values*. Cambridge, UK: Cambridge University Press.

Johannesson, M., & Weinstein, M. C. (1993). On the decision rules of cost-effectiveness analysis. *Journal of Health Economics, 12,* 459–467.

Kamlet, M. S. (1992). *The comparative benefits modeling project: A framework for cost-utility analysis of government health care programs*. Washington, DC: U.S. Department of Health and Human Services.

Raiffa, H., Schwartz, W. B., & Weinstein, M. C. (1977). Evaluating health effects of societal decisions and programs. *Decision making in the environmental protection agency* (Vol. 2b, pp. 1–81). Washington, DC: National Academy of Sciences.

Stinnett, A. A., & Paltiel, A. D. (1996). Mathematical programming for the efficient allocation of health care resources. *Journal of Health Economics, 15,* 641–653.

Torrance, G. W., Thomas, W. H., & Sackett, D. L. (1972). A utility maximization model for evaluation of health care programs. *Health Services Research, 7,* 118–133.

Weinstein, M. C., & Fineberg, H. V. (1980). *Clinical decision analysis*. Philadelphia: W. B. Saunders.

F

Factor Analysis and Principal Components Analysis

On the surface, the methods of factor analysis and principal components analysis (PCA) share similarities and common purposes. In particular, they both involve the characterization of multiple variables into components, or factors. However, factor analysis is much more ambitious than PCA in that it involves modeling assumptions, in particular the modeling of latent, unobservable factors.

Principal Components

PCA can be used to reduce the dimensionality of data in the sense of transforming an original set of variables to a smaller number of transformed ones. Such a purpose is desirable as it allows for the parsimonious explanation of the systematic variation of data with as few variables as possible. Obtaining parsimonious representations of data is especially useful when confronted with large numbers of variables, such as those found in survey data or genetics data. Socioeconomic variables have been combined into a smaller number through PCA as well. Furthermore, in regression analyses, multicollinearity can be a serious concern when there are a large number of variables to model. Reducing the number of variables used in an analysis or transforming the original variables to make them uncorrelated, as PCA does, can alleviate this problem.

PCA involves rotating multivariate data, which involves transforming the original variables into a new set of variables that are linear combinations of the original variables. This rotation process yields a new set of variables with desirable properties.

Let $X_1, \ldots, X_p$ (the Xs) denote the original variables. For instance, the Xs could be clinical variables, such as X_1 being weight measurements, X_2 being heights, X_3 being systolic blood pressure, and so on. Each X_i, $i = 1, \ldots, p$, is a vector with n elements, representing, for instance, n observations of the variable X_i from n subjects. A linear combination of the Xs would take the form $a_1X_1 + \ldots + a_pX_p$, for some constant weights $a_1, \ldots, a_p$. Loosely speaking, one object of PCA is to find uncorrelated linear combinations of the Xs that maximize the variance, a measure of the variability in data. Weights for the linear combinations being considered are restricted so that the sum of their squared values is 1. This restricts possible solutions under consideration to be derivable from rotations. Based on elegant theories from linear algebra, a sketch of how they are derived is given below (for more details, see Tatsuoka, 1988).

Given variables $X_1, \ldots, X_p$, one can construct a $p \times p$ matrix $\mathbf{A}$ that is composed of sample covariances A, with the i,jth entry in $\mathbf{A}$ corresponding to the sample covariance between X_i and X_j. Covariances measure the degree to which two variables vary together, or are correlated. We can solve what is known as the characteristic equation for the matrix $\mathbf{A}$ and generate p nonnegative roots (although it is possible that some roots are equal,

or even zero, depending on the rank of **A**). This equation is derived based on the objective of finding linear combinations of the *X*s that maximize variance. These roots are known as the eigenvalues of the matrix **A**. Furthermore, given these eigenvalues, corresponding vectors, called eigenvectors, can be derived where the elements in the eigenvectors give the weights for the desired linear combinations. Moreover, the eigenvalues are equal to the corresponding variance of the linear combination of the *X*s, with weights corresponding to the eigenvector. Hence, the eigenvalues and eigenvectors that are generated provide the essential practical information in attaining the objectives of PCA.

Denote the eigenvalues in descending order as $\lambda_1, \lambda_2, \ldots, \lambda_p$ so that $\lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_p$. More explicitly, PCA will generate a new set of variables, or components, $Y_1, \ldots, Y_n$ (the *Y*s), with

$$Y_1 = a_{(11)}X_1 + \ldots + a_{(1n)}X_n,$$
$$Y_2 = a_{(21)}X_1 + \ldots + a_{(2n)}X_n, \ldots,$$

where $a_{\{ij\}}$ are constants such that the sum of the squared values of $a_{\{ij\}}$ is 1. Again, these constants are derived from the elements in the associated eigenvector. Importantly, the new components have conditionally maximal variances in the following sense: Y_1 has maximum variance among all such linear combinations of the *X*s, Y_2 has maximum variance among all such linear combinations of the *X*s that are uncorrelated with Y_1, Y_3 has maximum variance among all such linear combinations of the *X*s uncorrelated with Y_1 and Y_2, and so on. Moreover, the variance of Y_1 is λ_1, the variance of Y_2 is λ_2, and so on.

An important result of this transformation is that the sum of the λs is equal to the sum of the variances of the *X*s. Thus, the variation of the *X*s can be viewed as "reshuffled" among the *Y*s, and this variation is concentrated on as few variables as possible. It is in this variation that the statistical information provided by the variables is contained. A subset of the *Y*s can be parsimoniously selected for subsequent analyses, and such a subset indeed does represent much of the variation in the *X*s. We can determine this as follows.

If the *X*s span a linear subspace with dimension *r*, with $r < p$, then PCA will find $(p - r)$ degenerate components (all zero weights for those components). While in practice, purely nondegenerate components won't be found due to random variation, components nonetheless could appear to be "essentially" degenerate, for instance, as measured by relatively small associated eigenvalues. In such cases, the components with the larger associated eigenvalues would contain most of the variation in data across the *X* variables, and hence little information would be lost by retaining only those components.

So a key methodological issue in applying PCA involves determining which components to keep for an analysis and which to discard. There are many approaches and criteria in helping make this decision. Two basic rules of thumb for selecting components are as follows: (1) retain components with the highest associated variances (eigenvalues) such that the total variation ratio, which is equal to the ratio of the sum of eigenvalues associated with the retained components to the sum of all eigenvalues, is greater than .85 or .90, and (2) choose components with a corresponding eigenvalue of at least 1.0. Another approach is the scree test, where eigenvalues are plotted in order of magnitude and truncation of components is determined by identifying a cutoff for when the changes in associated eigenvalue magnitudes appear to begin leveling off as the eigenvalues get increasingly smaller. Components are truncated when associated eigenvalues that are of the smallest magnitude are deemed to be beyond the cutoff.

Importantly, for PCA, there are no distributional assumptions that have to be made about the *X*s. Moreover, there are no modeling assumptions to validate either. PCA is thus a widely applicable statistical tool with a clearly defined and attainable purpose.

Factor Analysis

Factor analysis attempts to describe an observed set of variables (the *X*s) in terms of a linear model of unobservable factors (the *Y*s), much as in a regression model. However, a key difference is that the *Y*s are latent and unobservable. Factor analysis can thus be used to explore or reveal an internal structure or hidden relationships between observable variables by linking them to underlying latent constructs. Because of the presence of latent factors and the key role they play, factor analysis presents a difficult and ambitious statistical modeling problem. Yet it is a commonly used method

because there are a range of problems in which it is desirable to model observable phenomena as a function of unobservable factors.

For instance, in psychologically related applications, such as in psychiatry, quality-of-life measurement, and neuropsychological assessment, it is sometimes posited that underlying constructs are a driving force in the behavior of observed variables. For example, collections of neuropsychological measures (the *X*s) could be employed to assess the impact of a treatment on cognitive functioning. Such measures could assess different aspects of cognition—for example, through tasks that require memory or strategizing. Factor analysis could be used to assess what types of underlying cognitive functions (the *Y*s), as represented by latent constructs, are in fact being tested. Examples of latent constructs that are identified in such applications include various types of memory functions, motor skills, and executive functions, which are posited as higher-order functions used to regulate other cognitive abilities.

A factor analysis model can be written as follows to relate the observable variables X_i, $i = 1, \ldots, p$, to unobservable factors Y_j, $j = 1, \ldots, k$:

$$X_i = a_{i1}Y_1 + \ldots + a_{ik}Y_k + d_iU_i,$$

where the $a_{\{ij\}}$, $j = 1, \ldots, k$, are called the factor loadings and are constants; U_i is a unique factor associated with X_i; and d_i is the factor loading for U_i. It is assumed that U_i are statistically independent from the *Y*s.

Note then that these models of the observable *X* variables have two components: (1) that which can be attributed to, or explained by the latent *Y* factors that are common among all observable *X* variables and (2) that which is unique to each variable, not part of the common-factor structure of the *Y*s. The uniqueness of each X_i is described by d_iU_i. Generally, it is assumed that X_i represents a standardized variable, in the sense that the mean of X_i is 0 and the variance is 1. This can be achieved by transforming the original variables through subtracting from each X_i variable its mean and then dividing this difference by the standard error of X_i.

For each variable X_i, the variance that is attributable to the common latent factors is known as its communality. Denote this communality by h_i^2. Thus, $1 - h_i^2$ is the variance unique to the given X_i variable.

Since the unique factors are independent from the common factors, if the respective communalities can be substituted along the diagonal of the correlation matrix between the *X*s, this modified correlation matrix thus represents the correlation between the common factors, given that the linear model is true. Such a modified matrix is called the reduced correlation matrix. Unfortunately, communalities are not known in advance and must be estimated.

Communalities must be estimated iteratively, since we must have an understanding of the factor structure (number of factors and loadings) first before estimating them. On the other hand, communalities must be known to create the reduced correlation matrix on which estimation of the factor structure depends. Generally, prior estimates of the communalities are made, then the factor structure is estimated, the communalities are then reestimated, and this process is iterated until convergence is met, as defined by some criteria that indicate that the estimates have stabilized from iteration to iteration.

Based on the reduced correlation matrix, the same matrix theory as the one used for PCA can be employed to derive uncorrelated factors and associated factor loadings. This is done by solving for eigenvalues and eigenvectors, as before. Again, the number of factors to keep in the model must be determined, by using the scree test or other methods.

Since factor analysis is used to assess the underlying latent structure and relationships between the observable variables, it is desired to understand and characterize the nature of the latent factors that are found. This is done by interpreting the sign and magnitude of the factor loadings and identifying the patterns that arise in terms of how the factor loadings are associated with the observable variables. Generally, though, the factors and their relationship to the observable variables do not easily lend themselves to interpretation. This drawback also is shared with PCA, which also can be used to assess if there is some pattern, or structure, between variables. Yet such interpretation is a major aim for many practitioners as they set out to conduct factor analysis. Interpretation can be improved through rotation, since rotations will

change the factor loadings. Transformed factors may no longer be uncorrelated, but the reduced correlation matrix for the transformed factors remains unchanged, and hence the rotated factors are an equally plausible statistical formulation of the factor structure.

L. Thurstone described target criteria for selecting a rotation such that the transformed factors have certain characteristics that make interpretation easier. The main ones are roughly as follows: (a) each row should contain at least one zero so that not all variables are related to all factors; (b) each column should contain zeros as well so that each factor is not related to all variables; and (c) every pair of columns should contain rows whose loadings are zero in one column but nonzero in the other, so that factors are differentially related to the variables. Of course, no real-life factor-loading pattern will satisfy these criteria exactly, but it is certainly still desirable that they be approximately satisfied. There are a number of rotational techniques that have been developed to enhance interpretability of factor loadings, such as the varimax rotation.

As an example, using factor analysis, suppose two main underlying factors are identified among six measures, with the following factor loadings:

	Factor 1	*Factor 2*
Measure 1	0.8	0.05
Measure 2	0.9	0.05
Measure 3	–0.8	0.04
Measure 4	0.05	0.8
Measure 5	0.04	0.8
Measure 6	0.05	–0.9

The above factor structure approximately satisfies Thurstone's target criteria. Indeed, one could now attempt to identify an underlying common theme within the collection of the first three measures and within the collection of the last three measures in order to give interpretations to the two latent factors, respectively. Note that in practice, factor loadings can be negative. In terms of interpretation, this would imply that larger values of the observed factor are associated with smaller values of the underlying associated factor.

Factor analysis is used in validating scales, where, for instance, certain questions are grouped together and the response scores are combined to generate a scale score. For instance, in the quality-of-life survey SF-36, subscales can be generated from subsets of questions relating to body pain, social functioning, physical functioning, mental health, and so on. Justification for such groupings is supported if grouped questions share similarly high factor loading values on the same factor relative to other variables. Of course, such groupings must also be justified clinically.

Factor analysis is dependent on the information provided by correlations to estimate underlying relationships and factor loadings. Correlations are suited for measuring the strengths of linear relationships. Hence, nonlinear relationships between latent factors and observable variables may not be modeled well. Interpretation of factors is not clear-cut, nor is the selection of an appropriate rotation. Another critical subjective decision that must be made concerns the number of factors to keep in the model. These ambiguities make the task of detecting underlying structure more difficult. Modeling latent factors is an ambitious endeavor, and hence, model fit must be validated in a thorough manner, such as through cross-validation. Replication of findings may be elusive given all these issues. In sum, one should be cautious in drawing conclusions through factor analysis.

Curtis Tatsuoka

See also SF-36 and SF-12 Health Surveys; Variance and Covariance

Further Readings

Catell, R. (1966). The scree test for the number of factors. *Multivariate Behavioral Research, 1,* 245–276.

Kaiser, H. (1958). The varimax criterion for analytic rotation in factor analysis. *Psychometrika, 23,* 187–200.

Kaiser, H. (1960). The application of electronic computers in factor analysis. *Educational and Psychological Measurement, 20,* 141–151.

Tatsuoka, M. (1988). *Multivariate analysis: Techniques for educational and psychological research* (2nd ed.). New York: Macmillan.

Thurstone, L. (1947). *Multiple factor analysis.* Chicago: University of Chicago Press.

Fear

Fear and anxiety can alter decision making in a wide range of domains, not least of all decisions about one's own health or the health of patients under a physician's care. Past research has demonstrated this influence, including in medicine. Understanding the impact that these basic emotions have on medical decisions, particularly those involving risky and uncertain options, is essential to understanding medical decision making and building accurate predictive models of choice. Traditional economic models of decision making, such as expected utility theory, propose that patients and physicians weigh decision options rationally and choose an action based on the likelihood and the payoff of outcomes. These models rarely include psychological influences on behavior, particularly the emotional ones. In the medical context, an important omission from these models is the effect of patients' and physicians' emotions as they weigh the options associated with treating a serious medical condition and choose an action.

Patients and physicians must consider the possible consequences of treatment decisions, and how likely these would be to occur. Decisions involving risky, uncertain outcomes are especially susceptible to the influence of emotions such as anxiety. Anxiety is common in patients with serious illness who must make risky treatment decisions with major consequences: death, functional disability, diminished quality of life and psychological well-being. Their fear and anxiety can significantly alter their decisions. Both patients and physicians can be affected by fear and anxiety when making these decisions.

Influence on Decision Making

Fear and anxiety are related emotions that can influence decision making in multiple ways. Two potential formulations for the role of anxiety are that (1) anxiety and fear about risks alter the evaluation process (such as probability assessments) and (2) anxiety and fear lead to seeking relief from the state. There appears to be a curvilinear relationship between escalating anxiety and performance. Under this conception, anxiety is emotional arousal, and it places a load on central cognitive processing, so that anxious decision makers evaluate evidence differently than nonanxious ones. At low levels, arousal can improve task performance, likely by recruiting additional cognitive resources, initiating coping strategies, and increasing motivation for success. However, when arousal becomes sufficiently high to be appreciable anxiety and fear, it then exceeds the cognitive analytic capacities and leads to greater use of problem simplification. This is most problematic if the decision maker has limited information, as many patients do, or if one has many complex problems and uncertain factors to consider, as many physicians do.

Additionally, immediate strong (negative) emotions (i.e., "hot states") can overwhelm cognitive goals and affect the way future dispassionate risks (i.e., "cold states") are evaluated. Initial, primitive reactions to personally relevant information consist of a rudimentary "good versus bad" interpretation. Fearful reactions to risk have been shown to cause decision making to diverge from cognitive-based assessments of risk. Anxiety is formulated as a psychic-physiologic state that one is highly motivated to alleviate and from which one wishes to return to a nonanxious, or less anxious, baseline.

Influence on Medical Decision Making

In the field of medicine, anxious individuals make decisions to alleviate existing anxiety states as well as to avoid new situations that cause anxiety. Although statistical odds might indicate that continuing watchful waiting, in lieu of initiating a risky treatment, is advisable at an early stage of a disease, patients and physicians may fear the consequences of not treating so acutely that the evidence-based statistical guidelines are overruled. Likewise, patients may avoid indicated treatment due to the anxiety that it evokes. Thomas Denberg and colleagues' investigation of men's treatment choices for localized prostate cancer yielded many cases where patients considered risky surgery as "dreadful" and associated with likely death.

William Dale and Joshua Hemmerich's work on watchful-waiting scenarios supports George Loewenstein's hypothesis that the vividness of an anxiety-provoking outcome increases the emotional response to like situations and changes physician behavior. Dale and Hemmerich investigated how a preceding premature abdominal aortic aneurysm rupture can influence vascular surgeons' and older adults' subsequent decisions about the timing of surgery. They found that experiencing a rupture during watchful waiting accelerated people's decision to end watchful waiting, even when statistical guidelines suggest patients should continue with watchful waiting. Laboratory-based follow-up studies show that participants in the simulation are significantly anxious following the rupture.

Detecting the presence of anxiety does not complete the task of explaining its influence on decision making. The locus of anxiety is another key determinant of how it will affect treatment decisions. Fear and anxiety can be tied to erroneous beliefs or realistic, well-founded concerns. The fear of treatment can be as influential as the fear of a disease, and it is possible that decision makers have multiple and potentially conflicting worries, anxieties, and fears. One must understand the specific sources of fear and anxiety if one is to intervene and manage the behavior they influence.

Another difficulty is that people have a poor appreciation for how emotions such as fear and anxiety can alter their decision making about the future; to put it succinctly, people are poor affective forecasters. People poorly predict what they would do when placed in a state of anxiety as an impending dreaded event approaches. It is important for medical-decision-making researchers to know what patients and physicians are afraid of and how afraid they are of it. Attempts to model treatment decisions where uncertainty and risk are involved will likely be inaccurate unless anxiety is appropriately incorporated into the model.

Fear and anxiety are underappreciated influences on medical decisions. Anxiety causes patients and physicians to make different choices about risky, uncertain decisions. Anxiety can distort decision makers' ideas of risk and valuation of possible options, and it is also a psychophysiological state that people take steps to avoid. Many medical decisions involve dreaded potential outcomes that provoke fear, leading to the avoidance of those situations. Understanding this influence is important for implementing evidence-based recommendations in practice.

William Dale and Joshua Hemmerich

See also Decision Making and Affect; Decision Psychology; Emotion and Choice; Mood Effects

Further Readings

Becker, G. S. (1976). *The economic approach to human behavior*. Chicago: University of Chicago Press.

Berns, G. S., Chappelow, J., Cekic, M., Zink, C. F., Pagnoni, G., & Martin-Skurski, M. E. (2006). Neurobiological substrates of dread. *Science, 312,* 754–758.

Dale, W., Bilir, P., Han, M., & Meltzer, D. (2005). The role of anxiety in prostate carcinoma: A structured review of the literature. *Cancer, 104,* 467–478.

Dale, W., Hemmerich, J., Ghini, E., & Schwarze, M. (2006). Can induced anxiety from a negative prior experience influence vascular surgeons' statistical decision-making? A randomized field experiment with an abdominal aortic aneurysm analog. *Journal of the American College of Surgeons, 203,* 642–652.

Denberg, T. D., Melhado, T. V., & Steiner, J. F. (2006). Patient treatment preferences in localized prostate carcinoma: The influence of emotion, misconception, and anecdote. *Cancer, 107,* 620–630.

Gilbert, D. T., & Ebert, J. E. J. (2002). Decisions and revisions: The affective forecasting of changeable outcomes. *Journal of Personality and Social Psychology, 82,* 503–514.

Hastie, R., & Dawes, R. M. (2001). *Rational choice in an uncertain world: The psychology of judgment and decision making*. Thousand Oaks, CA: Sage.

Kahneman, D., & Tversky, A. (2000). *Choices, values, and frames*. New York: Cambridge University Press.

Loewenstein, G. (1996). Out of control: Visceral influences on behavior. *Organizational Behavior and Human Decision Processes, 65,* 272–292.

Loewenstein, G. F., Hsee, C. K., Weber, E. U., & Welch, N. (2001). Risk as feelings. *Psychological Bulletin, 127,* 267–286.

Fixed Versus Random Effects

The terms *fixed* and *random* are commonly used in the regression modeling literature and pertain

to whether particular coefficients in a model are treated as fixed or random values. A statistical model is classified as a fixed effects model if all independent variables are regarded as fixed, a random effects model if all independent variables are regarded as random, and a mixed effects model if the independent variables constitute a mix of fixed and random effects. Analytic methods vary depending on the model. The approach selected depends on the nature of the available data and the study objectives.

A fixed variable is one that is assumed to be measured without error. The values of the fixed variable from one study are assumed to be the same as the values in any attempted replication of the study; that is, they are the only levels of a factor that are of interest (hence the term *fixed*). Gender and marital status are examples of fixed variables because they have a small fixed number of categories (levels). There is no larger population of gender categories that the levels male and female are sampled from. Fixed effects regression and analysis of variance (ANOVA) refer to assumptions about the independent variable and the error distribution. The independent variables are assumed to be fixed, and the generalization of results applies to similar values of the independent variable in the population or in other studies.

A random variable is one whose levels are assumed to be a random sample from a larger population of levels for that variable. Subjects, hospitals, physicians, schools, and litters are examples of random factors since investigators usually want to make inferences beyond the particular values of the independent variable that were captured to a larger population. Designation of variables as fixed or random is not always straightforward. Some basic questions an investigator should ask are the following: (a) Is it reasonable to assume that the levels of an independent variable were randomly sampled from some population? (b) Is the goal to make inferences to a population from which the levels of the variable were selected or from the particular levels on hand? Treatments or drug doses from a clinical trial are usually considered fixed variables since they represent all levels of interest for a study; however, they can be considered as random if their levels are a subset of the possible values one wants to generalize to.

Random effects models are referred to as variance component models, hierarchical linear models, multilevel regression models, nested models, generalized linear mixed models, and random coefficient or mixed models (using both fixed and random effects). These models can be considered as extensions of linear models and have gained popularity with advances in computing and software availability.

Models

The underlying goal of much clinical research is to evaluate relationships among a set of variables. In an experiment, a change, or experimental condition, is introduced (the independent variable) to a subject or some experimental unit, and the effect of this change is studied on a characteristic of the subject (the outcome, dependent, or response variable). An experimental condition can be a treatment or combination of treatments or factors. Multiple factors are considered in the experimental design, such as the levels of treatment or experimental condition, patient population and selection of patients, assignment of treatment condition, and the response variable of interest.

A linear statistical model where the response variable (Y_i) is modeled as a function of the independent variables ($X_1, X_2, \ldots, X_k$) is given below:

$$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k + \varepsilon_i,$$

where β_0, the intercept term, is a constant. $X_1, X_2, \ldots, X_k$ are fixed variables assumed to be observed without error. The β parameters are fixed effects of treatment or experimental condition on response and are regarded as constant, although unknown. The response variable is subject to error (denoted by ε_i) and is most often, but not necessarily, assumed to be normally distributed with zero mean and constant variance, σ^2. It represents unexplained variation in the dependent variable. The error terms are assumed to be uncorrelated for different subjects. The unknown parameters $\beta_0, \beta_1, \beta_2, \ldots, \beta_k$ characterize the relationship and are estimated from this equation to provide the best fit to the data. The method of least squares is used to obtain the best-fitting model. This is done by minimizing the sum of squares of the distances between the observed responses and those given by

the fitted model. The least squares estimator is unbiased regardless of whether the error distribution is normally distributed or not. When the error distribution is normally distributed, the least squares estimates are equivalent to maximum likelihood estimates. The independent variables are also sometimes referred to as regressors, explanatory variables, exogenous variables, and predictor variables.

In a random effects model, an independent variable with a random effect has an infinite set of levels (a population of levels). The levels present in a study are considered a sample from that population. This induces random variation between subjects or experimental units. An investigator's interest is in drawing inferences that are valid for the complete population of levels. A specific example is patients treated in a multicenter study whereby a sample of hospitals across a region is studied as opposed to all hospitals in that region. The goal is to make inference on the population of hospitals from which the sample was drawn. This is a two-level data structure, with patients at Level 1 and hospitals at Level 2. In this setting, there are two kinds of random variation that need to be accounted for: (1) that between patients within a hospital and (2) that between different hospitals. Extending the notation from above, the random effects model that can account for the variability due to a single center can be expressed as follows:

$$Y_{ij} = \beta_0 + \beta_1 X_{1j} + \beta_2 X_{2j} + \ldots + \beta_k X_{kj} + t_j + \varepsilon_{ij},$$

where β_0 is an intercept that applies to all patients in the study, t_j is a random quantity for all patients in the jth hospital, and ε_{ij} is a random quantity for the ith patient in the jth hospital. In this model, it is assumed that the error and random effects (t_j) are independent and normally distributed with zero mean and constant variance, σ_j^2 and σ_e^2, respectively. Therefore, the residual variance is partitioned into two components: (1) a between-hospital component, the variance of the hospital-level residuals that represent unobserved hospital characteristics that affect patient outcomes, and (2) a within-hospital component, the variance of the patient-level residuals. The additional term for the random effect is what distinguishes this model from the ordinary regression model described earlier. If there is no hospital-to-hospital variation, then the parameter estimates from the random effects model will be identical to those from the ordinary regression model. Inferences may be made on the fixed effects, random effects, or variance components using either least squares or maximum likelihood estimation and likelihood ratio tests. This model can also be fit with data from a repeated measures or longitudinal design, where random variation may be due in part to multiple measurements recorded on a single experimental unit or multiple measurements taken over time.

There are many extensions to the basic random effects model outlined above, including random intercept, random slope, nested, cross-classified, and generalized linear mixed models. A random intercept model would allow for the intercept term in the regression equation to vary randomly across hospitals (or higher-level units). The effects of the independent variables are assumed to be the same for each hospital. In this setting, a plot of the predicted hospital regression lines would show parallel lines for each hospital. If the assumption that the effects of explanatory variables are constant across hospitals does not hold, then one can fit a random slope model (also referred to as a random coefficient model), where the hospital prediction lines can have different slopes. A nested random effects model would be fit with data from three levels. For example, suppose one was interested in studying the outcomes of patients treated by surgeons in hospitals. If it is unreasonable to assume that the data are truly hierarchical, or nested—that is, if surgeons typically operate at more than one hospital—then surgeons and hospitals are nonnested. A cross-classified random effects model can be fit with an additional random effect for surgeon included in the model.

If the response or outcome is binary, the methods are somewhat less well developed and computationally more burdensome than for normally distributed data, primarily due to the lack of a discrete multivariate distribution analogous to the multivariate normal. An extension of the random effects model described above, proposed by Breslow and Clayton and by Wolfinger and O'Connell, which can accommodate random effects for a logistic model, can be used. Such a model is referred to as a generalized linear mixed model. Complex algorithms are required for estimation of the fixed and random effects; hence,

these models are computationally burdensome and may be impracticable in some settings. For binary outcomes, a common estimation procedure is the quadrature method using numerical approximations. Different adaptations for binary data have been presented in the literature, such as those of Breslow and Clayton; Wolfinger and O'Connell; Stiratelli, Laird, and Ware; and Zeger and Karim.

Katherine S. Panageas

See also Logistic Regression; Ordinary Least Squares Regression; Report Cards, Hospitals and Physicians; Risk Adjustment of Outcomes

Further Readings

Breslow, N. E., & Clayton, D. G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association, 88,* 9–25.

Kleinbaum, D. G., Kupper, L. L., & Muller, K. E. (1988). *Applied regression analysis and other multivariate methods*. Belmont, CA: Duxbury Press.

Searle, S. R., Casella, G., & McCulloch, C. E. (1992). *Variance components*. New York: Wiley.

Snijders, T., & Bosker, R. (1999). *Multilevel analysis*. London: Sage.

Stiratelli, R., Laird, N., & Ware, J. H. (1984). Random effects models for serial observations with binary response. *Biometrics, 40,* 961–973.

Verbeke, G., & Molenberghs, G. (2000). *Linear mixed models for longitudinal data*. Springer Series in Statistics and Computing. New York: Wiley.

Wolfinger, R., & O'Connell, D. (1993). Generalized linear mixed models: A pseudo-likelihood approach. *Journal of Statistical Computation and Simulation, 48,* 233–243.

Zeger, S. L., & Karim, M. R. (1991). Generalised linear models with random effects: A Gibbs sampling approach. *Journal of the American Statistical Association, 86,* 79–102.

Frequency Estimation

Frequency estimation is a judgment task in which one conceptualizes and conveys the anticipated likelihood of an event. It is often used to measure perceptions of personal risk of disease or benefits of treatment in quantitative terms and is therefore an important component of medical decision making. In this entry, a frequency format is distinguished from other formats used to present probabilistic information, the skills needed to estimate frequency are highlighted, and the following pertinent issues related to frequency estimation are discussed: (a) the reasoning strategies used to estimate frequency, (b) the biases associated with frequency estimation, and (c) the importance of response scale and format in frequency estimation.

Frequency Format

A frequency format is one way to represent a probabilistic statement. Other formats commonly used to represent the likelihood of an event are a percentage format (with a range of 0–100%) and a probability format (with a range of 0.0–1.0). Frequency estimation requires consideration of both a numerator (the anticipated number of times the event will occur) and a denominator (the total number of times at risk for the event to occur). Representing risk in a frequency format may be a more intuitive way to communicate risk information for certain types of judgment tasks than using other probability formats.

Needed Skills

Accurate frequency estimation requires some knowledge about the outcome being estimated and the ability to understand probabilistic information. Accurate frequency estimation also requires skills in numeracy, including a conceptual understanding of the concepts of probability. People are often inaccurate in frequency estimates of the likelihood of their developing or dying from a given disease or the benefit of a given treatment. For example, women tend to overestimate their personal risk of dying from breast cancer. In contrast, smokers tend to underestimate their risk of dying from lung cancer.

Types of Reasoning Used

There are two general types of reasoning used in frequency estimation: deliberative reasoning and experiential reasoning. In deliberative reasoning, people will attempt to integrate knowledge of relevant probabilities in formulating an estimation of

frequency. In experiential reasoning, people will rely to a greater degree on intuition, emotion, and affect in formulating an estimate of frequency. One aspect of experiential reasoning is use of the availability heuristic. The availability heuristic incorporates personal experience and exposure to the outcome in question in making a frequency estimate. The use of a pictograph with a spatial array to convey frequency information has been found to decrease the bias that can be associated with anecdotal information presented alongside frequency information in the context of a medical decision. Frequency estimates may also be influenced by optimistic bias, which reflects people's tendency to view themselves as being at lower risk than others. One theory that explains how people formulate frequency estimates is fuzzy-trace theory. Fuzzy-trace theory holds that people will naturally conceptualize frequency estimates in the most general way possible in order to solve a problem or make a decision.

Importance of Response Scale and Format

Numeric estimates of frequency are influenced by additional factors including the magnitude of the risk assessed, the response scale used, and whether the frequency estimate is made in isolation or in comparison with other risks. There is a tendency to overestimate small-frequency occurrences and to underestimate large-frequency occurrences. One approach to assist people with estimates of small frequencies is the use of a scale that has a "magnifying glass" to represent probabilities between 0% and 1% on a logarithmic scale or to use other response scales with greater discrimination among smaller probabilities. The choice of response scale can influence the magnitude of the frequency estimates assessed. Specifically, frequency estimates have been found to differ when using a percentage versus frequency format scale. Frequency estimation can also be assessed using a scale with a $1/X$ format, with an increasing value of X indicating a lower frequency. However, the $1/X$ format has been found to be a more difficult format for judgment tasks in which a person is asked to compare risk magnitudes. In frequency judgments, people may find the task easier and be more accurate when comparing their risk with that of others versus providing a frequency estimate for their risk of a given outcome in isolation.

Conclusion

Frequency estimation is a judgment task that conveys perceptions of risk using quantitative terms. Accurate frequency estimation involves some knowledge as well as numeric skills, including knowledge of the concepts of probability. Frequency estimation is an important aspect of risk communication and decision making. However, when assessing frequency estimates or conveying frequency information, one must be cognizant of the role of critical and experiential reasoning in frequency estimation as well as the biases associated with response scales and numeric and graphic formats used to convey probabilistic information.

Marilyn M. Schapira

See also Biases in Human Prediction; Decision Making and Affect; Numeracy; Risk Perception

Further Readings

Ancker, J. S., Senathirajah, Y., Kukafka, R., & Starren, J. B. (2006). Design features of graphs in health risk communication: A systematic review. *Journal of the American Medical Informatics Association, 13,* 608–618.

Fagerlin, A., Wang, C., & Ubel, P. A. (2005). Reducing the influence of anecdotal reasoning on people's health care decisions: Is a picture worth a thousand statistics? *Medical Decision Making, 25,* 398–405.

Hoffrage, U., & Gigerenzer, G. (1998). Using natural frequencies to improve diagnostic inferences. *Academic Medicine, 73,* 538–540.

Lichtenstein, S., Slovic, P., Fischhoff, B., Layman, M., & Combs, B. (1978). Judged frequency of lethal events. *Journal of Experimental Psychology: Human Learning and Memory, 4,* 551–579.

Reyna, V. F., & Adam, M. B. (2003). Fuzzy-trace theory, risk communication, and product labeling in sexually transmitted diseases. *Risk Analysis, 23,* 325–342.

Slovic, P., Monahan, J., & MacGregor, D. G. (2000). Violence risk assessment and risk communication: The effects of using actual cases, providing instruction, and employing probability versus frequency formats. *Law and Human Behavior, 24,* 271–296.

Tversky, A., & Kahneman, D. (1982). Availability: A heuristic for judging frequency and probability. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (Chap. 11, pp. 163–178). Cambridge, UK: Cambridge University Press.

Frequentist Approach

The frequentist (or classical) approach is a branch of statistics that currently represents the predominant methodology used in empirical data analysis and inference. Frequentist statistics emerged as a prevailing method for inference in the 20th century, particularly due to work by Fisher and, subsequently, by Neyman and Pearson. Given that distinct differences exist between the research conducted by these authors, however, frequentist inference may also be subcategorized as being either Fisherian or Neyman-Pearson in nature, although some view the Fisherian approach to be a distinct philosophy apart from frequentist statistics altogether.

Frequentist methods are often contrasted with those of Bayesian statistics, as these two schools of thought represent the more widely considered approaches through which formal inference is undertaken to analyze data and to incorporate robust measurements of uncertainty. Although frequentist and Bayesian statistics do share certain similarities, important divergence between the approaches should also be noted. In this context, the central tenets that differentiate the frequentist paradigm from other statistical methods (e.g., Bayesian) involve (a) the foundational definition of probability that is employed and (b) the limited framework through which extraneous information (i.e., prior information) is assessed from sources outside of the immediate experiment being conducted. Ultimately, these characteristics affect the breadth of research design and statistical inference. By formally focusing primarily on data that emanate from an immediate experiment being conducted (e.g., a randomized clinical trial) and not on additional sources of information (e.g., prior research or the current state of knowledge), results of a frequentist analysis are essentially confined to an immediate study. Reliance is thus often placed on a more informal process to consider extraneous data from sources beyond the immediate study. Although this issue has, in part, led to its theoretic appeal among both regulatory agencies and scientists as being an "objective" method of inference (e.g., the conclusions of a single study do not allow for other findings to affect statistical inference), the frequentist approach has also been viewed as lacking a full rigor that parallels the comprehensive aspects of scientific inquiry and decision theory. Despite a lengthy debate concerning these philosophical issues, the frequentist approach remains the most commonly used method of statistical inquiry. When correctly applied and interpreted, frequentist statistics also represent a robust standard for point estimation, interval estimation, and statistical/hypothesis testing. Consideration of the frequentist approach is additionally important when addressing the overall study design, sample size calculations, and effect sizes.

Within the frequentist paradigm, *probability* is defined as a long-run expected limit of relative frequency within a large number of trials or via a *frequency concept of probability* that denotes the proportion of time when similar events will occur if an experiment is repeated several times. Hence, classical statistical analysis and inference yields interpretations only within a context of repeated samples or experiments. While the theory of infinitely repeatable samples may be viewed as a largely hypothetical issue for an analyst (i.e., because researchers typically obtain only one random draw from a population), the concept becomes of fundamental importance in interpreting results within the frequentist paradigm. Furthermore, the assumption of infinite repeated samples imparts asymptotic properties (e.g., the law of large numbers, convergence, the central limit theorem), which are required for robust inference under the frequentist approach.

Samples and populations are key concepts in frequentist statistics. Researchers use frequentist analysis and inference to generalize findings from a given sample to a broader population. In this context, research questions often focus on obtaining a point estimate, interval estimate, or statistical/hypothesis test concerning a population parameter whose value is assumed to be both fixed and unknown.

Point Estimation

Point estimation is undertaken to find a statistic that is calculated from the sample data and ultimately used for inference concerning the fixed, unknown population parameter. A common nomenclature for this research question involves denoting the population parameter θ and its estimator statistic $\hat{\theta}$. Importantly, the frequentist paradigm defines

an *estimator* of the population parameter, $\hat{\theta}$, as a random variable that provides inference concerning the fixed, unknown population parameter θ under the assumption that infinite random samples are drawn from the population itself. The exact value that an estimator $\hat{\theta}$ takes for any given sample is termed an *estimate*. Procedures for obtaining point estimations of population parameters include methods of moments (MoM), maximum likelihood estimation (MLE), and ordinary least squares (OLS), among others. Also contingent on the assumption of infinite random sampling, a theoretical *sampling distribution* (i.e., the probability distribution of a statistic under repeated sampling of the population) exists for the estimator, $\hat{\theta}$, from which a researcher ultimately obtains a random draw. Figure 1 graphically presents the concept of an estimator, a likelihood function that may be estimated via maximum likelihood, and a sampling distribution that may be represented via infinitely repeated samples.

Robust results in frequentist point estimation are produced with minimal bias when the expected value of an estimator $\hat{\theta}$ equals the value of the population parameter, θ, via infinite repeated sampling. For an estimator to be deemed *unbiased*, the mean value of the sampling distribution would be equal to that of a true population parameter. In frequentist statistical theory, emphasis is placed on obtaining unbiased estimators because it is under these conditions that these statistics equal that of a true population parameter, specifically when its average value is found across an infinite random sampling of that population. Given that considerable difficulties may emerge in calculating a sampling distribution, it is also common to rely on asymptotics to approximate infinite sampling requisites in frequentist statistics.

Interval Estimation

In addition to point estimation, researchers often seek to obtain a *confidence interval* (CI) (i.e., a range of values represented by a lower and an upper bound) of estimators that have an a priori (i.e., "before the fact") probability of containing the true value of the fixed unknown population parameter. Based on a given a priori significance level chosen for an analysis, α, the straightforward "$(1 - \alpha)\cdot 100\%$" CI defined for a population parameter θ is

$$\Pr(L_{\text{bound}} < \theta < U_{\text{bound}}) = 1 - \alpha,$$

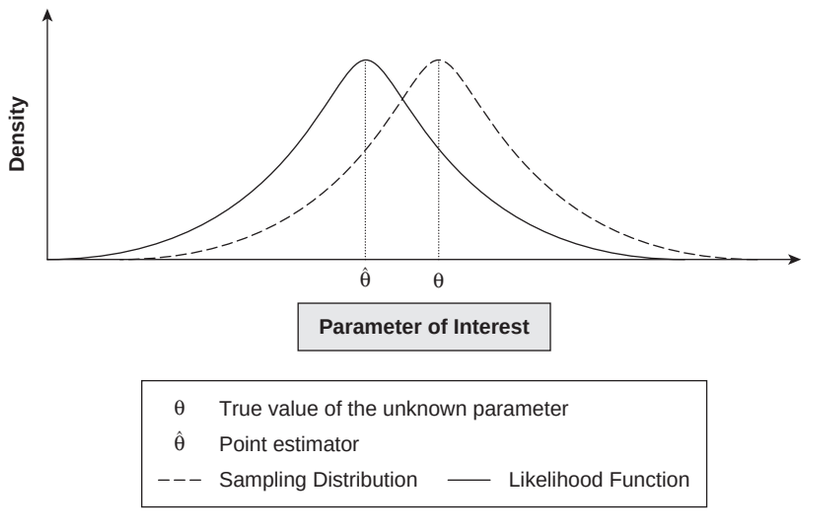


Figure 1 Frequentist sampling distribution

Sources: Skrepnek, G. H. (2007). The contrast and convergence of Bayesian and frequentist statistical approaches in pharmacoeconomic analysis. *PharmacoEconomics, 25,* 649–664. Kennedy, Peter. *A Guide to Econometrics, fifth edition,* figure 13.1, p. 231, © 2003 Peter Kennedy, by permission of The MIT Press.

where Pr denotes probability; L_{bound} and U_{bound} are the lower and upper bounds of the CI, respectively; θ is the population parameter being estimated; and α is the a priori significance level chosen for the analysis. Under conditions wherein a sampling distribution is approximately normally distributed, the CI is

$$\hat{\theta} = \pm c \times SE(\hat{\theta}),$$

where $\hat{\theta}$ is the coefficient estimate, c is the critical value obtained from a t or Z table (depending on sample sizes and adherence to a level of confidence and degrees of freedom), and $SE(\hat{\theta})$ is the standard error of the mean, equal to the standard deviation divided by the square root of the sample size. While the most typical CI in frequentist analyses is 95%, other CIs may be calculated for 90%, 99%, or 99.9%. In instances of large sample sizes, the critical values for 90%, 95%, 99%, and 99.9% CIs are approximately 1.645, 1.96, 2.58, and 3.27, respectively, from the standard normal distribution table (i.e., Z table). Thus, under the condition of a large sample size, the 95% CI would be

$$\hat{\theta} \pm 1.96 \times \sigma/-\sqrt{n},$$

where $\hat{\theta}$ is the coefficient estimate, 1.96 is the critical value for a 95% CI, and $\sigma/-\sqrt{n}$ is the standard deviation divided by the square root of the sample size (i.e., standard error of the mean). Figure 2 presents a graphical depiction of a 95% CI for a normal sample distribution whose mean value of the point estimate is 0.

An area of concern among researchers involves the correct interpretation of a CI. Importantly, it is *incorrect* to infer a probability statement concerning a calculated interval itself in that it might be a "probability that the true value of a parameter is contained within its lower and upper bounds." Rather, CIs are correctly interpreted in terms of a certain percentage of the intervals (e.g., 95%) that will contain the true parameter in the long run. Thus, for example, a 95% CI is properly presented as representing 95% of the intervals derived from infinite sampling of the underlying population of interest that would include the true value of the fixed unknown population parameter. The rationale

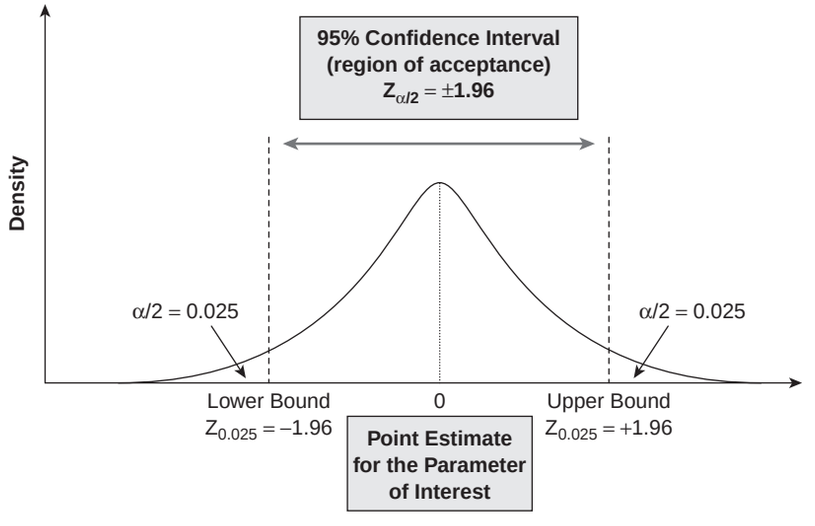


Figure 2 95% confidence interval

Source: Gujarati, D. N. (1995). *Basic econometrics* (3rd ed.). New York: McGraw-Hill. Reproduced with permission of The McGraw-Hill Companies.

behind this interpretation is that the probability level (e.g., .05) of a CI refers to the interval itself and not to the parameter, because the parameter is a fixed unknown and not a random variable. Furthermore, the lower and upper bounds of a CI are considered random only prior to sampling of the population. After sampling, the explicit values obtained for the CI are not random, and thus, it does not have a probability associated with it.

Several methods are available for calculating CIs, each of which may be appropriate for a particular sampling distribution. Researchers, for example, have developed methods to obtain exact CIs for linear functions of a normal mean and variance. Given the extensive requirements to calculate exact CIs, approximate CIs are often used wherein a strong reliance is placed on the assumptions of the law of large numbers and the central limit theorem.

Significance and Hypothesis Testing

Whereas Fisher developed and popularized *significance testing* to weigh evidence against a given hypothesis, Neyman and Pearson developed *hypothesis testing* as a method to assess two directly competing hypotheses. Central to these inferential techniques is the assessment of whether the findings observed within a given experiment were based on chance alone. Additionally, central to both significance and hypothesis testing is the concept of the null hypothesis H_0, which is used to describe the lack of a treatment effect. Importantly, frequentist approaches can never "accept" the null hypothesis. Rather, research can either "reject" or "fail to reject" the null, suggesting that a treatment effect either was or was not observed, respectively. The rationale for this decision rule concerning the null is one of rigorous scientific method—if an investigation fails to reject a null hypothesis, it cannot necessarily be concluded that the null is true under all circumstances.

Significance and hypothesis testing both require that an appropriate test statistic (e.g., *t* test, *F* test, regression) be employed to summarize the sample data relevant to the research hypothesis being evaluated. CIs are also related to hypothesis testing in that if the CI does not include a null hypothesis, then a hypothesis test will reject the null, and vice versa. Although significance and hypothesis testing are closely related, differences do exist between the concepts, so the two are not synonymous. Despite this, the term *hypothesis testing* is routinely used to describe the general process of testing for the presence of a treatment effect.

Initially, Fisher developed significance testing to assess the *direct probabilities* (i.e., changes in observed data $\hat{\theta}$ within an immediate experiment leading to rejection of a hypothesis H, or $\Pr(\hat{\theta}|H)$) rather than relying on the *indirect probabilities* (i.e., the probability of a hypothesis given observed data, or $\Pr(H|\hat{\theta})$). Neyman-Pearson hypothesis testing built on Fisher's work by explicitly formalizing the specification of a rival alternate hypothesis H_A, which had only been indirectly addressed in frequentist statistics until that point. The specification of an alternate hypothesis allowed Neyman and Pearson to formalize issues concerning sample size, power, and effect size. This occurred, in part, because the concepts of Type I and Type II errors complemented the development of a formal rival hypothesis against the null.

Type I and Type II errors involve the potential of either incorrectly rejecting or incorrectly failing to reject a null hypothesis, respectively, and are concepts that play an important role in the broader design and interpretation of experiments. The probability of committing a Type I error, represented as α, is the probability of a statistical test to incorrectly reject a null hypothesis when the null is actually true (i.e., committing a false positive). Conversely, the probability of a Type II error, denoted by β, is the probability of a statistical test to incorrectly fail to reject a null hypothesis when the null is actually false (i.e., committing a false negative). The *power of a test*, calculated as $1 - \beta$, is defined as the probability of rejecting a false null hypothesis when the null is actually false (i.e., a correct decision) or, stated differently, the ability of a test to detect a statistical relationship. In practice, the power of a test is often calculated prior to conducting an experiment to ascertain sufficient sample sizes. Alternatively, post hoc power analyses may be computed to determine if a sufficient sample size had been obtained and to determine effect sizes for interpretation of a study's results. Beyond establishing if the treatment effect is statistically significant, *effect sizes* are measures that

represent the actual magnitude of a treatment effect. In describing the relationships between hypothesis testing, Type I and Type II errors, and power, Figure 3 graphically presents these aforementioned concepts relating to Neyman-Pearson hypothesis testing.

Importantly, the probabilities of committing a Type I or Type II error are inversely related (i.e., the smaller the probability of one, the higher the probability of the other). Thus, the smaller the significance level specified for an investigation, the greater the probability of failing to reject a false null hypothesis. As such, the researcher must weigh the importance of protecting from committing a false positive versus a false negative when establishing an appropriate significance level versus the power of a test. Depending on the research question being addressed, either a Type I or a Type II error may be considered to be the most important to avoid. To illustrate, a Type I error would occur if a research study concluded that the treatment was observed to yield a statistically significant effect compared with the placebo control when, in reality, there was no difference between them. Conversely, a Type II error would occur if no difference was observed in the study when a difference actually existed. Committing a Type I error in this instance concerning efficacy may result in the use of an ineffective therapy, while a Type II error would suggest that a potentially efficacious therapy would not be used. If the research question involved safety, however, it would be crucial to minimize the potential of committing a Type II error (i.e., suggesting that safety existed when it actually did not) rather than a Type I error. Pragmatic methods to reduce the probability of incurring either type of error emphasize following a robust study design with appropriate sample sizes. Figure 4 presents a graphical depiction of the relationship between a Type I error, a Type II error, and the power of a test for distributions of a null hypothesis H_0 and an alternate hypothesis H_A—noting that shifting the critical value ultimately affects each of the representative

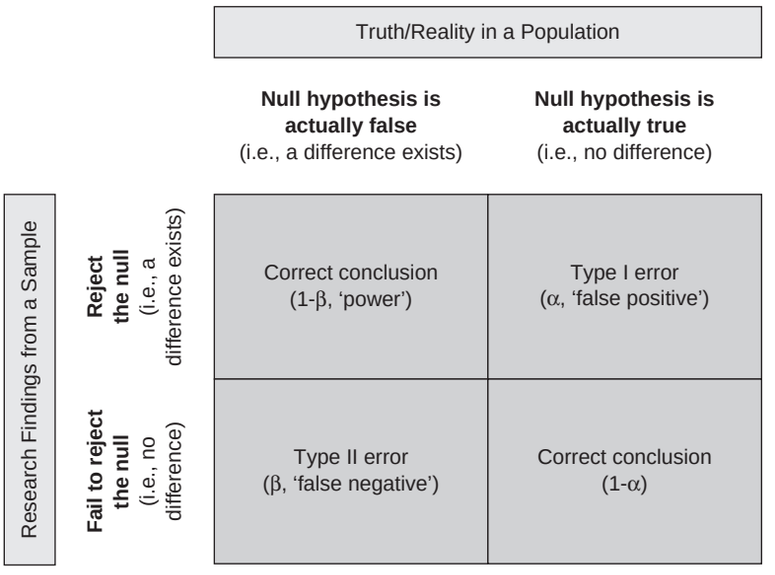


Figure 3 Neyman-Pearson hypothesis testing

Sources: From Hays, W. L. *Statistics,* 5th ed. © 1994 Wadsworth, a part of Cengage Learning, Inc. Reproduced by permission. www.cengage.com/permissions; Young, R. K., & Veldman, D. J. *Introductory Statistics for the Behavioral Sciences*, 4th ed. © 1981 Wadsworth, a part of Cengage Learning, Inc. Reproduced by permission. www.cengage.com/permissions.

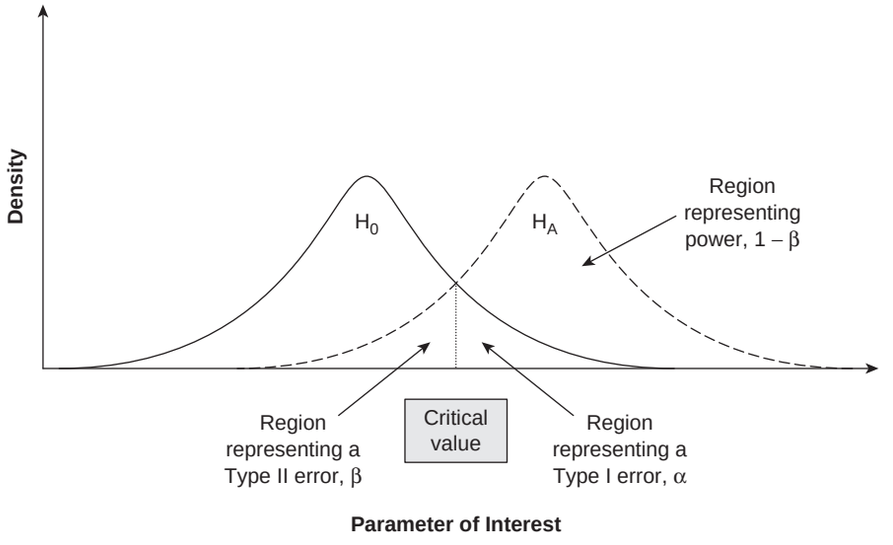


Figure 4 Type I and Type II errors in hypothesis testing

Source: From Hays, W. L. *Statistics,* 5th ed. © 1994 Wadsworth, a part of Cengage Learning, Inc. Reproduced by permission. www.cengage.com/permissions.

regions. Additionally, Figure 5 illustrates the concept of the inverse relationship between Type I and Type II errors according to the probability of rejecting a null hypothesis versus the ratio of the true variances among two populations, σ_x^2 / σ_y^2, noting again that power is the probability that a false null hypothesis will actually be rejected. Herein, the probability of rejecting the null increases when the probability of committing a Type I error, α, increases and the probability of a Type II error, β, decreases (i.e., with increasing power 1 – β, there is an increasing ability for a statistical test to detect a difference when one truly exists).

In practice, significance and hypothesis testing both use *p* values, or probability values, to express the likelihood that results may have been observed by chance alone. A concept addressed extensively by Fisher, a *p* value may be formally defined as the probability of obtaining a test statistic at least as extreme as a calculated test statistic if a null hypothesis were true, and thus representing a measure of strength against the null itself. Stated differently, the *p* value is the probability that a result at least as extreme as that which was observed in an experiment would occur by chance alone. Notably, *p* values have been misinterpreted to be "the probability that a null hypothesis is true." Overall, a *p* value is the lowest significance level wherein a null hypothesis can be rejected.

The *p* value or a priori α level of .05 as an acceptable value for significance or hypothesis testing remains a contentious area of discussion, albeit corresponding to the most commonly chosen figure used to designate statistical significance in scientific research. Furthermore, criticism concerning the reliance on *p* values or α levels for statistical testing appears in both a theoretical and an applied context, particularly concerning their association with sample size. Additionally, adjustments of *p* values or α levels to more conservative figures may be warranted in instances of sequential, subset, or multiple-comparison analysis (e.g., via Bonferonni, Holm, Sidak, or other corrections). Beyond these debates, results from an analysis wherein *p* values are calculated to be equal to or below an a priori α level chosen for significance suggest a statistically significant relationship concerning a treatment effect that is being researched. When reporting results, analysts may choose to explicitly present an exact value of a computed *p* value that is obtained from a statistical test or, alternatively, report whether a null hypothesis

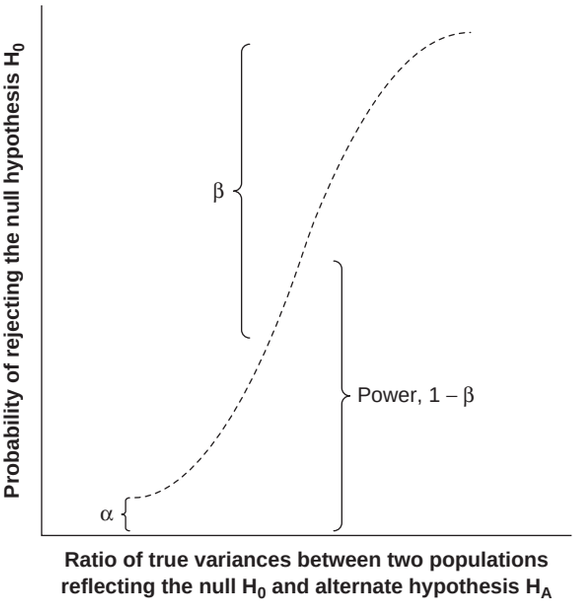


Figure 5 Relationship of the probability of rejecting a null hypothesis relative to the ratio of true variances between two populations

Source: Young, R. K., & Veldman, D. J. *Introductory Statistics for the Behavioral Sciences*, 4th ed. © 1981 Wadsworth, a part of Cengage Learning, Inc. Reproduced by permission. www.cengage.com/permissions.

is rejected based on an a priori α level chosen for statistical significance (i.e., $\alpha = .05$) and if the computed *p* value of that statistical test is below this α level.

Beyond the statistical significance of a test, assessing the *clinical significance* of a result is also warranted. To illustrate, when a statistic is found to be significant, it suggests that the statistic itself is a reliable estimate of the population parameter or that some treatment effect exists (i.e., that such a finding is unlikely due to chance alone). For example, an investigation may find a statistically significant difference of .5 mmHg between two groups. This in itself does not prove that the finding is relevant, important, or able to support final decision making. Determining clinical (or practical) significance involves assessing the broader aspects of clinical practice, the study design employed in a given investigation, and identifying the smallest magnitude of an effect that is typically associated for a clinically beneficial or harmful impact.

Conclusion

Frequentist methods currently constitute the most widely used approach to empirical data analysis and statistical inference. The hallmarks of this philosophy involve a definition of probability that emphasizes an interpretation over long-run, repeated trials and a focus on results that are confined to an immediate empirical investigation. While the foundation of frequentist statistics does allow for robust inference under several conditions, other statistical approaches may additionally offer sound frameworks with which to engage in scientific inquiry. To fully capture the positive elements of any statistical methodology, researchers must remain fully cognizant of the specific elements associated with each approach concerning appropriate application and interpretation.

Grant H. Skrepnek

See also Confidence Intervals; Hypothesis Testing; Maximum Likelihood Estimation Methods; Sample Size and Power; Statistical Testing: Overview

Further Readings

Berger, J. O., & Sellke, T. (1987). Testing a point null hypothesis: The irreconcilability of *p*-values with evidence (with discussion). *Journal of the American Statistical Association, 82,* 112–139.

Bloom, B. S., de Pouvourville, N., & Libert, S. (2002). Classical or Bayesian research design and analysis: Does it make a difference? *International Journal of Technology Assessment in Health Care, 18,* 120–126.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum.

Cowles, M., & Davis, C. (1982). On the origins of the .05 level of significance. *American Psychologist, 37,* 553–558.

Cox, D. R. (2005). Frequentist and Bayesian statistics: A critique. In L. Lyons & M. K. Unel (Eds.), *PHYSTAT 05: Proceedings of statistical problems in particle physics, astrophysics and cosmology* (pp. 3–7). London: Imperial College Press.

Fisher, R. A. (1935). *The design of experiments.* Edinburgh, UK: Oliver & Boyd.

Goodman, S. N. (1999). Toward evidence-based medical statistics: 1. The *p*-value fallacy. *Annals of Internal Medicine, 130,* 995–1004.

Hays, W. L. (1994). *Statistics* (5th ed.). Austin, TX: Harcourt Brace.

Hubbard, R., & Bayarri, M. J. (2003). Confusion over measures of evidence (p's) versus errors (α's) in classical statistical testing. *American Statistician, 57,* 171–182.

Kennedy, P. (2003). *A guide to econometrics* (5th ed.). Cambridge: MIT Press.

Neyman, J., & Pearson, E. S. (1933). On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London, Series A, Containing Papers of a Mathematical or Physical Character, 231,* 289–337.

Skrepnek, G. H. (2005). Regression methods in the empiric analysis of health care data. *Journal of Managed Care Pharmacy, 11,* 240–251.

Skrepnek, G. H. (2007). The contrast and convergence of Bayesian and frequentist statistical approaches in pharmacoeconomic analysis. *PharmacoEconomics, 25,* 649–664.

Sterne, J. A. C., & Smith, D. (2001). Sifting the evidence: What's wrong with significance tests. *British Medical Journal, 322,* 226–231.

Fuzzy-Trace Theory

Fuzzy-trace theory explains how people remember, reason, and decide. The theory has been applied to a variety of domains in health and medical decision making, including HIV prevention, cardiovascular disease, surgical risk, genetic risk, and cancer prevention and control. Within these domains, it explains the mysteries of framing effects, ratio bias, frequency-formatting effects, and base-rate neglect, among other classic phenomena. Fuzzy-trace theory has led to the discovery of new, counterintuitive effects too. For example, studies show that adolescents think about risks more logically and quantitatively than mature adults do, which, paradoxically, promotes risk taking—a surprising but predicted effect.

Fuzzy-trace theory has been applied to a variety of populations, including patients and physicians. As a developmental theory, it focuses on changes in memory, reasoning, and decision making with age (differences among children, adolescents, young adults, and the aged). It also specifies when age does not make a difference; for example, adolescents and expert physicians perform equally poorly on base-rate neglect problems involving medical diagnosis (underestimating the effects of prior probabilities of disease on subsequent probabilities once a diagnostic test result is known). Most recently, fuzzy-trace theory has been used to characterize the changes in cognition that accompany disease processes, such as in Alzheimer's and Parkinson's disease, as well as mild cognitive impairment.

The phrase *fuzzy trace* refers to a distinction between verbatim memory representations that are vivid and gist memory representations that are "fuzzy" (i.e., vague and impressionistic). The distinction between verbatim and gist representations was initially borrowed from psycholinguists, who had amassed substantial evidence for it and had applied it to the representation and retention of verbal materials. However, despite the continued use of the term *verbatim* in fuzzy-trace theory, these types of representations were extended to describe memories of nonverbal stimuli, including numbers, pictures, graphs, and events.

For example, if a physician tells a patient that she has a 22% chance of having a stroke in the next 3 years, she forms two kinds of memories for that information: (1) a memory of the precise details of what was said ("22% chance of stroke"), which fades rapidly and is subject to interference (e.g., from anxiety), and (2) a memory of the bottom-line meaning, or gist, of what was said (e.g., there is a good chance of having a stroke in the next few years). Multiple gist memories are typically encoded into memory for a single piece of information.

Research on the major paradigms of judgment and decision making and of developmental psychology have shown a common pattern of results with respect to verbatim and gist memories: Individuals encode parallel representations of information along a continuum of precision that is anchored at each end by gist and verbatim representations, or memory traces. Verbatim traces preserve veridical details at the precise end, and gist traces preserve extracted meanings and patterns at the fuzzy end. This first tenet of fuzzy-trace theory is not an assumption, in the usual sense of that term, but, rather, is based on the results of numerous experiments that tested alternative hypotheses regarding memory representations.

A second tenet of the theory, central to understanding reasoning, is the idea that retrieval of

either verbatim or gist representations is cue dependent, a conclusion that is also based on empirical evidence. That is, the two types of traces are stored separately and retrieved independently, and their successful retrieval depends on cues, or specific reminders, in the environment. As many studies have demonstrated, two different cues presented to the same individual can elicit contradictory responses about what is stored in memory (such as those found in false-memory reports). Different values and reasoning principles are retrieved from memory, depending on cues in the environment, which helps explain why reasoning and decision making are so variable.

A number of factors conspire to make gist traces the default representations used in reasoning. Verbatim traces become rapidly inaccessible and are sensitive to interference. Reasoning therefore gravitates to using gist (or fuzzy) representations, which minimizes errors due to the fragile and cumbersome verbatim representations. Moreover, this adaptive tendency to use gist representations—the fuzzy-processing preference—increases with development as individuals gain experience at a task. Studies of children (comparing older with younger children) and of adults (comparing experts with novices in a domain of knowledge) have demonstrated that reliance on gist representations increases with development. For example, a study comparing medical students and physicians varying in expertise in cardiology showed that the more expert processed fewer dimensions of information and processed it in an all-or-none (gist based) manner (i.e., patients with chest pain were seen as either requiring intensive care or safely discharged with a 72-hour follow-up).

People think using simple gist representations of information, often processing them unconsciously, and engage in parallel rather than serial processing of that information (leaping ahead based on vague gist impressions of the relations and patterns in information without fully encoding details). This kind of thinking is what is meant by "gist-based intuitive reasoning." The third tenet of the theory is that people exhibit a fuzzy-processing preference (a preference for reasoning with the simplest gist representation of a problem). This preference produces more coherent thinking (because working with gist representations is easier and less error-prone) and more positive decision outcomes (e.g., less unhealthy risk taking in adolescents). Recent research has linked gist-based intuitive reasoning to lower HIV risk. The reliance on gist as a default mode of processing is associated with more adaptive responses to risk as people mature.

From these tenets, it can easily be seen why fuzzy-trace theory's prescriptions to improve health communication and medical decision making differ from those of standard utility or dual-process theories. The goal of informed consent, for example, is to reach an understanding of the bottom-line gist of risks and benefits (e.g., of surgery) rather than to regurgitate verbatim facts. Similarly, the goal of prevention programs in public health is to inculcate rapid and unconscious recognition of the gist of risky situations and to retrieve relevant values (e.g., involving unprotected sex) rather than to consciously deliberate about the details and degrees of risk. Thus, contrary to other dual-process theories, gist-based intuition is an advanced form of thought.

Valerie F. Reyna

See also Gain/Loss Framing Effects; Risk Communication; Risk Perception

Further Readings

Reyna, V. F. (2004). How people make decisions that involve risk: A dual-processes approach. C*urrent Directions in Psychological Science, 13,* 60–66.

Reyna, V. F., & Adam, M. B. (2003). Fuzzy-trace theory, risk communication, and product labeling in sexually transmitted diseases. *Risk Analysis, 23,* 325–342.

Reyna, V. F., & Brainerd, C. J. (2008). Numeracy, ratio bias, and denominator neglect in judgments of risk and probability. *Learning and Individual Differences, 18,* 89–107.

Reyna, V. F., & Farley, F. (2006). Risk and rationality in adolescent decision-making: Implications for theory, practice, and public policy. *Psychological Science in the Public Interest, 7*(1), 1–44.

Reyna, V. F., & Lloyd, F. (2006). Physician decision making and cardiac risk: Effects of knowledge, risk perception, risk tolerance, and fuzzy processing. *Journal of Experimental Psychology: Applied, 12,* 179–195.

Reyna, V. F., Lloyd, F., & Whalen, P. (2001). Genetic testing and medical decision making. *Archives of Internal Medicine, 161,* 2406–2408.

G

Gain/Loss Framing Effects

Amos Tversky and David Kahneman's work in the 1980s on framing (presentation) effects was a stimulus for other researchers to examine how these effects affect medical decision making. Interestingly, the work by Tversky and Kahneman in framing effects was based on consideration of a transmissible infectious disease in a population.

Tversky and Kahneman's use of the term *frame* was in the arena of type of description applied to data. In its most basic sense, framing refers to the way in which medical decision making alternatives are presented. For example, in one frame, all data might be presented in terms of *survival*; in the second frame, all data could be presented in terms of *mortality*. Here, the term *framing effect* would be similar to the term *presentation effect*, where a frame is a type of presentation of data to study subjects in a research survey or research questionnaire. Presenting the data in terms of survival would be an example of gain framing; presenting the data in terms of mortality would be an example of loss framing.

Risky and Riskless Contexts

Tverksy and Kahneman describe this aspect of their work as research on the cognitive and psychophysical determinants in risky and riskless contexts. For these authors, framing refers to the cognitive point at which decision problems can be described (framed) in multiple ways, giving rise to different preferences being elicited that are dependent on the frame.

They further argue that framing effects can help explain some of the anomalies found in consumer behavior. Other researchers have extended their point to medical decision making in that caution needs to be used in deciding how decision problems are presented to patients.

Early Research in Framing

Attention to the use of data in decision making was brought into the medical-decision-making arena in a scientific article by Barbara J. McNeil, R. Weichselbaum, and S. G. Pauker appearing in the *New England Journal of Medicine* in 1978 on the fallacy of 5-year survival in lung cancer. McNeil and colleagues focused attention on the 5-year survival data in lung cancer. This article focused attention on the importance of choosing therapies not only on the basis of objective measures of survival but also on the basis of patient attitudes. However, while McNeil and colleagues derived their data from existing data on 5-year survival from the published medical literature, they did not present graphical displays of 5-year survival curves to study participants. Rather, McNeil and colleagues presented data derived from 5-year survival for lung cancer in terms of *cumulative probabilities* and *life-expectancy data* in this study.

In a subsequent article published in the *New England Journal of Medicine* in 1979, McNeil, Pauker, H. C. Sox, and Tversky asked study participants to imagine that they had lung cancer and

to choose between two therapies on the basis of either cumulative probabilities or life-expectancy data. In this study, different groups of respondents received input data that differed in the following ways: whether or not the treatments were identified (as surgery and radiation therapy) and whether the outcomes were framed in terms of the probability of living or the probability of dying. The authors found that the attractiveness of surgery as a treatment choice, relative to radiation therapy, was substantially greater (a) when the treatments were identified rather than unidentified, (b) when the information consisted of life expectancy rather than cumulative probability, and (c) when the problem was framed in terms of the probability of living (survival frame) rather than in terms of the probability of dying (mortality frame). The authors in their conclusion suggest that an awareness of such influences among physicians and patients could help reduce bias and improve the quality of medical decision making.

Yet two questions can be asked of both studies: First, how useful did the study participants find cumulative probabilities and life-expectancy data? Second, are there other forms of data displays that patients may find as useful or more useful to consider for their own choice of therapy in cases where surgery is to be contrasted with radiation therapy? In the case of Stage 3 lung cancer, surgery has a better long-term (and worse short-term) survival than radiation therapy, while radiation therapy has a better short-term (and worse long-term) survival than surgery for the same Stage 3 lung cancer.

Graphical Displays Comparing 5-Year Survival Curves

In the 1990s, Dennis J. Mazur and David H. Hickam studied graphical displays of 5-year data as survival curves. These graphical displays and comparison of survival curves out to 5 years illustrate how framing effects can be illustrated in graphical displays of data.

In medical decision making, framing has been most typically depicted in comparisons of scenarios within which the data are depicted in one of two frames, Frame 1 or Frame 2. Frame 1 depicts all outcomes in terms of survival, and the other frame, Frame 2, depicts all outcomes in terms of mortality.

The time line over which the survival data and the mortality data are typically provided to volunteers in these research studies goes from the time of the initial treatment with a medical intervention (T_0) to a time 5 years after the initial treatments (T_5). Five-year survival data are a common form of data used by physicians in oncology, and the first types of medical conditions (disease processes) studied to look for framing effects were more aggressive cancers, for example, Stage 3 lung cancer. The 5-year survival curve would not be appropriate for a cancer such as prostate cancer, where the chance of survival goes well beyond 5 years.

Underlying assumptions of the above study design include the point that patients cannot shift from one treatment to another and must remain with the treatment they choose throughout. For example, if the two time lines of treatment of Treatment 1 and Treatment 2 cross at some midpoint, the participant cannot shift from the treatment that has a better survival over the time line T_0 to the midpoint to the other treatment that has a better survival from the midpoint to the time T_5.

An example of two treatments where one treatment has a better short-term (T_0 to midpoint) survival and a worse long-term (midpoint to T_5) survival are surgery and radiation therapy for Stage 3 lung cancer. Here, with surgery, there is an initial chance of patients dying with a surgical intervention at time T_0 but a better chance of their still being alive at T_5; with radiation therapy, there is little to no chance of dying from the radiation therapy itself at time T_0, but fewer patients are alive at T_5 with radiation therapy than with surgery.

Here, the following assumptions come into play. First, there is the assumption that the patient will not die as a result of radiation therapy. In fact, a rare patient may die during radiation therapy; therefore, it is not necessarily guaranteed that all patients will survive the radiation therapy. Second, there is the assumption that the patient incurs a high risk of dying during surgery. In fact, the chance that a patient may die with a surgical intervention is highly dependent on the status of the individual patient's cardiovascular and respiratory condition. (The risk of dying from anesthesia is built into the surgical death rate, as a surgery cannot be performed without anesthesia.)

In a typical framing study in medical decision making, study participants are randomized to one

of two frames, Frame 1 or Frame 2. Participants randomized to Frame 1 receive all data in terms of survival (number of patients alive after treatment); participants randomized to Frame 2 receive all data in terms of mortality (number of patients dead after each treatment).

In the typical verbal study of framing effects, the participant is provided information about the number of patients being alive after the initial intervention (surgery and radiation therapy) at T_0 and the number of patients still alive after surgery and radiation therapy 5 years later at T_5.

Researchers present data to the study participants in two distinct frames. Frame 1 depicts data to participants only in terms of survival (the chance of still being alive after the initial intervention and the chance of still being alive 5 years after that intervention); Frame 2 depicts data to participants only in terms of mortality (the chance of dying during the initial intervention and the chance of dying within 5 years after that intervention).

Framing Effects in Tabular and Graphic Displays

The initial framing study of McNeil and colleagues provided study participants with frames presented in terms of quantitative descriptions of data (cumulative probabilities and life-expectancy). Since that time in research on human study participants, such effects with framing have been demonstrated using tabular and graphic expressions of chance (likelihood) as well, where tabular or graphical expressions of chance or likelihood depict the survival and mortality data the participants are asked to consider.

The depiction of data in terms of *words and numbers*, as contrasted with *graphical data displays*, brings into play the following considerations. When attempting to see if framing effects are present in a particular study where data are provided to participants in terms of words and numbers, the researchers may simply provide participants with the number of patients alive after the surgical or radiation therapy interventions at time T_0 and time T_5 and ask each participant to choose which treatment he or she prefers.

However, if data are provided to participants in terms of graphical comparisons of two 5-year survival curves (one 5-year survival curve for surgery and one 5-year survival curve for radiation therapy), the researchers are providing patients with much more data than simply the number of patients alive and dead at T_0 and T_5. The patients being provided the data in terms of 5-year survival curves are also being presented with midpoint data that can influence their choice.

The other point about the depiction of framing effects in terms of words and numbers as contrasted with graphical data displays is that even if a researcher devises a graphical display of only two sets of data (the number of patients alive at T_0 and the number of patients alive at T_5), there may well be a point where one survival curve crosses the other, and participants could use this point within their decision, making the interpretation of the graphical data comparison more complex than that of the choice depicted in terms of words and numbers at T_0 and T_5 only.

Yet, despite the above considerations, the results of such framing studies in medical decision making have shown a marked consistency.

Results of Typical Framing Studies

The results of typical framing studies in medical decision making are as follows: Participants who are presented with data framed solely in terms of survival choose the treatment that gives patients a better chance of being alive 5 years after the initial treatment (T_5), that is, the treatment with the better long-term (and worse short-term) result; participants who are presented with data framed solely in terms of mortality choose the treatment that has a better chance of being alive after the initial treatment at T_0, that is, the treatment with the better short-term (and worse long-term) result.

Graphical Displays Comparing Survival and Mortality Curves

Karen Armstrong and colleagues studied the effect of framing in terms of survival versus mortality on understanding and treatment choice, randomizing study participants to receive one of three questionnaires: one presenting survival curves, one presenting mortality curves, or one presenting both survival and mortality curves. Armstrong and colleagues found that study participants who received only survival curves or who received both survival

and mortality curves were significantly more accurate in answering questions about the information than participants who received only mortality curves ($p < .05$). They found that participants who received only mortality curves were significantly less likely to prefer preventive surgery than participants who received survival curves only or both survival and mortality curves ($p < .05$).

As with the Mazur and Hickam studies with graphical displays of 5-year survival curves, Armstrong and colleagues' study participants who received information in terms of graphical displays of 5-year survival data preferred the long-term survival of surgery rather than the short-term survival advantage with radiation therapy. Armstrong and colleagues also showed that adding a graphical display of mortality curves to a survival curve comparison yields similar effects as those that result from providing study participants with survival curves alone. Here, the suggestion remains that a well-discussed graphical display of 5-year survival curves with appropriate information may provide useful information to patients in representing treatment decisions related to surgery versus radiation therapy for a disease entity such as lung cancer. The research question that remains about graphical displays using comparisons of survival curves alone (or survival curves and mortality curves together) is how to best provide that discussion and explanation to patients without unfairly influencing their choices about survival (mortality) in the short term, medium term, and long term.

Questions About Framing Effects

The main questions about framing effects today are not related to whether these effects are demonstrable in paper-and-pencil type research settings where study participants are asked to choose between or among treatments in hypothetical scenarios. Rather, the questions focus on the actual impact that frames have in advertising in general and in direct-to-consumer advertising of medical products (prescription medicines and medical devices).

Future Research on Framing

Further research on framing needs to focus on how to best present data to patients in ways that minimize the influence of framing on the choices they must make. Initial work on the control of framing effects has focused on tabular displays of data. This work needs to be extended to all verbal and graphical data displays and all presentation formats to offer patients the best display of data for their decision making with the minimum intrusion of influences on the choices they are considering at all times.

Dennis J. Mazur

See also Bias; Cognitive Psychology and Processes; Decision Psychology

Further Readings

Ancker, J. S., Senathirajah, Y., Kukafka, R., & Starren, J. B. (2006). Design features of graphs in health risk communication: A systematic review. *Journal of the American Medical Association, 13,* 608–618.

Armstrong, K., Schwartz, J. S., Fitzgerald, G., Putt, M., & Ubel, P. A. (2002). Effect of framing as gain versus loss on understanding and hypothetical treatment choices: Survival and mortality curves. *Medical Decision Making, 22,* 76–83.

Kahneman, D., & Tversky, A. (1984). Choices, values, and frames. *American Psychologist, 39,* 341–350.

Kahneman, D., & Tversky, A. (2007). *Choices, values, and frames*. New York: Cambridge University Press.

Mazur, D. J., & Hickam, D. H. (1990). Interpretation of graphic data by patients in a general medicine clinic. *Journal of General Internal Medicine, 5,* 402–405.

Mazur, D. J., & Hickam, D. H. (1990). Treatment preferences of patients and physicians: Influences of summary data when framing effects are controlled. *Medical Decision Making, 10,* 2–5.

Mazur, D. J., & Hickam, D. H. (1993). Patients' and physicians' interpretations of graphic data displays. *Medical Decision Making, 13,* 59–63.

Mazur, D. J., & Hickam, D. H. (1994). The effect of physicians' explanations on patients' treatment preferences: Five-year survival data. *Medical Decision Making, 14,* 255–258.

Mazur, D. J., & Hickam, D. H. (1996). Five-year survival curves: How much data are enough for patient-physician decision making in general surgery? *European Journal of Surgery*, *162,* 101–104.

McNeil, B. J., Pauker, S. G., Sox, H. C., Jr., & Tversky, A. (1982). On the elicitation of preferences for alternative therapies. *New England Journal of Medicine*, *306,* 1259–1262.

McNeil, B. J., Weichselbaum, R., & Pauker, S. G. (1978). Fallacy of the five-year survival in lung cancer. *New England Journal of Medicine, 299,* 1397–1401.

O'Connor, A. M., Boyd, N. F., Tritchler, D. L., Kriukov, Y., Sutherland, H., & Till, J. E. (1985). Eliciting preferences for alternative cancer drug treatments: The influence of framing, medium, and rater variables. *Medical Decision Making, 5,* 453–463.

Tversky, A., & Kahneman, D. (1981). The framing of decisions and the psychology of choice. *Science*, *211,* 453–458.

GAMBLES

The standard gamble is a method for eliciting a person's preferences for different outcomes where an outcome may be a physical object, a monetary gain, a medical condition, or some other state of affairs. It uses simulated choices among various outcomes of known preference to quantify the value (or utility) of the target outcome.

To determine how much a myopic (nearsighted) patient values his or her vision, for instance, the following standard gamble may be presented: "Imagine that your doctor offers you a treatment that is known to always permanently reverse myopia with 100% effectiveness and safety. It would provide you with perfect vision without the need for corrective lenses. The treatment does not cause pain, have side effects, wear off, or cost money. Would you accept the treatment?" As most patients respond "yes" to this all gain/no risk scenario, an element of gambling is introduced: "Now imagine that the treatment is successful 50% of the time, but causes immediate painless death in the other half of patients. Your doctor cannot predict whether you will have a success or not. Would you accept this treatment?" If the patient says "yes," the probability of success is decreased (and the complementary probability of death is increased) and the gamble re-presented. If the patient says "no," the probability of success is increased. The gamble is repeatedly presented with new probabilities until the patient is indifferent between the two choices.

The probability of success at which the patient is indifferent is a quantitative expression of the patient's value for his or her vision on a scale that is anchored by death at 0.0 and perfect vision at 1.0. The patient who is willing to accept a treatment with 95% efficacy and 5% chance of death, but not at a higher probability of failure, is said to have a utility of .95 for their vision deficit. Another patient willing to accept a 60% chance of death to attain perfect vision has a utility of .40 for their current vision.

Uses

Preferences (also called values or utilities) are one of the critical aspects of a decision that must be specified for decision analysis or cost-effectiveness modeling. (The others are the choices that are under consideration, their potential outcomes, and the probabilities of those outcomes.) The standard gamble is one of several methods for eliciting preferences.

For instance, the decision to undergo laser eye surgery to repair myopia can be modeled as a choice between continued use of corrective lenses and surgery. Each choice is associated with different monetary costs and different probabilities of achieving perfect vision, requiring lenses into the future, and experiencing complications (including pain, blindness, and, rarely, death). These outcome states and their probabilities may be more or less accurately described by an expert such as a physician. However, some patients will place a greater value on some of the outcomes than other patients will. A myopic actor with a tendency to eye infections from contact lenses and the need to eschew eyeglasses on stage may value achieving perfect vision more than a graduate student on a limited budget. The way people feel about the outcomes is important to understanding both how they behave when faced with decisions and how medical providers might advise them.

Utilities are the gold standard measure of quality of life; combined with life-expectancy, they are used to calculate quality-adjusted life years (QALYs).

Assumptions

Standard gambles are derived directly from the assumptions of John von Neumann and Oskar Morgenstern's expected utility theory, which asserts that decision makers must be able to express a preference (or indifference) among any two health

states; preferences are transitive (if $A > B$ and $B > C$, then $A > C$); if $A > B > C$, there is a probability p such that the decision maker is indifferent between B for certain or A with probability p and C with probability $1 - p$; the decision maker must prefer the gamble with the highest probability of attaining the preferred health state; and the decision maker is indifferent among results attained through any combination of simultaneous gambles.

Assume that A is perfect vision without spectacles, B is current vision with spectacles, and C is total binocular blindness. If we can assign a numeric value to perfect vision (A = perhaps 1.0) and to blindness (C = 0.0), the assumptions allow us to estimate the value of spectacles. (Remember that B must lie between A and C.) The gamble is a choice between keeping the status quo (B) for sure or accepting the chance of getting either A or C. If $p = .95$, there is a 95% chance of achieving perfect vision, but a 5% chance of death. Certainly, some, but not all, eyeglass wearers would accept that gamble. They are willing to accept at least a 5% mortality to improve their vision. Some would be willing to accept a higher mortality (perhaps 10% or 20%), and some would refuse the gamble unless p were higher (perhaps .99) and the chance of mortality ($1 - p$) correspondingly lower (.01). The lowest p that the subject is willing to accept is called the utility. A subject who is willing to accept a 90% chance of success, but not 89%, has a utility for spectacles of .90.

Standard gambles assume that the target health state lies on a coherent number line between the two anchor states. If the target is much disfavored (perhaps persistent vegetative state), some subjects may value it less than 0 on a scale anchored by death and perfect health. Standard gamble values less than 0 (or greater than 1) are not easily interpretable.

Standard gambles further assume that the value of a health state is independent of the time spent in that state ("constant proportional trade-off"). A headache is no more or less severe for lasting 1 hour or 1 month. (The effect of duration is modeled independently and combined with the utility to generate the expected utility.)

Limitations

The standard gamble actually measures two values that are conflated in the single result. In addition to the value of the health state itself, the patient must deal with loss of control and other risky aspects inherent to the gamble. Some patients avoid risk so assiduously that they cannot be made to assign a utility less than 1.0 to any health state, no matter how dire. Others seem to have the attitude of inveterate gamblers or adolescents and seek out risky alternatives out of all proportion to the benefits, apparently because they desire the risk itself.

Some investigators feel that the standard gamble is difficult for subjects to complete. It is not clear if this is an accurate representation. If so, the difficulty may stem from dealing with the concept of death or the necessity to carefully consider their values when answering. In fact, many subjects report stopping to think about family obligations, personal aspirations and goals, and other essential issues during the standard gamble exercise. Although emotionally and intellectually challenging, it is unlikely that "easier" methods that avoid such introspection yield results that are as valid.

Some subjects (about 3% to 7% in most studies) fail to rank health states in an apparently rational order. For instance, in assessing various vision states, some subjects place a higher standard gamble value on binocular (total) blindness than blindness in one eye. In fact, some assessors use these two health states to introduce the method to subjects and test comprehension of the task. If the utility of binocular blindness is greater than the utility of monocular blindness, all the other standard gamble results for that subject are considered invalid.

In common with other methods of eliciting utilities, the standard gamble is subject to *framing effects*. People tend to be risk-averse when considering gains but risk tolerant when avoiding equivalent losses. For instance, many people are willing to take a higher risk of death to prevent becoming blind than to cure existing blindness.

Alternative Approaches

Utilities can be assessed in a number of ways. The most straightforward methods, such as asking the patients directly or having them mark a visual analog scale, tend to be subject to a remarkable number of cognitive biases. Perhaps chief among these is that patients don't seem to know their preferences, at least in quantitative form. Of course, they know

what they like and what they don't like, but putting numeric utility values on these preferences is not part of their everyday experience. Therefore, the answers they give to direct assessments are often grossly misleading when used in formal models.

Observing real-world choices is limited by the difficulties and expense of data collection and by the fact that people often get bargains. A patient who is bothered by corrective lenses a great deal will make the same observable decision (to have surgery) as one who has only a moderate distaste for glasses as long as the risks and costs of surgery are low enough. In this case, the first patient is getting more value for the same risk as the second, but an observer cannot discern the difference.

Simulated decisions, such as the standard gamble, allow the decision to be repeated multiple times with slight variations in the conditions of the choices until the pattern of decisions reveals the underlying value trade-offs. Other approaches to utility assessment that take advantage of the simulated trade-off use different anchors to frame the decision. For instance, the time trade-off replaces the varying risk of death in the standard gamble with varying survival times. Rather than "What risk of death would you accept to improve your health?" the question is "What reduction in length of life would you accept to improve your health?" Because the time trade-off specifies the time of death, it avoids the probabilistic or risky aspects of the standard gamble. This has the advantage of eliminating risk tolerance from the assessment but the disadvantage that the outcomes are delayed. Many decision makers value postponed outcomes less than current outcomes of similar value. This discounting effect, like the risk-tolerance effect in the standard gamble, may partially obscure the value of the underlying health state.

Willingness-to-pay methods express utilities in monetary terms. "How much money would you pay to avoid the health state?" Unfortunately, it is difficult to compare the values derived from a multibillionaire with those from an indigent farm laborer.

Benjamin Littenberg

See also Chained Gamble; Disutility; Expected Utility Theory; Quality-Adjusted Life Years (QALYs); Utility Assessment Techniques; Willingness to Pay

Further Readings

Gafni, A. (1994). The standard gamble method: What is being measured and how it is interpreted. *Health Services Research, 29,* 207–224.

McNeil, B. J., Weichselbaum, R., & Pauker, S. G. (1978). Fallacy of the five-year survival in lung cancer. *New England Journal of Medicine, 299,* 1397–1401.

Pauker, S. G., & Kassirer, J. P. (1987). Decision analysis. *New England Journal of Medicine, 316*, 250–258.

Sox, H. C., Blatt, M. A., & Higgins, M. C. (2007). *Medical decision making*. Philadelphia: ACP Publications.

Torrance, G. W., & Feeny, D. (1989). Utilities and quality-adjusted life-years. *International Journal of Technology Assessment in Health Care, 5,* 559–575.

Genetic Testing

Genetic testing includes prenatal and clinical diagnosis or prognosis of fetal abnormalities, predicting risk of disease, and identifying carriers of genetic disorders. Genetic testing is distinguished from genetic screening, which is offered to healthy people who might benefit from more information about their genetic risk profile. In this context, the initiative for a genetic test comes from the healthcare professional and not from the counselee. Learning about the perceptions of genetic risks and the way people make decisions with regard to these risks is becoming increasingly relevant, given the rapidly growing knowledge about the human genome. Recent developments in molecular genetics have led to speculations that we are moving into a new era of predictive medicine in which it will be possible to test for a variety of genes to determine the chances that an individual will at some point in the future develop a disease. At this moment, genetic counseling mostly takes place in clinical genetic centers. In the future, this will probably be done in other settings, and clinicians, primary care physicians, and other healthcare professionals will be more and more often confronted with persons seeking advice about their genetic risks. Furthermore, an increasing number of tests will be available for screening for genetic diseases (self-help kits are already being offered on the Internet).

The aim of individual genetic counseling, but also of population genetic screening, is to provide

people with information about (future) diseases or possible diseases in their offspring and help them process the information in such a way that they can make informed decisions or take some action, for example, therapy or preventive measures. In this view, providing probability information in an understandable way is one of the most essential components of genetic education and counseling. An accurate risk perception is assumed to be an important determinant for taking preventive measures, or it may provide a basis on which counselees can make informed decisions about important personal matters such as childbearing. However, inaccurate perceived risk may also lead to unwanted behavior, such as excessive breast self-examination in case of heritable breast cancer, and an increased vulnerability to worry and distress or an unhealthy lifestyle; or it may lead to false reassurance. The underlying assumption that accurate perception of risk could help counselees make informed and individual decisions seems to be based on the principle of the autonomous and rational decision maker. However, the characteristics of genetic risk information as well as of the decisions to be made complicate a rational trade-off of pros and cons of genetic testing.

Genetic Risk Perception

Genetic risk information is often complex and replete with uncertainties, associated not only with the hereditary nature of the disease but also with the informativeness of the test results, the effectiveness of possible preventive measures, and the variability of expression of the disease. The uncertainties involved in genetic information differ between different genetic testing settings. The following examples illustrate the complexity of genetic risk information and the many ways in which these risks can be expressed.

Genetic counseling for hereditary cancer, such as hereditary breast cancer, includes education regarding the genetics of cancer; the probability of developing cancer and of carrying a genetic mutation; the benefits, risks, and limitations of genetic susceptibility testing; and prevention strategies. A woman who wants to have a DNA test for hereditary breast cancer may receive a nonconclusive test result, meaning that although she has a family history of breast cancer, a genetic disposition is not found. If, on the other hand, she receives a positive test result, she knows for certain that she is a carrier of a mutant breast cancer gene, but she is not certain about the chance (between 45% and 65%) of developing breast cancer during her life. She also does not know when she may get breast cancer. If she decides for a prophylactic breast amputation to prevent the development of breast cancer, she is not even certain that this drastic intervention will reduce her chances of developing breast cancer to zero. Besides, she has to consider the potential impact that the information about hereditariness of breast cancer might have on her family.

Different uncertainties are involved in prenatal testing. Pregnant women have to make a decision whether or not to have a prenatal screening test on congenital disorders, in particular, Down syndrome. There are many probabilities that a woman might consider before making this decision. If she is 36 years old, she has an increased age-related chance of having a child with Down syndrome of 1 out of 250. A woman of her age has a large chance (about 20%) of receiving a positive test result if she opts for screening, that is, an increased risk of larger than 1 out of 250 of having a child with Down syndrome. However, a negative—favorable—test result does not mean that she is not carrying a child with Down syndrome. If she decides not to have prenatal screening, her chance of being pregnant with a child with Down syndrome is based on her age. This chance is quite large compared with younger women. In case the test result is positive, she must also decide whether or not she wants to have an amniocentesis, which has a risk of 1 out of 300 of inducing an abortion of either an affected or a healthy fetus. In addition to this information, she might also consider the severity of the handicap of a child with Down syndrome. The severity of mental retardation cannot be predicted. However, it is known that about 40% of children with Down syndrome experience heart problems. Deafness and other health problems are also possible.

Multifactorial diseases such as type 2 diabetes and cardiovascular disease are caused by a complex interplay of many genetic and nongenetic factors. Although genetic testing for susceptibility genes for many multifactorial diseases, such as type 2 diabetes, is not yet warranted in clinical practice, an increased susceptibility can be determined using family history information. Family history is

an important risk factor that may be used as a surrogate marker for genetic susceptibility and is seen as a useful tool for disease prevention in public health and preventive medicine. Family history reflects the consequences of a genetic predisposition, shared environment, and common behavior. Based on family history and other factors such as lifestyle, an individual's risk of disease can be determined. The information may be used either to identify high-risk groups or as an intervention tool to tailor behavioral messages. For an individual with an increased susceptibility for cardiovascular disease, for example, a healthy diet, physical exercise, and not smoking are even more important than for a person with a population risk of disease.

Research has shown that the perception of genetic risks tends to be inaccurate. Genetic counseling for hereditary cancer, for example, has been shown to improve accurate risk estimation in some women with a family history of breast cancer, although the majority of women still have an inaccurate perception of their lifetime risk after counseling. Meta-analysis of controlled trials showed that genetic counseling improved knowledge of cancer genetics but did not alter the level of perceived risk. Prospective studies, however, reported improvements in the accuracy of perceived risk. Studies about risk perception and prenatal screening showed that pregnant women do not have an accurate perception of their risk of being pregnant with a child with Down syndrome, which is assumed to be important for the decision to have the prenatal testing performed. Research showed that the decision to undergo prenatal screening for Down syndrome was mainly determined by the woman's attitude toward undergoing prenatal screening and not her perceived risk of having a child with Down syndrome. The increased risk due to family history of, for example, type 2 diabetes, is also underestimated. Recent studies indicate that fewer than 40% of people with a positive family history of type 2 diabetes actually perceive themselves to be at risk.

Genetic Decisions

Genetic testing may enable early disease detection and surveillance leading to effective prevention strategies, among other benefits. Genetic decisions are decisions for which informed decision making is seen as particularly important due to the lack of curative treatment for certain conditions. However, in reality this aim is not always achieved. A Dutch study in which prenatal screening for Down syndrome was offered in an experimental setting showed that only 51% of the pregnant women made an informed and reasoned choice despite detailed information leaflets. Studies in other countries have shown even more pessimistic results. Informed decision making in the context of genetic testing for monogenetic diseases or diseases caused by specific genes, such as hereditary breast and colon cancer, has not been studied. For these settings, it is even unclear whether knowing one's risk really increases the freedom of choice or the rationality of a decision. It has been argued that counselees are not really interested in knowing the probability of getting the genetic disease. Traditionally, genetic counseling has been concerned with communicating information about risk largely within the context of reproductive decision making, and risk was seen as a stimulus that elicited a largely predictable response. It was assumed that counselees given risk information would make reproductive plans that reflect the risk level of the birth defect. Investigations of outcomes of genetic counseling, however, have not consistently supported these expectations. Research confirmed that it was not the risk factor that influenced reproductive decisions but the burden of risk to that family and the personal experience with the disorder. The magnitude of the genetic risk was of relative importance only.

Knowledge of genetic risk for multifactorial diseases, that is, susceptibility to a disease, may motivate people to engage in risk-reducing behaviors and might be more motivating than other types of risk information because of its personalized nature. On the other hand, a genetic susceptibility may be perceived as a fixed, unchangeable self-attribute, especially when it is established by DNA testing, and may trigger feelings of fatalism, the belief that little can be done to change the risk, and may adversely affect motivation to engage in risk-reducing behavior. Evidence concerning responses to this kind of genetic risk information is limited and inconclusive. Some studies show that genetic risks are perceived as less controllable and less preventable, while others find no support for this.

For genetic information for risk of disease, whether it is hereditary breast cancer, Down

syndrome, or familial diabetes, dilemmas that generally are not part of patient decision making in many nongenetic contexts play a role. The familial quality of genetic information may raise ethical dilemmas for physicians, particularly related to their duty of confidentiality, especially when multiple family members are seen at the same clinic. It is therefore important to use careful procedures to ensure that results and other sensitive information are not inadvertently communicated to a third party. The principle of confidentiality can expand to the social level as well, given the potential for genetic discrimination. Testing for multifactorial diseases implies different dilemmas. Genetic prediction of disease is based on testing for multiple genetic variants. A person's risk of disease will be based on this genetic profile and may have a range of different probabilities, mostly associated with an increased risk for more than one disease. Because multiple genes are involved, family members most likely will not share the same profile and their susceptibility to diseases will probably differ.

Because of the far-reaching consequences of genetic information, it is particularly important that decisions be autonomous decisions, that clients make an informed, noncoerced testing decision, and that they understand the benefits, risks, and limitations of testing. Informed decision making presupposes adequate knowledge and should be based on the participants' values. Decision aids are a promising type of intervention to promote informed decision making, although research is not conclusive about whether they indeed lead to better-informed decisions. For multifactorial diseases, providing genetic information about an increased susceptibility for disease is not enough to motivate behavioral change to reduce this risk, and tailored education strategies might be needed.

Danielle R. M. Timmermans and Lidewij Henneman

See also Informed Decision Making; Patient Decision Aids; Shared Decision Making; Uncertainty in Medical Decisions

Further Readings

Bekker, H., Thornton, J. G., Airey, C. M., Connolly, J. B., Hewison, J., Robinson, M. B., et al. (1999). Informed decision making: An annotated bibliography and systematic review. *Health Technology Assessment, 3,* 1–156.

Berg, M., Timmermans, D. R. M., Ten Kate, L. P., Van Vugt, J. M., & Van der Wal, G. (2006). Informed decision making in the context of prenatal screening. *Patient Education and Counseling, 63,* 110–117.

Braithwaite, D., Emery, J., Walter, F., Prevost, A. T., & Sutton, S. (2006). Psychological impact of genetic counselling for familial cancer: A systematic review and meta-analysis. *Familial Cancer, 5,* 61–75.

Burke, W. (2002). Genetic testing. *New England Journal of Medicine, 347,* 1867–1875.

Collins, R. S. (1999). Shattuck lecture: Medical and societal consequences of the human genome project. *New England Journal of Medicine, 341,* 28–37.

Emery, J. (2001). Is informed choice in genetic testing a different breed of informed decision-making? A discussion paper. *Health Expectations, 4,* 81–86.

Green, J. M., Hewison, J., Bekker, H. L., Bryant, L. D., & Cuckle, H. S. (2004). Psychosocial aspects of genetic screening of pregnant women and newborns: A systematic review. *Health Technology Assessment*, *8*(33), 1–109.

Heshka, J. T., Palleschi, C., Howley, H., Wilson, B., & Wells, P. S. (2008). A systematic review of perceived risks, psychological and behavioral impacts of genetic testing. *Genetics in Medicine, 10*, 19–32.

Khoury, M. J., McCabe, L. L., & McCabe, E. R. B. (2003). Population screening in the age of genomic medicine. *New England Journal of Medicine, 348*, 50–58.

Yoon, P. W., Scheuner, M. T., Peterson-Oehlke, K. L., Gwinn, M., Faucett, A., & Khoury, M. J. (2002). Can family history be used as a tool for public health and preventive medicine? *Genetics in Medicine, 4,* 304–310.

Government Perspective, General Healthcare

The perspective of the decision maker is a very important element of the decision-making process. Awareness of the decision maker's perspective can help guide the decision prior to it being made or understand the decision after it has been finalized. This is especially important for the decisions made by government agencies because of the important role that those agencies assume with respect to healthcare.

The emphasis here is on government agencies rather than the government as a whole because, contrary to what some believe, governments are usually not monolithic entities with a single societal perspective. Rather, they are collections of individual organizations, each with a distinct role (or set of roles) and, hence, perspective. The U.S. government is a good example, as it has multiple agencies with unique, and sometimes conflicting, roles and perspectives. Understanding "the government's position" requires one to first determine which agency is involved and what role it is assuming with respect to the issue at hand. That role then determines its perspective—the primary viewpoint that goes along with that role.

When it comes to healthcare, the federal government has several different roles that it fills through its various agencies. It is a driver of innovation, a protector of public health, a regulator, a payer, a payer-provider, and a stimulator of system change and quality improvement. In what follows, these different roles and the perspectives that they imply are discussed. Although this entry focuses on the federal government, similar examples could be taken from state and local governments as well as from large, democratically elected governments in other countries.

Driver of Innovation

The federal government drives innovation in a few different ways. First, it awards patents to the developers of new technologies, such as new drugs and new medical devices. This process is handled by the U.S. Patent and Trademark Office (USPTO). The perspective of the USPTO focuses on getting potentially useful inventions to market as quickly as possible so that society can benefit from the new innovation. It is also focused on the protection of intellectual property so as to maintain the incentives for innovation.

The second way the government stimulates innovation is through the conduct and sponsoring of basic research. Consider, for example, the U.S. National Institutes of Health (NIH), a research organization that is primarily focused on conducting basic research aimed at understanding diseases and identifying potential biologic interventions that could eliminate or reduce the burden of those diseases. As such, the perspective of the NIH revolves around the generation of new information about the biologic causes of disease. Because this frequently takes a lot of time, its perspective is also more long term in nature.

Protector of Public Health

The federal government also acts as a protector of public health. One of the lead agencies in this role is the U.S. Centers for Disease Control and Prevention (CDC), which investigates disease outbreaks, sponsors programs to improve the health of populations, and conducts applied research that is focused on specific interventions that are known to reduce the burden of a disease—especially those that are concerned with behavior.

Another example of a government agency that acts as a protector of public health is the U.S. Food and Drug Administration (FDA). Specifically, the FDA regulates healthcare technologies, especially when they are first being introduced to the healthcare marketplace. Its goal is to ensure that those technologies are safe and effective, and its perspective toward healthcare is centered on that goal.

Because of their roles as protectors of public health, the perspectives of these two agencies are centered on the health of populations rather than the health of single individuals. They also tend to value safety above all other concerns.

Regulator

Regulation is an important part of government business, and healthcare is one of the areas subject to government regulation. The role of the FDA as a protector of public health has already been discussed. Unlike the CDC, though, the FDA also has a major role in regulation, with the authority to deny market access to drugs and medical devices that it deems unsafe.

The USPTO in its role as a driver of innovation has also been discussed. Like the FDA, it also has regulatory power insofar as it can grant or deny patents to the inventors of new technologies. This implies that the USPTO acts to regulate the competitiveness of healthcare markets that are created by new innovations.

Another example of an agency that assumes the role of regulator in the healthcare marketplace is the U.S. Federal Trade Commission (FTC). The

FTC works to ensure that markets remain competitive (except for those protected by a patent). It tries to prevent one firm or a set of firms from gaining market power—the power to set prices at a higher level than one would otherwise expect in a competitive marketplace. This applies to the market for healthcare just as it does to other industries. An example would be the prevention of anticompetitive mergers, such as those involving hospital chains or healthcare plans. The FTC perspective is focused on ensuring a competitive marketplace, even if that means denying mergers that could generate efficiencies that lower healthcare costs.

Payer and Payer-Provider

The U.S. Centers for Medicare and Medicaid Services (CMS) is an example of a government agency that takes on the role and perspective of a payer, as it has assumed financial responsibility for the healthcare given to older citizens (Medicare) and those living in poverty (Medicaid). Thus, unlike some of the other agencies, CMS is very concerned with the cost and quality of healthcare, particularly that given to its beneficiaries. This is especially true for new technologies that have received a patent from the USPTO and been approved for use by the FDA. As these new innovations are adopted as part of routine care, CMS (and its financial backer, the U.S. Congress) must be concerned with the impact the new technologies have on its overall budget.

Another agency that shares the payer perspective is the U.S. Office of Personnel Management (OPM), the organization that is responsible for overseeing the Federal Employee Health Benefit Program (FEHBP). As with CMS, it also has a keen interest in the cost and quality of healthcare given to its "members"—federal employees and their families.

The U.S. Department of Defense (DOD) and the U.S. Department of Veterans Affairs (VA) also have a payer perspective, with a strong focus on the cost and quality of healthcare. However, these two agencies also play the role of provider insofar as they have clinics, hospitals, and personnel who actually provide some of the care to their members. This allows them to directly control some of the cost and quality of healthcare—an element of control that CMS and OPM do not have.

Stimulator of System Change and Quality Improvement

Several of the agencies discussed above have some role in stimulating system change and quality improvement. For example, CMS has pilot projects that are meant to determine whether the current healthcare system and the level of quality created by it could be improved. However, these kinds of efforts are not the focus of those agencies. One agency that does have this as its primary focus is the U.S. Agency for Healthcare Research and Quality (AHRQ). AHRQ is committed to help improve the U.S. healthcare system, primarily through health services research that evaluates the quality, effectiveness, and efficiency of specific medical interventions targeted to individual patients. As such, its perspective reflects an interest in the care received by individual patients, with less emphasis on the health of populations.

Edward C. Mansley

See also Government Perspective, Informed Policy Choice; Government Perspective, Public Health Issues; Medicaid; Medicare

Further Readings

Agency for Healthcare Research and Quality. (n.d.). *About AHRQ*. Retrieved August 29, 2008, from http://www.ahrq.gov/about

Centers for Disease Control and Prevention. (n.d.). *About CDC*. Retrieved August 29, 2008, from http://www.cdc.gov/about

Centers for Medicare & Medicaid Services. (n.d.). *About CMS*. Retrieved August 29, 2008, from http://www.cms.hhs.gov/home/aboutcms.asp

Federal Trade Commission. (n.d.). *About the Federal Trade Commission*. Retrieved August 29, 2008, from http://www.ftc.gov/ftc/about.shtm

National Institutes of Health. (n.d.). *About NIH*. Retrieved August 29, 2008, from http://www.nih.gov/about/index.html

U.S. Department of Veterans Affairs. (n.d.). *About VA Home*. Retrieved August 29, 2008, from http://www.va.gov/about_va

U.S. Food and Drug Administration. (n.d.). *About the Food and Drug Administration*. Retrieved August 29, 2008, from http://www.fda.gov/opacom/hpview.html

U.S. Office of Personnel Management. (n.d.). *Federal Employees Health Benefits Program*. Retrieved

September 2, 2008, from http://www.opm.gov/insure/HEALTH/index.asp
U.S. Patent and Trademark Office. (n.d.). *Introduction.* Retrieved August 29, 2008, from http://www.uspto.gov/main/aboutuspto.htm

Government Perspective, Informed Policy Choice

Governmental perspectives on the individual's right to choice in healthcare and, in particular, the concept of informed choice, are relatively recent. Choice is now seen as being integral to healthcare reforms (patient-led care) taking place in the United Kingdom, the United States, and elsewhere. These policies contrast with some previous policies in which the government made choices on behalf of the population. Current government policy toward informed choice, in countries such as the United States and the United Kingdom, is based on the premise that people have an individual responsibility for their own health and are able to make their own choices. The provision of information (particularly evidence-based information) is viewed as being the key to enabling people to make rational choices.

Responsibility for Health

Individual choice in modern health policy in the West has its origins in the intellectual and economic revolutions of the 18th century. Before industrialization, attitudes on health and disease were largely defined by religion. The effect of the Enlightenment and the scientific revolution was to replace this with reason. At the same time, the industrial revolution brought with it a new wave of disease and death related to rapid urbanization. Liberal, laissez-faire economic theory initially influenced thinking on health, which as a result was seen as the responsibility of the individual rather than the government. However, investigations and a report by British politician Edwin Chadwick in 1842 linked ill health to water, overcrowding, and other environmental problems. Thereafter the U.K. government took greater responsibility for the health of the population, particularly the sick poor. Public health measures included centralization of the provision of drainage, water, and sanitary regulations. Such centralization was also seen in parts of Europe but took longer to happen in the United States.

Responsibility by the government for the health of the population was to achieve its greatest expression in the United Kingdom with the creation of the National Health Service (NHS) in 1946. By the 1970s, however, the triumph of neoliberal ideology within the leadership of the Conservative Party led to the primacy of ideas of competition, deregulation, and individual choice, ideas that were to continue to influence the new Labour governments from 1997 onward and to become the dominant ideology across the political spectrum. Their application to health meant that users of health services were seen as individualized consumers who (it was assumed) would place a high value on choice and particularly the concept of informed choice.

Informed Choice and Informed Consent

The concept of informed choice is largely based on the principles of informed consent, but there are significant differences between the two concepts. Informed choice usually (although not always) implies the stage before a decision has been made and concerns providing information for people to make rational choices, for example, about the place of care, whether to have health screening, or what type of treatment to have. Informed consent implies that a decision has already been made and concerns the disclosure of the risks involved in, for example, undergoing surgery or an invasive procedure. Both informed choice and informed consent in medical care (and in research), however, have one overarching principle: promoting patient autonomy by providing information on risks and benefits of a healthcare choice, intervention, or treatment.

The doctrines of informed consent and informed choice have ethical, legal, and clinical interpretations. It has been argued that medicine has long been committed to ethics and morality when dealing with the patient, although this commitment may be incomplete at times. Ethical concerns over informed consent have been around since the early 20th century. However, it was not until the mid-1950s that an autonomy model rather than a

beneficence model (which depicts the physician's primary obligation as providing medical benefit) governed the justifications for informed consent.

Legal interest in informed consent and the rights of patients has been evolving alongside the ethical interest. Initially, this concerned consent to treatment, however uninformed. However, it became recognized that patients are autonomous human beings with rights and interests independent of the medical profession. The judicial doctrine of informed consent in healthcare is based primarily on decisions about treatments.

The doctrine of informed consent in clinical care emerged to some extent because of the perception that patients were uninformed and thus powerless in healthcare (i.e., without autonomy). Informed consent was developed in clinical care from an obligation by doctors to disclose information on the risks of procedures and to act with beneficence and nonmalfeasance. One way to redress the imbalance of power between patient and clinician is to inform the patient. The concomitant of the doctrine of informed consent is therefore the right to refuse treatment. Thus, there was necessarily a shift (in theory), away from paternalism and beneficence in medicine (however benign), toward a partnership between patient and physician, with the transference of information playing a crucial role. However, what the information should comprise, and how best to inform people (or make them informed), is the subject of a large body of research.

Information Required to Make an Informed Choice

As an ethical principle, provision of unbiased information is seen as being the key to respecting patient autonomy. However, it has been recognized that the provision of information alone will not necessarily ensure that people become autonomous or fully informed; the information also needs to be evidence-based, understandable, unbiased, and relevant. Research indicates that people still lack knowledge in certain areas, despite information given to them. Although policy makers cannot assume that the provision of information necessarily results in an informed (i.e., knowledgeable) population, they still have a responsibility to provide such information and ensure that it is the information that the population wants and needs to make an autonomous choice.

There is also debate and uncertainty as to what constitutes "sufficient" information to make someone informed. Recent research suggests that people want information through which to contextualize their choices, not just information on risks and benefits. In addition, many of the organizations that develop information resources are the same organizations that are involved in delivering the services. This may mean that the information is highly regulated or that it lacks independent scrutiny.

Informed choice is gaining prominence in two particular areas of government health policy making: giving patients choice about the location or types of care (e.g., Hospital A or Hospital B) and encouraging healthy choices (e.g., going for health screening, exercising more).

Choice of Location of Care

In the United Kingdom, for example, the government's policies are increasingly aimed at offering, and expecting, people to make choices about the location of their care (e.g., choice of hospital). Choice is promoted as being a positive element of the health services experience, something that all patients want. In 2004 and 2005, a number of U.K. government policies were set out that required NHS organizations to offer a choice of four to five hospitals to patients requiring elective care ("choose and book"). The aim of this policy was to achieve one of the standards set out in the NHS Improvement Plan. This policy of choice of hospital was thought to benefit both patients and the NHS. For the patients, it was viewed as a way of providing them with more personalized, flexible, convenient care. For the NHS, it was viewed as reducing administrative bureaucracy and patient cancellations and providing a more standardized and improved service.

The policy makers recognized that people would need support to make informed choices about which hospital to choose and identified four key areas of information that they would need. These were waiting times, location and convenience of the hospital, patient experience, and clinical quality.

Although such government policies involve encouraging people to make informed choices about various aspects of their healthcare, it is not

clear as to what extent people are able, or indeed want, to make rational, informed choices about their place of care at times when they may be already overwhelmed with other information about their disease or condition. There is some evidence to suggest that not all people see choice as a positive element of healthcare, as it can create stress and anxiety, especially if the choice that they make is later deemed to be the "wrong" choice. Also, the choices that they are being offered (e.g., choice of hospital) do not necessarily reflect the choices that they want (e.g., care at home). There is also scant evidence for the impact of choice policies on other issues such as equity—for example, some people may be more likely than others to get their hospital of choice or the maternity service that they request.

Informed Choice for Health Screening

Until recently, the focus of health screening programs and policies has been to maximize cost-effectiveness by achieving the highest coverage and uptake possible. The benefits of screening for cancer were deemed to be so great that any potential harm or limitations were given little attention. However, there is increasing recognition by U.K. government policy makers that, even when it is accepted that screening has a net beneficial effect (to the population), one of the inherent limitations is that some individuals will be harmed. In the United Kingdom, screening policy makers now, in principle, at least, consider informed choice alongside more conventional screening parameters such as quality assurance procedures and improvements in survival. The reasons given by the U.K. National Screening Committee in 2000 for promoting informed choice were the recognition of change in social attitudes and the acknowledged risks and consequences. It was observed that the advantage of increasing informed choice is that it prevents people feeling coerced. It was also seen as having economic advantages in that an informed choice policy may create opportunities for selective screening based on individual risk profiles. However, there may be conflicting or even contradictory motives at work among those involved in screening. For example, although an informed choice policy may be presented as one that promotes individual choice and autonomy, other factors (such as target payments for uptake of cervical screening) may discourage health professionals from actively implementing the policy. Therefore, policy makers in cancer screening may need to decide whether they really want to increase and promote informed choice, or whether they want to increase informed participation—the choice to decline screening is neither promoted nor endorsed.

Informing Healthier Choices

Government policies toward public health may involve a combination of encouraging the population to take responsibility for their own health by making informed, healthier choices (e.g., using health promotion initiatives to tell people to stop smoking) and the government taking responsibility for health (e.g., through legislation and regulation). For example, policies in the United Kingdom include "Informing Healthier Choices: Information and Intelligence for Healthy Populations." These policies encourage informed choice, but the overarching aim is population health improvement by people choosing to live a healthier lifestyle.

One of the difficulties for government policies promoting individual informed choice is that there is only one "choice" that the government wants: people choosing healthy lifestyles. As in health screening, it is unlikely that the government actually wants people to exercise autonomous choices that result in nonparticipation in screening or to choose unhealthy lifestyles. However, as mentioned previously, the concept of informed choice is based on the premise that both choices are equal and people can make an autonomous choice. If the principles of informed choice are considered valid by policy makers, then the implication is that people should be able to choose not to take up the healthy living advice if they are informed and autonomous.

Current increases in mortality and morbidity attributable to unhealthy behaviors (e.g., rise in heart disease, diet-related cancers) suggest that people are not making healthy choices even when informed by mass media campaigns. One way that government policies try to respond is to take back responsibility for health and try and enforce healthy behaviors by, for example, banning smoking in public places. However, this approach results in people losing their autonomy in exercising choice.

Tensions Between Informed Choice and Public Health Policies

Public health policies such as health screening, immunization, and health promotion initiatives are often only effective if a high percentage of the population complies with the government policies and directives. Tensions and dilemmas arise when the government is trying to implement policies at a population level (e.g., health screening) in tandem with policies aimed at individuals within that population (e.g., promoting informed choice and autonomous decision making). What is not clear is whether a policy of informed choice can operate within a structure where both information and choice are, to some extent, regulated by the need to benefit populations (the public good) rather than individuals. The central issue is whether government-sponsored health initiatives are compatible with the concept of respect for individual autonomy. Many public health policies, such as health screening, are grounded in positions based on outcomes (the theory of utilitarianism). Thus, there may be a tension between policies aimed at the benefit of the population and other policies promoting individual autonomous decision making.

Government policies designed to increase (informed) choice are difficult to reconcile with policies aimed at improving the health of populations. Those aimed at promoting choice will not ensure that desired population health outcomes are achieved. Those designed to benefit populations will mean that a degree of individual choice is lost.

Ruth Jepson

See also Government Perspective, Public Health Issues; Informed Consent; Informed Decision Making

Further Readings

Faden, R. R., & Beauchamp, T. L. (1986). *A history and theory of informed consent*. New York: Oxford University Press.

Jepson, R. G., Hewison, J., Thompson, A., & Weller, D. (2005). How should we measure informed choice? The case of cancer screening. *Journal of Medical Ethics, 55,* 20–25.

Mazur, D. J. (2003). Influence of the law on risk and informed consent. *British Medical Journal, 27,* 731–734.

National Screening Committee. (2000). *Second report of the National Screening Committee*. Retrieved January 8, 2009, from http://www.nsc.nhs.uk/library/lib_ind.htm

O'Neill, O. (2003). Some limits of informed consent. *Journal of Medical Ethics, 29,* 4–7.

Petersen, A., & Lupton, D. (1996). *The new public health: Health and self in the age of risk*. London: Sage.

Porter, R. (1997). *The greatest benefit to mankind: A medical history of humanity from antiquity to present*. London: HarperCollins.

Government Perspective, Public Health Issues

The perspective of any decision-making entity determines which costs and outcomes will be included in a decision and is determined by who the audience to the decision is and how that audience will use the information contained in the decision or disseminated as a result of the decision. In questions of government or public health policy, it is usually most appropriate to adopt a societal perspective, because a more narrowly defined perspective will lead to an inefficient allocation of scarce public resources. A government public health perspective, therefore, would include an audience that tends to be society in general, including those who do and those who do not benefit from a program, policy, or intervention and including those who do and those who do not pay for the program, policy, or intervention. Therefore, from the government public health perspective, all benefits and costs, regardless of those to whom they accrue, would be included in the decision-making process. For example, when assessing costs of a mandatory human papillomavirus (HPV) screening program, a public health perspective would include not just the costs to deliver the screening but the losses in productivity associated with the persons receiving the immunization or experiencing illness and the long-term costs to society for transmission of disease in the absence of disease prevention. Society will use information stemming from government public health decision making to decide how to improve community-level health with societal resources.

Decision making from the government perspective is often based on interventions that are

population based rather than individual based. For example, suppose that the outcome of interest is a reduction in lung cancer. Individual interventions to reduce the incidence of lung cancer could include a decision between therapeutic nicotine patches, gums, or lozenges. Public health interventions from the government perspective, however, could consider such population-based policies as requiring designated smoking spaces within public buildings or banning smoking in public buildings altogether. Thus, interventions, programs, or policies considered from a government public health perspective, by definition, are intended to provide a positive return to the population, in general, and not to just one individual at a time. In the smoking example, the public health intervention is designed to limit the smoking opportunities of the individual and reduce the secondhand smoke affecting the nonsmokers in society.

Decision making from a government public health perspective may also include community-based rather than clinical interventions. Community-based interventions are those interventions intended to promote the community's health and prevent disease and include decisions about interventions typically delivered in states or provinces, local agencies, healthcare organizations, worksites, or schools. Examples of community-based interventions designed to increase physical activity and reduce obesity include the promotion of school-based physical activity programs, urban design and land use policies, and social support services in community settings. Examples of community-based interventions designed to prevent violence include early home visitation to prevent violence against the child, school-based programs to prevent violent behavior, and group-based cognitive behavior therapy to reduce the harmful effects of traumatic events.

Guidance on how to deliver evidence-based community interventions often complements the evidence-based guidance available for clinical settings. For example, evidence-based community interventions that have been recommended to reduce tobacco use initiation include increasing the unit price for tobacco units, restricting access for minors, and conducting mass media campaigns combined with clinical interventions and community mobilization campaigns combined with clinical interventions.

Regardless of intervention setting, the value of government perspective decision making in public health is a focus on prevention versus treatment in promoting health. In an entirely systematic way, the public health perspective examines the effectiveness, economic efficiency, and feasibility of interventions to combat risky behaviors such as tobacco use, physical inactivity, and violence; to reduce the impact of specific conditions such as cancer, diabetes, vaccine-preventable diseases, and motor vehicle injuries; and to address social determinants of health such as education, housing, and access to care.

A focus on prevention, however, makes public health decision making often incommensurate with clinical decision making. Prevention efforts typically result in costs that occur in the short term, while benefits occur in the longer term. For example, the benefits of a nutrition and exercise program may not be realized until many years later, with future reductions in cardiovascular disease and diabetes, whereas clinical decisions often pertain to interventions whereby both the costs and benefits are realized in the short term: for example, statin treatment to lower cholesterol level and reduce the incidence of coronary heart disease and vascular events such as heart attack and stroke. When considering the present value of future costs and outcomes, treatment efforts will always appear more favorable than prevention efforts, if benefits are realized in different time frames.

Another important consideration in public health decision making is the challenge of establishing a causal link between intervention and outcomes because of a lack of longitudinal data to show future outcomes or sustained outcomes. For example, without longitudinal data, it may be difficult to establish the long-term benefits to a community that may result from an environmental improvement plan to provide more green space and exercise facilities for its residents. Thus, decision making from a public health perspective may need to be considered separately from decision making in a clinical setting, where the value of the decision-making information serves different purposes.

Phaedra Corso

See also Government Perspective, General Healthcare; Moral Choice and Public Policy; Trust in Healthcare

Further Readings

Owens, D. K. (2002). Analytic tools for public health decision making. *Medical Decision Ma king*, *22*(5), S92–S101.

Schneider, M. (2006). *Introduction to public health* (2nd ed.). Sudbury, MA: Jones & Bartlett.

ZaZa, S., Briss, P., & Harris, K. (Eds.). (2005). *The guide to community preventive services: What works to promote health?* New York: Oxford University Press.

H

Hazard Ratio

The hazard ratio in survival analysis is the effect of an explanatory variable on the hazard or risk of an event. In this context, the hazard is the instantaneous probability of the event (such as death) within the next small interval of time, assuming that one has survived to the start of that interval. The hazard ratio then compares the hazard of the event under one condition (e.g., treatment for a disease) with the hazard of the same event under a second (baseline) condition (e.g., placebo) by taking the ratio of one hazard over the other. A hazard ratio greater than 1 indicates an increase in the hazard of the event under the first condition over the hazard of the event under the second condition.

Survival Analysis

Survival analysis is a class of statistical methods that deals with the timing of the occurrence of particular events. These methods focus on modeling the *time* to an event such as onset of a particular disease. Survival analysis methods were originally designed to study death, hence the name. However, an event can be defined as the first diagnosis of cancer, the failure of a manufacturing machine, the progression of disease from one stage to another, and attrition times among criminals. An event can also signify a positive occurrence such as marriage, pregnancy, or cure from a disease. Survival analysis is also termed *reliability analysis* or *failure time analysis* in engineering, *duration analysis* or *transition analysis* in economics, and *event history analysis* in sociology. In general, survival analysis involves the modeling of time-to-event data.

Survival Data

Since survival analysis deals with data collected over time until an event occurs, the time origin and event or end point of interest need to be clearly defined. In clinical research, the time origin is typically the time at which a patient is recruited into a study. The event or end point would then be the occurrence of a particular condition such as an adverse event or even death. In another such study, the event of interest could be being cured of the disease. In general, an event is a clearly definable transition from one discrete state to another. Examples of these transitions include the following: from being in pain to pain relief, from being disease-free to having a disease, or from being free to being incarcerated. However, in survival analysis, it is not sufficient to know only who is disease-free and who is not; one also needs to know *when* the transition occurred. Exact times of the event are sometimes known, but often, the timing of an event may only be known within a range. For example, in a well-monitored clinical trial, the onset of an adverse event may be pinpointed to a particular day (e.g., 10 days after study entry). On the other hand, in a study of menarche, only the year (e.g., age 13) of first menstruation may be collected.

There are several reasons why survival data cannot be suitably analyzed by standard statistical methods. First, survival data are typically not symmetrically distributed. For example, in a study of actual death times, the distribution of time to death will often be positively skewed (a histogram of the data will have a long tail to the right of where the majority of observations lie). Hence, those data are not readily amenable to standard statistical procedures that require data to have a normal distribution. Second, many variables that may influence the event or outcome of interest may change over time. These are called time-varying covariates. For example, a patient's increase in blood pressure over time may affect his or her risk of cardiovascular disease. These changes in the blood pressure variable can be easily accommodated in survival analysis models. Finally, and most important, survival analysis methods can deal with censored observations that are described next.

Censoring

One primary feature of survival data that is difficult to deal with using conventional statistical methods is censoring. The survival time of an individual is said to be censored when the end point of interest has not been observed for that individual. If the end point of interest is death, then an individual's survival time may be censored because that individual has been "lost to follow-up." For example, a patient participating in a clinical trial may unexpectedly move to another city before the end of the study and may no longer be contacted. The only survival information that would be available on that patient is the last date that that patient was known to be alive, which may be the date that the patient was last seen at the clinic. On the other hand, an individual's survival time may be censored because the patient is still alive at the end of the study period and his death date is not observed. An observed survival time (i.e., time to death) may also be regarded as censored if the death is known to be unrelated to the treatment under study. For example, a person's death due to a car accident is most likely unrelated to the chemotherapy that the patient was receiving in a clinical trial. However, in instances where it is not clear whether the death is unrelated to the treatment under investigation, it is more appropriate to consider survival time until death due to all causes; or it may be of interest to analyze the time to death from causes other than the primary condition for which the patient was being treated.

There are three primary types of censoring: right censoring, left censoring, and interval censoring. If we let T be a variable that represents the time of occurrence of a particular event, then T is said to be right censored if the only information we have on T is that it is greater than some value, c. For example, if T represents age at death, but a patient was lost to follow-up at age 65, then for that patient we only know that $T > 65$, in which case the patient's event time is right censored at age 65. The right-censored survival time is less than the actual, but unknown, survival time (the censoring occurs to the right of the last known survival time). Left censoring occurs when the actual survival time is less than that observed. One example in which left censoring may occur is in a follow-up study of cancer recurrence in which patients are seen by their oncologist 6 months after their initial treatment for their primary cancer. At the 6-month visit, a patient is examined for disease recurrence. Some patients will have evidence of recurrence at that visit, but the recurrence may have occurred at any time prior to that clinic visit. Hence, the recurrence time is said to be left censored (the censoring occurs to the left of the known examination time). Interval censoring is a combination of both right and left censoring. Revisiting the cancer recurrence example, if patients are followed by their oncologist every 6 months, and cancer recurrence is detected at their third follow-up visit but not at prior visits, then we know that the actual recurrence time is between their second and third clinic visit. The observed recurrence time is said to be interval censored. Right censoring is the most common type of censoring and is handled more readily than other types of censoring when using standard analytic software packages.

Because censored observations are a common occurrence in time-to-event data, all survival analysis approaches provide ways to deal with these types of observations. The most commonly used survival analysis model that allows for censored observations is the proportional hazards model described next.

Estimating the Hazard Ratio

Often, the objective of a survival analysis study is to compare two groups (e.g., those who are given a treatment for a disease vs. those who are administered a placebo) on their risk or hazard of death. The hazard is defined as the instantaneous probability of death within the next small interval of time, assuming that one has survived to the start of that interval. When comparing the hazards of two groups, an assumption is commonly made that the ratio of the hazards (the hazard of death in those treated divided by those given placebo) is the same at all possible survival times. This is called the *proportional hazards assumption*. If the hazards in the two groups at time t are denoted as $h_0(t)$ and $h_1(t)$, then proportional hazards implies that $h_1(t)/h_0(t) = \varphi$ at all survival times, t, and where φ is a constant that does not change over time. This constant is called the *hazard ratio*. Since hazards are always positive, the hazard ratio can conveniently be expressed as $\varphi = e^{\beta}$, where β is a parameter that can be positive or negative. For two individuals who differ only in their group membership (e.g., treatment vs. placebo), their predicted log-hazard will differ additively by the relevant parameter estimate, which is to say that their predicted hazard rate will differ by e^{β}, that is, multiplicatively by the antilog of the estimate. Thus, the estimate can be considered a hazard ratio, that is, the ratio between the predicted hazard for a member of one group and that for a member of the other group, holding everything else constant.

Proportional Hazards Regression Models

The parameter β can be estimated by regression models that treat the log of the hazard rate as a function of a baseline hazard $h_0(t)$ and a linear combination of explanatory variables. Such regression models are classified as proportional hazards regression models and include the Cox semiparametric proportional hazards model and the exponential, Weibull and Gompertz parametric models. These models differ primarily in their treatment of $h_0(t)$. The proportional hazards model first introduced by Cox in 1972 is the most widely used regression model in survival analysis. The main advantage to this model is that it does not require a particular form for the survival times; specifically, the baseline hazard does not need to be specified.

Interpretation

Statistical software packages used to fit a proportional hazards model will generally provide point estimates of the hazard ratio and of the parameter β. A hazard ratio with a value of 1 (corresponding to a value of 0 for β) can be interpreted to mean that there is no apparent difference in hazard of death under the treatment versus the placebo. A hazard ratio less than 1 indicates that the treatment group has a reduced hazard of death over the placebo group, and a hazard ratio greater than 1 indicates an increased hazard of death for those in the active treatment group. In addition to a point estimate, statistical packages will also provide standard errors that allow one to better access the accuracy of the hazard ratio estimate. These standard errors can be used to obtain approximate confidence intervals for the unknown β parameter. In particular, a $100(1 - \alpha)\%$ confidence interval for β is the interval with limits $\hat{\beta} \pm z_{\alpha/2} SE(\hat{\beta})$ where $\hat{\beta}$ is the estimate of β and $z_{\alpha/2}$ is the upper $\alpha/2$ point of the standard normal distribution. If the confidence interval (usually a 95% confidence interval) for β does not include 0, then this is evidence that the value of β is nonzero. The corresponding confidence interval for the hazard ratio can be found simply by exponentiating the confidence limits of β. If the 95% confidence interval for the true hazard ratio does not include 1, then one can be fairly confident that the value of the hazard ratio is not 1. One can also test the hypothesis that there is no difference in hazards between two groups by testing the null hypothesis that $\beta = 0$. This can be tested using the statistic $\hat{\beta}/SE(\hat{\beta})$, whose value can be compared with the percentage points of the standard normal distribution to obtain the corresponding p value. This corresponds directly to testing whether the hazard ratio is equal to 1.

Nandita Mitra

See also Cox Proportional Hazards Regression; Log-Rank Test; Parametric Survival Analysis; Survival Analysis

Further Readings

Collett, D. (2003). *Modelling survival data in medical research* (2nd ed.). London: Chapman & Hall.

Cox, D. R., & Oakes, D. (1984). *Analysis of survival data*. London: Chapman & Hall.
Hosmer, D. W., Lemeshow, S., & May, S. (2008). *Applied survival analysis: Regression modeling of time to event data* (2nd ed.). New York: Wiley.
Kalbfleisch, J. D., & Prentice, R. L. (2002). *The statistical analysis of failure time data* (2nd ed.). New York: Wiley.
Klein, J. P., & Moeschberger, M. L. (2005). *Survival analysis: Techniques for censored and truncated data*. New York: Springer.
Kleinbaum, D., & Klein, M. (2005). *Survival analysis: A self-learning text*. New York: Springer.
Therneau, T. M., & Grambsch, P. M. (2001). *Modeling survival data: Extending the Cox model*. New York: Springer.

Health Insurance Portability and Accountability Act Privacy Rule

Protections in health and medical care are not limited to the protection of an individual's right to make decisions about his or her own body and mind. Protections also extend to the release of an individual's privacy-protected information in medical care and medical research settings. An example of such an extension of protection is the Health Insurance Portability and Accountability Act of 1996 (HIPAA) in the United States.

HIPAA (Public Law 104-191), as enacted in the United States on August 21, 1996, required the Secretary of the Department of Health and Human Services (HHS) to issue privacy regulations governing individually identifiable health information if Congress did not enact privacy legislation within 3 years of the passage of HIPAA.

Since Congress did not enact privacy legislation within that time frame, HHS developed a proposed rule and released it for public comment on November 3, 1999. After review of 52,000 public comments, the final regulation—the Privacy Rule—was published on December 28, 2000. The Standards for Privacy of Individually Identifiable Health Information (Privacy Rule) established for the first time in the United States a set of national standards for the "protection" of all "individually identifiable health information" held or transmitted by a covered entity or its business associate in any form or medium (electronic, paper, or oral).

Distinctions in Data Identifications

When considering issues of identification, one must distinguish among the following concepts. First, one must distinguish between "anonymous data" and "nonanonymous data." Second, in the area of nonanonymous data and individual identification, one must distinguish between two types of individual identification: nonunique versus unique.

Anonymous Versus Nonanonymous Data

Anonymous data are data from which all unique identifiers have been removed. Ideally, it should be impossible to identify a unique individual from a data set composed of anonymous data. However, what may appear on the surface to be an example of anonymous data may become problematic after further examination, as is illustrated later in this entry.

Nonanonymous data are data where individual identifiers have not been removed, and therefore, the data can be traced back to individuals. Sometimes this tracing back will yield one individual; sometimes several; sometimes many.

All individual identifiers are not necessarily unique identifiers. Individual identifiers in HIPAA include the following:

- Names
- All geographic subdivisions smaller than a state, including street address, city, county, precinct, zip code, and their equivalent geocodes, except for the initial three digits of a zip code if, according to the publicly available data from the Bureau of the Census,
  - The geographic unit formed by combining all zip codes with the same three initial digits contains more than 20,000 people and
  - The initial three digits of a zip code for all such geographic units containing 20,000 or fewer people are changed to 000.
- All elements of dates (except year) for dates directly related to an individual, including birth

date, admission date, discharge data, and date of death; and all ages over 89 and all elements of dates (including year) indicative of such age, except that such ages and elements may be aggregated into a single category of age 90 or older
- Telephone numbers
- Fax numbers
- Electronic mail addresses
- Social Security numbers
- Medical record numbers
- Health plan beneficiary numbers
- Account numbers
- Certificate/license numbers
- Vehicle identifiers and serial numbers, including license plate numbers
- Device identifiers and serial numbers
- Web Universal Resource Locators (URLs)
- Internet Protocol (IP) address numbers
- Biometric identifiers, including finger and voice prints
- Full-face photographic images and any comparable images
- Any other unique identifying number, characteristic, or code

Unique Versus Nonunique Individual Identifiers

An individual identifier can be used to link an individual with a piece of health or medical information contained, for example, in a larger data set of health information, but this individual identifier may not be able to identify the individual uniquely. Thus, there are unique identifiers and nonunique identifiers and a range of identifiers in between.

For example, a visual illustration of an individual identifier is a full-face photograph of a patient. While the photo may be an individual identifier, it may not be a unique identifier. In the case of identical twins, the full-face photograph may be able to reduce the number of possibilities in the world to two candidates for the identity of the individual in the photo—that is, the twin or his or her identical sibling. However, someone examining the full-face photograph may not be able to distinguish between the two identical twins. A unique identifier would be able to pick out one or the other of the two twins on the basis of a property held by one twin but not by his or her identical sibling.

Why the Need to Identify Individuals Uniquely?

Why is there a need to identify individuals uniquely as associated with a data set based on a blood or tissue sample and/or a study test result? A data set can contain a result or a set of study results from a test or set of tests used to screen for disease in an asymptomatic patient or to diagnose disease in a symptomatic patient (in medical care) or results obtained after an individual has agreed to participate in a research study and is studied before and then again after a research intervention. In both settings, test results (especially abnormal test results) need to be traced back to the donor of the blood or tissue sample or the individual who agreed to participate in the study. If a blood sample yields a markedly abnormal result, the blood sample will (1) need to be repeated to check the accuracy of the first specimen and then (2) acted on as quickly as possible.

The following is an example. A high serum lead level in an infant being cared for by a provider or participating in a research study needs to be acted on immediately to prevent further damage to the infant from the lead. This requires being able to identify the individual uniquely so that the infant's parents can be told about the abnormality, to remove the infant from continued exposure, and to get the infant into a medical care facility to be treated.

Why the Need to Protect Individuals From Unique Identification?

Why is there a need to protect individuals from unique identification in medical care or medical research? First, there are ways to inappropriately use data derived from blood and tissue specimens to stigmatize individuals as having a specific disease. Individuals and families (in the case of a genetically linked condition) need to be protected in society from such stigmatization. HIPAA attempts to eliminate (reduce) harm to individuals based on misuse of individually identifiable health information, including attempts to exclude individuals from job and employment opportunities and attempts to exclude individuals from present and future entitlements within society.

Second, there are economic uses of blood and tissue samples derived from humans (which include

the development of medical products). While individuals may be willing to provide specimens for their own medical care, they may not be willing to donate a sample for research purposes. Or if they are willing to donate a sample for research purposes one time, they may not be willing to be pursued by a researcher or product manufacturer over time to provide additional specimens.

Third, by being uniquely identified as having a specific disease, an individual may be targeted for advertising related to medical products that can be used in managing and treating his or her disease. While certain individuals consider the receipt of such new product advertising an opportunity, other individuals may not want to be so targeted.

Finally, in the medical research setting, although an individual may be willing to allow his or her blood or tissue samples to be used in a study to test a particular scientific hypothesis or to donate blood or tissues to a data bank for future research (whose scientific hypotheses have not even been conceived of today), that individual may be willing to donate his or her specimen only if it is labeled in such a way that the data cannot be traced back to that donating individual (completely anonymous data).

Genetically Identifiable Health Information

This entry so far has considered the types of decisions that decisionally capable individuals are able to make on their own: the decision to participate in medical care and the decision to volunteer to participate in a research study. However, in the case of data of genetic origin, while the individual may not care about protecting himself or herself from possible harm related to data release and while the individual may be willing to donate a specimen for a present or future research study, the individual does not have the right, with only his or her own permission, to donate materials whose release could damage other genetically linked family members, even if the individual possesses decisional capacity. Those individuals genetically linked to one another can have this linkage identified on the basis of examination of DNA, RNA, unique proteins, and other biologic materials with the same shared characteristic. In research that requires as its substrate genetically linked material, the major question in need of clarification is the following: How does anyone secure the relevant informed consents from all relevant genetically linked individuals to allow this genetically based study to start up and then continue over time (or to allow specimens to be banked over time to allow future research on the specimens)? This key question still remains unanswered and open for continued debate and research.

Difficult De-Identification

Perhaps the most difficult case in which to attempt to de-identify data with respect to an individual involves clinical care of the one patient with a rare medical condition who is followed in one medical center. Here, simply the labeling of the individual with the name of the rare medical condition or disease process he or she has is enough to label that individual within that medical center uniquely. If that individual is the only individual with that rare disease in the town, city, county, state, or nation in which he or she lives, the problem of de-identification (nonuniquely identifying the individual) will remain with that individual throughout his or her life.

Dennis J. Mazur

See also Decisions Faced by Institutional Review Boards; Informed Consent

Further Readings

Charo, R. A. (2006). Body of research: Ownership and use of human tissue. *New England Journal of Medicine, 355,* 1517–1519.

GAIN Collaborative Research Group. (2007). New models of collaboration in genome-wide association studies: The Genetic Association Information Network. *Nature Genetics, 39,* 1045–1051.

Ginsburg, G. S., Burke, T. W., & Febbo, P. (2008). Centralized biorepositories for genetic and genomic research. *Journal of the American Medical Association, 299,* 1359–1361.

Kauffmann, F., & Cambon-Thomsen, A. (2008). Tracing biological collections: Between books and clinical trials. *Journal of the American Medical Association, 299,* 2316–2318.

Mazur, D. J. (2003). Influence of the law on risk and informed consent. *British Medical Journal, 327,* 731–734.

Mazur, D. J. (2007). *Evaluating the science and ethics of research on humans: A guide for IRB members*. Baltimore: Johns Hopkins University Press.

Milanovic, F., Pontille, D., & Cambon-Thomsen, A. (2007). Biobanking and data sharing: A plurality of exchange regimes. *Genomics, Society, and Policy, 3,* 17–30.

Moffatt, M. F., Kabesch, M., Liang, L., Dixon, A. L., Strachan, D., Heath, S., et al. (2007). Genetic variants regulating ORMDL3 expression contribute to the risk of childhood asthma. *Nature, 448,* 470–473.

Summary of the HIPAA Privacy Rule. (2003). Retrieved June 5, 2008, from http://www.hhs.gov/ocr/privacy/hipaa/understanding/summary/privacysummary

Topol, E. J., Murray, S. S., & Frazer, K. A. (2007). The genomics gold rush. *Journal of the American Medical Association, 298,* 218–221.

Yuille, M., van Ommen, G. J., Bréchot, C., Cambon-Thomsen, A., Dagher, G., Landegren, U., et al. (2008). Biobanking for Europe. *Briefings in Bioinformatics, 9,* 14–24.

Health Outcomes Assessment

Have population-wide death rates and disability levels attributable to disease *X* declined over the past decade? In a randomized clinical trial comparing drug *Y* with standard therapy, is there a clinically important difference in patient survival or in patient-reported outcomes (PRO) such as symptom bother? In an economic evaluation of screening for disease *Z* annually rather than semiannually, what are the estimated differences in quality-adjusted life years (QALYs) per dollar spent? If a patient, working closely with her physician, is considering two therapies with similar projected survival benefits, how might she determine which provides the better health-related quality of life (HRQOL)? These quite diverse queries share a central common feature: They involve health outcomes assessment.

Health outcomes assessment (HOA) is a systematic and frequently multistep analytical *process* that may entail (1) identifying the health-related issue or problem to be investigated and the relevant audiences for the assessment (which may or may not be an identified decision maker); (2) selecting health outcome measures applicable to the problem at hand; (3) establishing an appropriate study design and collecting and analyzing the health outcomes data, often in conjunction with additional data deemed necessary for the particular assessment being done (e.g., one may want to draw inferences about the potential determinants of health outcomes); and (4) translating findings from (3) into information useful to the audiences identified in (1). By implication, health outcomes *measurement* is an essential step in health outcomes assessment but is not synonymous with health outcomes assessment. The health outcomes of interest pertain generally to quantity of life (mortality, survival, disease-free survival), quality of life (to encompass a range of PROs, including HRQOL and symptom bother), or both (as indexed, say, by the QALY).

As thus defined, health outcomes assessment may be viewed as a central task of health outcomes research, which, according to the U.S. Agency for Healthcare Research and Quality, "seeks to understand the end results of particular health care practices and interventions." Such end results are to be distinguished from "intermediate" outcomes (e.g., disease-screening rates) and "clinical outcomes" (e.g., changes in the individual's underlying medical condition). To be sure, medical interventions are very frequently aimed at improving such clinical or intermediate outcomes; health outcomes assessment asks the bottom-line question of whether such improvements translate into a longer life or better health.

The sections that follow discuss the health outcomes assessment process, with particular attention to linkages between the purpose of the assessment, the selection of specific outcome measures, and the translation of findings into useful information for the intended audience. That said, it is not the intent here to provide a detailed examination of each of the components of the multistep HOA process but rather to indicate how these considerations can be jointly brought to bear in the conduct of an assessment.

Areas of Application

Most health outcome assessments are designed to inform decision making in one of five areas of application: (1) population-level health

surveillance, (2) randomized controlled trials, (3) observational (nonrandomized) studies of intervention effectiveness, (4) cost-effectiveness analyses, and (5) patient–clinician deliberations about interventions and outcomes.

Population-Level Health Surveillance

This includes international, national, or subnational (e.g., state, regional) studies of health outcomes, either at the individual disease level or across diseases. Depending on the purpose of the study and the data available, the focus may be on trends in mortality, survival, or various PRO measures, including morbidity levels, symptoms, functional status, or HRQOL. The primary purpose of such surveillance studies, which are conducted routinely in some form by most developed nations and by international organizations such as the World Health Organization, is to inform policy discussions and the research agenda by revealing successes, shortcomings, and issues requiring more intensive investigation.

Considerable progress has been achieved in North America and Europe in the calculation of mortality and survival rates in a consistent fashion and in the application of multidimensional HRQOL instruments. For example, the SF-12 instrument is routinely administered as one component of the U.S. Medical Care Expenditure Panel Survey (MEPS), while variants of the SF-36 instrument are used in the ongoing Health Outcomes Survey of enrollees in Medicare managed care plans, conducted by the U.S. Centers for Medicare & Medicaid Services. The EQ-5D, a HRQOL measure designed to incorporate population preferences for health outcomes, has been used in several representative surveys of the U.K. population by Kind and colleagues. The Health Utilities Index, another preference-based HRQOL measure, is being applied on an ongoing basis across Canada and also in the Joint Canada-U.S. Survey of Health (JCUSH).

Regarding the HOA process delineated earlier, most population-level assessments focus on Steps 1, 2, and 4; historically, there has typically been less emphasis on sorting out the determinants of variations in population health. However, there is a growing interest in identifying disparities in health (and access to healthcare) across population subgroups defined by race/ethnicity, demographics, or geography.

Randomized Controlled Trials

The purpose of health outcomes assessment in most experimental studies of drugs, devices, biologics, or other interventions is clear: to generate evidence on safety, efficacy, and clinical benefit to inform regulatory decisions about product approval and labeling and (subsequent) decision making by purchasers, providers, and patients.

Recent developments in oncology offer a particularly rich opportunity for examining these issues in a concrete way. As officials of the U.S. Food and Drug Administration have written, approval of cancer drugs is based on "endpoints that demonstrate a longer life or a better life." From these officials' published reviews of cancer regulatory decisions spanning the period 1990 to 2006 in total, clinical outcomes (primarily tumor response) clearly played an important role in the majority of approval decisions, although patient-reported outcome measures—particularly symptom relief—provided critical or supplementary support in a number of instances. These officials report, however, that in no case was a cancer drug approval based on a HRQOL measure.

In 2006, the FDA issued its own draft "guidance to industry" on the use of PRO data (including HRQOL) in medical product development to support labeling claims generally. In 2007, the National Cancer Institute (NCI) fostered the publication of a series of papers (in the *Journal of Clinical Oncology*) assessing the state of the science of PRO application in cancer trials supported through the NCI. Also in 2007, a series of papers interpreting and evaluating the FDA PRO draft guidance appeared in another scholarly journal (*Value in Health*).

In sum, this is an era of intense debate about health outcomes assessment in clinical trials, particularly regarding the choice of appropriate end points and the closely related issues of study design, data collection, and analysis (corresponding to Steps 2 and 3 in the HOA process). While oncology studies have been very much in the spotlight, similar issues arise in any clinical trial where the patient's own perspective is regarded as an essential element in the outcomes assessment. Moreover, these issues have received analogous critical attention outside the United States.

Observational Studies of Intervention Effectiveness

There is a vast literature examining the impact of interventions—ranging from prevention activities, to disease screening, to treatments that may be surgical, medical, or radiological, or other—on health outcomes in the real-world practice of medicine. Depending on the disease and the purpose of the study, the focus may be largely on survival outcomes (e.g., do AIDS patients receiving a certain drug cocktail have a longer life expectancy?); HRQOL outcomes (e.g., do rheumatoid arthritis patients receiving a new disease-modifying agent report better functioning and less pain than before?); or both (e.g., do patients with two-vessel heart disease have better quality-adjusted survival with angioplasty or with coronary artery bypass surgery?).

Still relatively rare are longitudinal health outcomes assessments to track over time the impact of interventions on HRQOL, satisfaction with care, and other PROs, in addition to survival. A noteworthy example is the NCI-supported Prostate Cancer Outcomes Study, which has followed more than 3,500 newly diagnosed patients for up to 60 months, attempting to survey each at four different time points regarding symptom bother, functional status, and other aspects of HRQOL as well as about satisfaction with care and with the outcomes being experienced.

Because the validity of such observational (nonrandomized) studies may be threatened by selection effects (i.e., the subjects choosing Intervention A may not be comparable to those choosing Intervention B, in ways that may not be observable to the analyst), certain statistical correctives are increasingly being applied. These include both instrumental variable and propensity scoring techniques, two approaches in pursuit of a common aim: namely, to permit valid inferences about the impact of some hypothesized causal factor (e.g., a healthcare intervention) on a dependent variable of interest (e.g., a health outcomes measure) when it is likely that the variables are codetermined (mutually causal).

In general, the steps within the HOA process requiring the greatest attention here are the choice of outcomes measure(s) and study design and data analysis issues (i.e., Steps 2 and 3). In contrast to most clinical trials, the majority of such nonrandomized intervention studies inform decision making (if at all) in a generally more indirect or diffused way.

Cost-Effectiveness Analyses

In economic evaluations of whether a candidate intervention (e.g., individualized smoking cessation therapy) offers good value for the money compared with some alternative (e.g., an antismoking ad campaign), health outcomes assessment plays a pivotal role. This is because the "value" component of the cost-effectiveness analysis (CEA) is measured in terms of health outcomes improvement—for example, life-years gained or (most commonly now) QALYs gained. To carry out the CEA, therefore, requires sound statistical evidence on the health impact of each competing intervention, as would typically be derived from randomized or observational studies. Also required is information on the associated costs of each intervention and on a host of other factors (covariates) that allow the health and cost calculations to be tailored to specific population subgroups.

A prominent example of such a CEA, carried out in close conformance to the recommendations of the U.S. Panel on Cost-Effectiveness in Health and Medicine, is the study by Ramsey and colleagues comparing lung volume reduction surgery with medical management for elderly emphysema patients. CEA ratios in terms of dollars per QALY gained were computed for patients at varying degrees of clinical severity and under a variety of other clinical, economic, and statistical assumptions. This CEA and the randomized clinical trial on which it was based were sponsored by the U.S. Centers for Medicare & Medicaid Services (CMS), which covers virtually all Medicare-eligible patients and had a direct interest in the findings. Following publication of the trial and CEA findings, CMS approved coverage of lung volume reduction surgery for Medicare-eligible patients meeting specific clinical and behavioral criteria.

Frequently, such CEAs are conducted to inform public- or private-sector clinical policies relating to practice guidelines development or coverage decisions. In these instances, the process of health outcomes assessment feeds into the larger process of health economic evaluation.

Patient-Clinician Decision Making

There is growing interest, experimentation, and real-world application of health outcomes assessment to enhance the substantive content and overall quality of communications between patients and their healthcare providers. The aim is to strengthen shared decision making about intervention strategies and, ultimately, to improve patient outcomes. In most applications to date, patients complete questionnaires—focusing typically on aspects of their health-related quality of life—and the information is fed back to clinicians to inform healthcare management decisions.

Compared with the other areas of application, this use of health outcomes assessment is still in its infancy. There are promising results from some studies, including at least one randomized, controlled trial, indicating that providing clinicians with feedback on the patient's HRQOL status can favorably influence the perceived quality of communications and the patient's subsequent HRQOL. However, for this application of health outcomes assessment to realize its potential, several challenges must be confronted. These include strengthening the theoretical basis for anticipating and interpreting the impact of PRO measurement on decision making in routine clinical practice; understanding better how HRQOL measures developed originally to assess the impact of interventions on *groups* of patients can be informative for *individual-level* decision making; identifying targeted, patient-appropriate interventions based on responses to HRQOL questionnaires; and developing more user-friendly software to facilitate data collection and sharing. Progress on all fronts is expected to accelerate in the years ahead.

Informing Decision Making

Health outcomes assessment may be viewed as a multistep process, which progresses through identifying the decision problem to be addressed, the selection of appropriate outcomes measures, and the design and execution of the assessment itself, to the translation of findings to the intended audience(s) of analysts and decision makers. An assessment that carefully considers these steps in turn has the highest likelihood of successfully informing decision making.

Joseph Lipscomb, Claire F. Snyder, and Carolyn C. Gotay

See also Cost-Effectiveness Analysis; Mortality; Oncology Health-Related Quality of Life Assessment; Outcomes Research; Propensity Scores; Quality-Adjusted Life Years (QALYs)

Further Readings

Agency for Healthcare Research and Quality. (2000). *Outcomes research fact sheet*. Retrieved January 20, 2009, from http://www.ahrq.gov/clinic/outfact.htm

Clauser, S. B. (2004). Use of cancer performance measures in population health: A macro-level perspective. *Journal of the National Cancer Institute Monograph, 33,* 142–154.

Greenhalgh, J., Long, A. F., & Flynn, R. (2005). The use of patient reported outcome measures in routine clinical practice: Lack of impact or lack of theory? *Social Science & Medicine, 60,* 833–843.

Lenderking, W. R., & Revicki, D. A. (Eds.). (2005). *Advancing health outcomes research methods and clinical applications*. McLean, VA: Degnon.

Lipscomb, J., Gotay, C. C., & Snyder, C. (Eds.). (2005). *Outcomes assessment in cancer: Measures, methods, and applications*. Cambridge, UK: Cambridge University Press.

Patrick, D. L., Burke, L. B., Powers, J. H., Scott, J. A., Rock, E. P., Dawisha, S., et al. (2007). Patient-reported outcomes to support medical product labeling claims: FDA perspective. *Value in Health, 10*(Suppl. 2), 125–137.

Shortell, S. M., & Richardson, W. C. (1978). *Health program evaluation*. St. Louis, MO: C. V. Mosby.

Tunis, S., & Stryer, D. (1999). *The outcomes of outcomes research at AHCPR: Final report*. Retrieved June 30, 2008, from http://www.ahrq.gov/clinic/out2res

Health Production Function

The health needs of any population will be considerable, and there will never be enough resources to meet them all. As resources are scarce in relation to needs, they must be allocated by some mechanism, driven by the government or the private market.

To understand the results of the various ways that resources can be allocated, economists have developed a concept called the production function.

A production function is a mathematical concept that expresses a relation between resources and the outputs which the resources produce. The venue of the production can be a department, a plant, or a business firm. An economic "actor" or "manager" is responsible for combining the resources so they yield the outputs. An exemplified expression for a production function is $Q(L, K, t)$, where Q is an output (e.g., bushels of wheat), L represents the quantity of labor input, and K represents the quantity of capital equipment. The term t represents the stage of technological knowledge, which generally changes over time.

Hypotheses

The production function was initially used to predict how resources were combined to produce physical outputs of plants or firms, such as steel producers, shoe manufacturers, or agricultural firms. There are four primary hypotheses that can be generated from this concept. These are as follows:

1. As one resource increases while the others are held constant, the additional output from this increase will eventually decline. As an example, successive increases in receptionist time in a clinic (with other resources such as physician time and equipment held fixed) will initially lead to increased clinic output (visits). But this will happen only up to a point. Beyond that level, as more receptionist time is successively added, the increase in visits will become smaller.

2. As all resources increase together, in equal proportions, the *additional* output resulting will initially increase in a greater proportion ("economies of scale"), level off ("constant economies"), and then decrease ("diseconomies of scale"). For example, a moderate size clinic will have a greater productivity (in terms of visits per resource unit) than will a very small one. However, there are limits to which the productivity will increase with expansions in clinic size.

3. Over time, the entire curve will increase, indicating that more output can be achieved with the same quantity of resources ("technological change" or "increases in productivity"). In recent years, the mechanization of lab services has led to a reduction in the total resources that are used to produce lab tests. This means an upward shift in the production curve that relates outputs (lab tests) to inputs. It should be noted that quality of care is a component of output, though one that is difficult to measure.

4. The incorporation of one set of resources in two different applications within the same organization (a nursing home and a hospital), rather than in separate organizations, can result in economies (or diseconomies) of scope. For example, if we divide the hospital into distinct diagnostic units (pediatric, geriatric, cancer care), then economies of scope will be evidenced if a multiservice unit that incorporates many of these is less costly than a series of separate specialized hospitals.

Application to Medical Care

The production function concept has been used to explain how changes in the use of resources can yield different volumes of output (i.e., services produced) of *medical care*. Thus, John Cawley has examined how the use of specific types of capital-intensive services or equipment (catheters, tube feeding, psychotropic drugs), when used in nursing homes, will influence the use of labor (nurses' time) per resident day. That is, the capital equipment can be substituted for labor, keeping the amount of services produced at the same level. Production functions can also be used to explain factor substitution in health maintenance organizations, physicians' offices, and hospitals.

In 1965, Gary Becker published an article on the allocation of time within the household. The production function, formerly used to analyze how business firms combined resources, could now be used to analyze how households could combine resources, including purchased inputs and personal time, to produce activities that yield consumer benefits. Michael Grossman used this model in 1972 to examine how persons could use resources such as time and medical care to produce the output "health."

The first empirical studies of health production functions used mortality rates as a measure of health output (now called outcomes). A 1971 study

by Charles Stewart characterized resources in four groups: treatment, prevention, information, and research. In estimating the production function, Stewart showed that in developed countries, physician inputs, when increased, had an insignificant impact on life expectancy. The statistical procedure used in this model was challenged by Edward Meeker and Ronald Williams; when population health status and more appropriate statistical measures were taken into account, the statistical models showed a statistically significant impact of physician density on mortality, though not a large one.

The most influential studies associated with health production come from longitudinal analyses of mortality by David Cutler and Frank Lichtenberg. Responding to the noticeable improvements in mortality rates of selected groups since the 1970s, Cutler conducted several analyses of health production for people with different health conditions, including heart attacks and lung cancer. The findings are that trends in medical care for heart attacks in the past 30 years have resulted in improvements in health productivity. The same has not been true for lung cancer. Lichtenberg analyzed the impact of new drugs on mortality with positive findings as well.

In the studies discussed, outcomes were expressed in physical units, such as years of life or age-adjusted mortality: Production functions are concepts that relate physical units of resources to physical measures of outputs. Some analysts have put a dollar value on the changes in mortality. Using results obtained from estimating the value that people put on changes in mortality from other studies, the investigators have compared the value of increased longevity with the costs of the resources. Valuations placed on changes in mortality do indicate what people are willing to pay for the increases in life spans, but they are not necessarily the same valuations that will be used for making policies, because they place a low value on the health of the poor and destitute. In addition, analysts have attached prices to individual resources, which has yielded measures of the cost of production. Behind the money cost and benefit measures that are obtained from these calculations lie the more fundamental measures of physical relationships between physical inputs and health states. That is, cost and benefit calculations are derivative from the production function.

Recently, investigators have extended the health production function by examining outcomes that are reflective of changes in health status, not only in mortality. Among the measures of health status that have been used are time spent working and health-related quality of life (HRQOL). Not all the studies have related the changes in resources during a time span (e.g., a year) with changes in health status over the same time span; some have measured a relation between the health status at the beginning, or end, of a period with the quantities of resources used during the same period. While the use of personal and medical resources during a given year can indeed have an impact on health status after (or before) the year is over, the use of the health status at the end of the period as an outcome measure is not an appropriate indicator to use in the health production function. The use of health resources will have an impact on *changes* in health status during the year or afterwards. The health status at the start of the year, in addition to how health status changes during the year, will affect the health status at the year's end. To get around this problem, investigators who use the health status at the end of the period in their studies have used the starting health status as an independent variable in their statistical analyses (e.g., Hakkinen, Jarvelin, Rosenqvist, & Laitinen, 2006; Lu, 1999).

Influence

The health production function has been widely used to conceptualize the impact of resources or changes in health states. The results of research inform, and probably influence, policy decisions. Earlier studies in this area raised skepticism about the overall effectiveness of adding more medical resources to the healthcare system. Policy makers and policy analysts were influenced by these results, and terms such as "flat of the curve medicine" became popular as descriptions of the state of use of health (and especially physician) services. More recent studies have shown that medical, pharmaceutical, and personal resources all have impacts on health status. If further studies corroborate these results, policy decisions will be influenced by this information.

Philip Jacobs

See also Cost-Effectiveness Analysis; Economics, Health Economics; Efficacy Versus Effectiveness; Efficient Frontier

Further Readings

Becker, G. (1965). The theory of the allocation of time. *Economic Journal, 299,* 493–517.

Cawley, J., Grabowski, D. C., & Hirth, R. A. (2006). Factor substitution in nursing homes. *Journal of Health Economics, 25,* 234–247.

Cutler, D. M., & McClellan, M. (2001). Is technological change in medicine worth it? *Health Affairs, 20,* 11–29.

Grossman, M. (1972). On the concept of health capital and the demand for health. *Journal of Political Economy, 80,* 233–255.

Hakkinen, U., Jarvelin, M. R., Rosenqvist, G., & Laitinen, J. (2006). Health, schooling and lifestyle among young adults in Finland. *Health Economics, 15,* 1201–1216.

Kenkel, D. S. (1995). Should you eat breakfast? Estimates from health production functions. *Health Economics, 4,* 15–29.

Lichtenberg, F. R. (2003). The economic and human impact of new drugs. *Clinical Psychiatry, 64*(Suppl. 17), 15–18.

Lu, M. (1999). The productivity of mental health care: An instrumental variable approach. *Journal of Mental Health Policy and Economics, 2,* 59–71.

Meeker, E. (1973). Allocation of resources to health revisited. *Journal of Human Resources, 8,* 257–259.

Rosen, A. B., Cutler, D. M., Norton, D. M., Hu, H. M., & Vijan, S. (2007). The value of coronary heart disease care for the elderly: 1987–2002. *Health Affairs, 26,* 111–123.

Stewart, C. T. (1972). Allocation of resources to health. *Journal of Human Resources, 6,* 103–122.

Williams, R. L. (1975). Explaining a health care paradox. *Policy Sciences, 6,* 91–101.

Woodward, R. M., Brown, M. L., Stewart, S. T., Cronin, K. A., & Cutler, D. M. (2007). The value of medical interventions for lung cancer in the elderly: Results from SEER-CMHSF. *Cancer, 110,* 2511–2518.

Health Risk Management

Decision making is a critical element in the field of medicine that can lead to life-or-death outcomes, yet it is an element fraught with complex and conflicting variables, diagnostic and therapeutic uncertainties, patient preferences and values, and costs. Judgments and decisions made daily in clinical work necessitate the assessment and management of risks. The physician must determine what may be wrong with a patient and recommend a prevention or treatment strategy, generally under less-than-optimal circumstances and time frames. A patient decides whether or not to follow this recommendation and, once under care, may or may not faithfully pursue a recommended strategy. Health policy makers and insurers must decide what to promote, what to discourage, and what to pay for. Together, such decisions determine the quality of healthcare, quality that depends inherently on counterbalancing risks and benefits and competing objectives such as maximizing life expectancy versus optimizing quality of life, or quality of care versus economic realities.

Therefore, diagnostic reasoning and treatment decisions are a key competence of physicians and have been attracting increasing interest. Reasoning skills are imperfect in many clinical situations, and it has been found that diagnostic errors are more frequently a result of failure to properly integrate clinical data than of inaccurate data. Diagnostic experts use relatively few clinical data, with modes of reasoning sometimes oversimplified. These limitations are connected to several aspects of clinical decision making; one of these aspects is to acknowledge components of knowledge used in clinical practice.

Moreover, the literature on the reasoning process is often unfamiliar to physicians, and studies of diagnostic reasoning are often simpler than the diagnostic reasoning in real-life situations. Although studies provide information about the outcomes of decisions, they provide little or no information about the process of the decision.

How are sound treatment decisions determined? Are they based on the value of the outcome or the probability of the outcome? Are judgments and decisions based on both variables, or are the simplifying strategies employed by experts based on only one of the variables? Judgments and decisions are made daily in clinical work, where the assessment of risk is necessary. Risk is involved in the choice of tests to use in reaching a diagnosis. There is also uncertainty and risk in interpreting test

results. With lab tests indicating an infection, what level of antibiotics should be used in treatment? What other factors should a doctor consider in a diagnosis and treatment?

With this uncertainty taken into consideration, how should information from clinical and biomedical knowledge be combined to reach a diagnosis? Is there an additive relationship between different sources of information, or is there a multiplicative relationship? That is, is the interpretation of risk dependent on risk in still another variable? A high temperature can be interpreted in a certain way in one context, but given another picture of symptoms, it is interpreted another way. With a diagnosis obtained with some certainty, what treatment should be chosen? In all these cases, there is risk involved for multistage decision problems.

Modeling and Risk

Mathematical modeling is used widely in economic evaluations of pharmaceuticals and other healthcare technologies. Clinical decision making may benefit from the same modeling approach, since the task of the healthcare provider is to provide care and to incorporate the probability of obtaining certain health outcomes, whether explicit or implicit; the latter varies with providers and in many cases may not be done at all. Weighting the value of an outcome by the probability of its occurrence provides both patient and provider with information about decision making.

A model based on values and beliefs provides a conceptual framework for clinical judgments and decisions; it also facilitates the integration of clinical and biomedical knowledge into a diagnostic decision. From this perspective, decision research in health has increasingly recognized evaluated value and probability of outcome in explaining judgments and decisions in various domains, seeing them as based on the product of these two parameters, termed *expected value*. This is a prescription approach, however, and is often inconsistent with how people generally make decisions.

In clinical decision making, the values are healthier outcomes in various variables. The outcome assessment variables in rheumatoid arthritis have been pain, disability, and overall health. These variables are assessed by the patient. For the patient with a heart attack, one assessment variable could be decreased pain and normalized electrocardiography another.

The probability for a certain outcome to occur will also have to be estimated in these diagnostic and treatment decisions.

Both value and probability are usually estimated values in clinical decision making. Therefore, model assumptions and parameter estimates should be continually assessed against data, and models should be revised accordingly. Estimated values and probabilities are involved sequentially for every step in the decision-making process. However, a dichotomous decision will have to be performed to reach a diagnosis and a treatment option. Moreover, there might be many differential diagnoses to exclude and also many treatment options. The number of differential diagnoses considered, and what they are, might have an influence on the diagnosis finally selected. The availability of treatment options might also affect what treatment is chosen.

Risk and Errors

One issue is the way clinical inferences generally are arrived at in making judgments and decisions. Theories have been provided about how doctors could include relevant information to improve decision making. Nonetheless, a reasoning error could be made in clinical inference, as it is characterized by backward reasoning, where diagnosticians attempt to link observed effects to prior causes. In contrast to this post hoc explanation, statistical prediction entails forward reasoning, because it is concerned with forecasting future outcomes given observed information.

Clinical inference uses information from prior periods to make a statement about today and tends to consider error as a nuisance variable. The statistical approach, on the other hand, accepts error as inevitable and, in so doing, probably makes fewer errors in prediction for periods extending over a relatively long time. Moreover, the statistical approach uses group data to arrive at a conclusion. The situation is different in clinical inference and decision making, where group data concerning risk constitute the basis for diagnostic and treatment choices regarding the individual patient.

It has also been found that doctors exhibit interindividual as well as an intra-individual variation

in judgments. One example in practical work is the outcome of clinical examinations, which may vary between doctors. Another example is the interpretation of radiological pictures, which may exhibit a variation between doctors.

Many people tend to overestimate how much they know, even about the easiest knowledge tasks. Overconfidence (i.e., greater certainty than circumstances warrant) leads to overestimating the importance of occurrences that confirm one's hypothesis. This impedes learning from environmental feedback, resulting in deleterious effects on future predictions. In many decision settings, inexperienced practitioners and even naive laboratory subjects perform as well (or as poorly) as performers with more experience. The performance of the patient could be as good or as bad as these subjects.

Daily work with patients implies considering risks at many stages of the decision process. How does one convey to patients this information about risk and error as an unavoidable condition in clinical work to reach a mutual agreement on treatment judgments and decisions? Through an awareness of errors that can be made, some errors can be counteracted. Thus, a challenge for clinical practice is to include different features of risk.

Shared Decision Making by Doctors and Patients

The application of evidence-based medicine requires combining scientific facts with value judgments and with the cost of different treatments. This procedure can be approached from the perspective of doctors or of individual patients. Doctors may not value various aspects of health the same way patients do, and studies on patient control have found that patients generally respond positively to increased information.

However, research in cognitive psychology has shown that people are quickly overwhelmed by having to consider more than a few options in making choices. Therefore, decision analysis, based on the concepts of value and risk, might be expected to facilitate clinical judgments and shared decision making by providing a quantifiable way to choose between options. Overall, likelihood of a specific adverse outcome should be one parameter affecting the estimate of future risk and its consequences. Risk estimates of future outcomes could be based on an outcome in the future having less importance than one in the present, where the adverse outcome may have different values for doctor and patient. Another parameter is that temporal distribution of risk is not homogeneous throughout the life span of the individual. Specific individual factors modify the risk for a specific person, and person-specific modifiers are likely to be distributed differently in time.

Patients dealing with chronic illness are increasingly knowledgeable. They must make multiple and repetitive decisions, with variable outcomes, about how they will live with their chronic condition. With rheumatoid arthritis, for instance, that demands lifelong treatment, doctor and patient share not one single decision but a series of decisions concerning treatment. Furthermore, in current healthcare, several doctors may be involved in the treatment.

Risk levels are adopted in a context, and their impact on decisions may be arbitrary when the norm for decisions involving risk is being set. With an uncertainty in diagnosis, at what risk level will treatment be chosen by both the doctor and the patient? How do patients and doctors estimate different variables? Psychological factors such as personal versus general risk, where personal risk relates to oneself and general risk to others and policy approaches, may have an impact on decisions. In a clinical decision-making situation, personal risks can be assumed to relate to the patient, while the doctor has a general perception of risk.

Perhaps patients might give a higher estimation of risk, being more conservative because the outcomes of decisions are more significant for them; it is their bodies and their lives that are affected. On the other hand, it is well-known that personal risks are underestimated; people judge their own risks from adverse health behaviors as smaller than the same risks for people in general. People's opinions about personal risk are generally too optimistic, whereas the perceived risk for others is more adequate.

In a study by Ayanian and Cleary, most smokers did not view themselves as being at increased risk of heart disease or cancer. The low perceived personal risk could tentatively be explained by risk denial. It has also been found that, in individual decision making, there is a preference for the

low-risk treatment. In societal choices, however, treatment of the high-risk patient groups is preferred. Social framing may therefore induce a propensity to prefer interventions that target high-risk populations. These preferences were performed by healthy individuals.

Patient Satisfaction

There is an increased awareness of patients' involvement in the clinical decision process, where patients and providers consider outcome probabilities and patient preferences. Assessing health values and beliefs may help providers understand their patients' treatment behavior and increase patients' satisfaction with services and their motivation to comply with treatment regimens.

With chronic conditions, patients are increasingly knowledgeable about their medical condition. The challenge is to balance advocacy for an active patient role with the preferences of individual patients concerning participation. Agreement between physicians and patients regarding diagnosis, diagnostic plan, and treatment plan has been associated with higher patient satisfaction and better health status outcomes in patients.

To be effective, the clinician must gain some understanding of the patient's perspective on his or her illness. Introducing decision-analytic modeling provides a more complete picture of variables that influence the decisions performed by doctor and patient and can contribute to skillful counseling around unhealthy or risky behaviors, an important aspect of the communication that should be part of healthcare visits.

Monica Ortendahl

See also Applied Decision Analysis; Errors in Clinical Reasoning; Risk-Benefit Trade-Off; Risk Communication; Shared Decision Making

Further Readings

Ades, A. E., Lu, G., & Claxton, K. (2004). Expected value of sample information calculations in medical decision modeling. *Medical Decision Making, 24,* 207–227.

Ayanian, J. Z., & Cleary, P. D. (1999). Perceived risks of heart disease and cancer among cigarette smokers. *Journal of the American Medical Association, 281,* 1019–1021.

Groves, M., O'Rourke, P., & Alexander, H. (2003). The clinical reasoning characteristics of diagnostic experts. *Medical Teacher, 25,* 308–313.

Kempainen, R. R., Migeon, M. B., & Wolf, F. M. (2003). Understanding our mistakes: A primer on errors in clinical reasoning. *Medical Teacher, 25,* 177–181.

Klayman, J., Soll, J. B., Gonzales-Vallejo, C., & Barlas, S. (1999). Overconfidence: It depends on how, what and whom you ask. *Organizational Behavior and Human Decision Processes, 79,* 216–247.

Slovic, P. (2000). *The perception of risk*. London: Earthscan.

Teutsch, C. (2003). Patient-doctor communication. *Medical Clinics of North America, 87,* 1115–1145.

Weinstein, M. C., O'Brien, B., Hornberger, J., Jackson, J., Johannesson, M., McCabe, C., et al. (2003). Principles of good practice for decision analytic modeling in health-care evaluation: Report of the ISPOR Task Force on Good Research Practices—Modeling Studies. *Value in Health, 6,* 9–17.

Health Status Measurement, Assessing Meaningful Change

Sensitivity is the ability of an instrument to measure change in a state irrespective of whether it is relevant or meaningful to the decision maker. Responsiveness is the ability of an instrument to measure a meaningful or clinically important change in a clinical state. Responsiveness, like validity and reliability, is not necessarily a generalizable property of an instrument and should be assessed for each population and for each purpose for which it is used. *Sensitivity* and *responsiveness* can refer to assessments of groups or individuals. Responsiveness is equivalent to longitudinal construct validity, where the ability of an instrument to measure a clinically meaningful change is evaluated. Sensitivity to change is a necessary but insufficient condition for responsiveness. Some dislike using *sensitivity* because it might be confused for terms often linked to sensitivity, such as specificity, positive predictive value, and negative predictive value, used in the description of diagnostic test performance. However, health status questionnaires are analogous to diagnostic tests in that they can be used in medical decision making to determine whether an individual has a condition or

disease; to screen individuals for incipient disease, disability, or risk of either; and to monitor the course of a disease or the response to treatment.

Measures of generic and disease-specific health status are sensitive to changes in clinical status when applied to groups of patients. They are as sensitive as or more sensitive than many traditional measures, such as performance tests and laboratory evaluation of disease activity. However, it is unclear whether these instruments can capture meaningful changes in subgroups or in individuals. Indeed, most instruments show that data cannot take on a value higher than some "ceiling" (ceiling effect) or lower than some "floor" (floor effect), which indicates that they cannot be used for the entire continuum of patients seen. Controlled studies evaluating the utility of providing health status data do not show that outcomes, health resource use, or costs are affected. A number of explanations are possible: (1) Physicians are not trained to interpret such data or to determine what should be done to improve function, (2) the information was not provided in a timely manner, (3) diminished function and well-being are distal end points in the chain of causation and present fewer opportunities to affect their course, or (4) measures used to assess groups of individuals are imprecise, insensitive, and unresponsive to clinically important changes.

Studies show that patients and their healthcare providers may disagree about health priorities, quality of life, functional ability, psychological state, and the importance or magnitude of the change captured by questionnaires. Patients can mean different things when they say they are "better." Response-shift or instrumentation bias, recall bias, and amnestic bias can also affect the measurement and the perception of change.

A clinically meaningful or important change can be defined and therefore evaluated from the perspective of the patient, his or her proxy, society, or the health professional. It implies a change that is noticeable, appreciably different, that is of value to the patient (or physician). This change may allow the individual to perform some essential tasks or to do them more efficiently or with less pain or difficulty. These changes also should exceed variation that can be attributed to chance.

Some investigators have defined a clinically significant change as a return to usual functioning, but this is a stringent criterion for many chronic conditions. Others have defined "clinically meaningful" as whether an individual has surpassed some absolute criterion, but this definition does not permit one to document a change that is important but is short of the absolute criterion. Roman Jaeschke and colleagues suggested that a clinically meaningful change could be defined as the minimal important difference. This could be defined as the smallest difference in score in the domain of interest that a patient perceives as a change and that would mandate, in the absence of side effects and excessive costs, modification in the patient's management. Others have advocated that the maximum improvement is more important clinically.

Methods for Evaluating Sensitivity

Statistical techniques to estimate the sensitivity of an instrument vary, and there is no apparent consensus regarding the preferred technique. Many are variants of the effect size statistic and resemble the *F* statistic in analysis of variance (see Table 1).

It is not clear that the methods would show the same rank order of sensitivity when different instruments are compared or whether the observed findings might have occurred by chance.

If all measures of sensitivity rank an instrument high or low relative to other instruments, then one may be relatively confident that that instrument is the most sensitive. In this case, the question of which method is best is moot because all agree. If instruments change their rank order depending on which measure of sensitivity is used, then the instruments are probably roughly equivalent. Which method is best must be determined by other means.

Methods for Evaluating Responsiveness

In contrast to the methodologic work evaluating sensitivity, the significance of these changes and the techniques for evaluating responsiveness to clinically important or meaningful change have received little attention. No one technique has been established as being superior. In fact, different methods for the assessment of responsiveness may lead to different conclusions.

One way to study "meaningfulness" is to ask the subject, the person's provider, or both

Table I Approaches to statistical evaluation of sensitivity

Effect size
Effect size index
Guyatt's method
F ratios, comparison of
Measurement sensitivity
Receiver operator characteristics
Relative change index
Responsiveness coefficient
Standard error of measurement
Standardized response mean (relative efficiency)

(a) whether a change has occurred ("transition question"), (b) how large the change is, (c) how important or relevant the change is, and (d) how satisfied the subject is with the change. The judgment of any or all of these could be done by patients, by an external judge, or by the use of a related construct. If patients are asked about a meaningful change, the framing and timing of the questions in relation to the intervention need careful consideration, because the extent of recall bias is unknown. An external judge could be a healthcare professional uninvolved with the subject's care, or a caretaker such as a family member or significant other, when the subject may be unreliable. Related constructs, such as patient satisfaction with the change or a change that allows resumption of normal work or necessitates assistance, are also possibilities. A problem for all methods is that a change in a state (e.g., function) derives its significance and meaning to the subject or to a proxy from the starting state as much as anything else.

Researchers have generally chosen test items appropriate to the content domain (domain sampling method) for the construction of health status in an attempt to maximize overall internal reliability (Cronbach's coefficient [alpha] of tests). This strategy tends to maximize reliability at or near the center of a scale, often by having more items with an average level of difficulty than items with very great or very slight difficulty on a test. With this, a test may not discriminate equally across its whole range. Thus, a subject who is near the middle of the range may change a small amount on true ability and yet change more than 1 point on the scale score because there are many items in the region where the subject is making the change. However, a subject who is at the high or the low end of a scale, where there are fewer items, may actually make a much larger or clinically meaningful change and not have it captured on the scale or have a small change as compared with the subject who started at the center. For example, in the Health Assessment Questionnaire, a measure of physical function, a subject with severe rheumatoid arthritis may rate all tasks as being maximally difficult; and yet this subject can still worsen to the point of being confined to home or completely dependent on others. An uncritical adoption of classic psychometric techniques for scale construction to maximize overall internal reliability has led to scales that may be more responsive in group applications (clinically meaningful) at the ends of the scale but more sensitive to change (statistically) at the center of the scale.

Developing a scale where items are equally spaced in terms of difficulty across the entire range of the scale (equidiscriminating) is one focus of item response theory, of which Rasch models provide a one-dimensional approach. With an equidiscriminating scale, when a patient moves a particular number of points, one can be relatively sure that he

or she has moved the same distance on some true scale of difficulty.

A problem for both types of scales is that the perception of change in a state, such as health status, derives its significance and meaning in comparison with the starting state as much as any other referent. Studies suggest that perceived change of physical and sensory states may be a power function. For instance, persons who start at a low level of function on a scale and change a relatively small distance along the dimension may perceive the change as clinically significant. However, persons who start with much higher physical function may view the same size change as a trivial improvement and would need a much larger change to judge it as clinically significant. Thus, even "equidiscriminating" scales beg the question of whether the same amount of change in an underlying dimension is clinically significant at all levels or a function of the level at which one starts. An inherent limitation in scales measuring health status is that one cannot collapse all the subtleties of change into a single linear scale. For instance, a patient with arthritis can have a change in pain and a change in mobility, but each patient may attach a different utility to these changes. Collapse of these different utilities into one scale often compromises the individual utility functions. This is important because classic domain sampling assumes a single dimension along which persons are being measured, but most health status instruments actually measure several dimensions.

Potential assays for the evaluation of what constitutes a meaningful change on an instrument might involve the measurement of states in certain clinical situations: (a) in clinical trials or cohorts where the intervention has varying effectiveness, such as the surgical and conservative management of lumbar spinal stenosis or total joint arthroplasty; (b) after an effective medication is stopped; and (c) during the washout period in crossover studies.

Kirshner and Gordon Guyatt suggested an examination of the response of patients to a treatment of known efficacy and a comparison of the responses of patients who had and had not responded by the physician's judgment. Mark Lipsey recognized the logistical difficulties in this. He suggested identifying a group whose average response would be approximately the same as the desired detectable change and administering the instrument just once to estimate the change in variance.

No single standard exists for the evaluation of responsiveness. The point of view from which responsiveness is being evaluated should be specified. Patients' judgments are influenced by their baseline health status, expectations and goals, illness duration, and actual need to perform some functions, as well as other factors. These judgments vary as compared with results of standardized measures of function. The physician's judgment usually includes knowledge of other patients with the same problem, knowledge of what domains are potentially treatable, and an appreciation for the significance of physiological (e.g., creatinine clearance) or physical findings that may not be symptomatic or apparent to the patient. Proxies, such as caretakers or significant others, may be preferred when the respondent or patient's status may not be reliably or validly reported. For measures of function and quality of life, responsiveness should be based on the subject's valuation of the magnitude and its importance. For measures of impairment or disease activity, the physician is the best judge.

Allen J. Lehman and Matthew H. Liang

See also Health Outcomes Assessment; Health Status Measurement, Floor and Ceiling Effects; Health Status Measurement, Minimal Clinically Significant Differences, and Anchor Versus Distribution Methods; Health Status Measurement, Responsiveness and Sensitivity to Change

Further Readings

Beaton, D. E., Tarasuk, V., Katz, J. N., Wright, J. G., & Bombardier, C. (2001). "Are you better?" A qualitative study of the meaning of recovery. *Arthritis Care and Research, 45,* 270–279.

Daltroy, L. H., Larson, M. G., Eaton, H. M., Phillips, C. B., & Liang, M. H. (1999). Discrepancies between self-reported and observed physical function in the elderly. *Social Sciences and Medicine, 48,* 1549–1561.

Fitzpatrick, R., Ziebland, S., Jenkinson, C., Mowat, A., & Mowat, A. (1993). Transition questions to assess outcomes in rheumatoid arthritis. *British Journal of Rheumatology, 32,* 807–811.

Hollon, S. D., & Flick, S. N. (1988). On the meaning and methods of clinical significance. *Behavioral Assessment, 10,* 197–206.

Kirshner, B., & Guyatt, G. (1985). A methodological framework for assessing health indices. *Journal of Chronic Disease, 38,* 27–36.

Liang, M. H., Larson, M. G., Cullen, K. E., & Schwartz, J. A. (1985). Comparative measurement efficiency and sensitivity of five health status instruments for arthritis research. *Arthritis and Rheumatism, 28,* 542–547.

Liang, M. H., Lew, R. A., Stucki, G., Fortin, P. R., & Daltroy, L. H. (2002). Measuring clinically important changes with patient-oriented questionnaires. *Medical Care, 40,* II45–II51.

Lipsey, M. W. (1993). A scheme for assessing measurement sensitivity in program evaluation and other applied research. *Psychological Bulletin, 94,* 152–165.

Redelmeier, D. A., & Lorig, K. (1993). Assessing the clinical importance of symptomatic improvements: An illustration in rheumatology. *Archives of Internal Medicine, 153,* 1337–1342.

Health Status Measurement, Construct Validity

Just as it is important to ascertain that a measurement instrument produces reliable results across different situations, it is crucial to assess whether it measures what it is intended to measure—its validity. In the area of health status measurement, construct validation of measurement instruments underlies sound medical decision making. Validity is established through a process involving a series of experiments designed to test various relevant hypotheses about the structure and nature of the construct and its logical manifestations. The results of these experiments inform the level of confidence with which researchers make conclusions about the persons under study and the interpretation of instrument scores.

Early measurement of health focused heavily on disease and mortality rates for populations and on clinical variables representing disease activity for individuals. Over time, with the mounting challenges presented by chronic diseases and disorders, many health interventions have focused more on levels of physical, mental, and social functioning than on length of life. Under these circumstances, sound medical decision making about the value of healthcare interventions depends increasingly on the validity of instruments for measuring health status. The impact of validity on interpretation of clinical trials has been demonstrated empirically in psychiatry, with evidence that clinical trials using unvalidated measurement instruments were more likely to report treatment effectiveness than those employing validated measures. This entry describes the specific challenges involved in assessing the construct validity of health status measures, addresses the evolving conceptual framework for validity and associated taxonomy, explains the main approaches used, and provides additional resources for more in-depth discussion of theory and methods.

Challenges in Validation of Health Status

As compared with the measurement of physical attributes such as height and weight, the measurement of health status comes with special challenges because it is not a directly observable quantity but a construct; a variable that must be defined to be measured. The definition of the construct may originate from theory, clinical or empirical observation, or a combination of the two. Essentially a construct is itself a theory about what makes up the construct and how its component parts relate. Based on the definition, instrument developers decide what attitudes, behaviors, or characteristics would be the best indicators of the construct. For example, it is widely accepted that health status is a multidimensional construct that includes at least physical status, emotional or mental status, and symptoms. Potential indicators can be observable manifestations of the construct, such as behaviors, or they can be attitudes. Measurement may be conducted by observation of behaviors through performance tests or by eliciting subjects' reports of their behaviors. The measurement of attitudes requires posing questions that represent the attitude in question. For many constructs, including health status, self-report is the preferred approach to measurement; therefore, questions are used to tap aspects of each of the dimensions of health status, and responses are provided to allow numerical description or scaling. In the case of physical status, developers ask themselves, "What behaviors would represent a lot of (or little) physical function?" An ideal measurement instrument would cover the full range of relevant functional

activities, with a sufficient number of increments in response categories to measure differences across the functional continuum.

Because there is no gold standard for the definition or measurement of health status, there is not one definitive test for the construct validity of a health status measure. Rather, construct validation is a process of accumulating evidence from empirical tests of hypotheses. These hypotheses rest fundamentally on the definition of the construct, the structure and relationships of its components, and relationships to other relevant concepts. There are many hypotheses that can be constructed about the structure or behavior of a construct; therefore, the validation process is incremental, requiring evidence from various studies to provide a reasonable level of confidence about the validity of conclusions made using the instrument. On the other hand, one sound negative validation study can put the validity of the construct in question. Using a hypothetical physical function scale *X* as an example, we might hypothesize that physical function would decline with increased health problems and test the correlation between scores on Scale *X* and those of a previously validated index of comorbid health problems. If a meaningful association between the two measures could not be demonstrated, this would be considered a negative study. Under these circumstances, the validity of physical function as a construct would be questioned, and the definition and theory underlying it would be reassessed. However, interpretation of construct validation studies is complicated by the fact that these empirical tests examine the validity of the construct *and* the validity of the measure itself in the application under study. In this example, perhaps physical function was a valid construct, but the instrument did not adequately represent it. Therefore, a negative study may mean that (1) the construct is not valid, (2) the construct is valid but the instrument is inadequate to measure it under the circumstances, or (3) both.

Validity and the Three Cs: Content, Criterion, and Construct Validity

The literature on validity contains references to many types of validity, which can cause confusion about the fundamental concepts involved. Historically, validity was often viewed as a characteristic of the measurement instrument, having three separate components: content, criterion, and construct validity. Criterion and construct validity have been further divided into several categories, with various naming conventions depending on study designs employed. *Content validity* addresses the degree to which the questions included in the instrument sample the full range of the attribute of interest. For example, for physical function, one would ask whether the instrument of interest provides questions at the lowest level of function and at the highest possible level, with adequate sampling in between. *Criterion validity* refers to the process of validation when there is a gold standard against which the instrument of interest can be compared. In contrast, *construct validity* refers to when there is no gold standard, and the measured variable is a construct. Finally, *face validity* is a term readers may find in validity studies, which refers to whether the instrument measures what it is meant to measure "on the face of it." Face validity can be placed within the realm of content validity. It usually signifies that researchers or instrument developers elicited the opinion of experts in the field on whether the instrument represents key components of the construct of interest.

More recently, there has been a shift toward viewing validity in terms of the inferences to be made from the data produced by an instrument. Ultimately, the goal of construct validation is to establish the level of confidence with which inferences can be made using the instrument. Therefore, the purpose of the application must be considered in developing hypotheses for testing validity. It is common for instruments to be applied for different purposes than those for which they were originally intended, and the evidence for the validity of inferences made for novel applications must be assessed.

Approaches to Assessing Construct Validity

Extreme Groups

One approach to assessing the construct validity of an instrument is to test it on two groups chosen for their divergent characteristics relative to the construct. This can be called *extreme groups, known groups,* or *discriminative validity testing*. In the case of health status measures, one group

would be chosen for its low level of health relative to a second group. The research question would be whether the scores for each group were different, indicating that the measure could discriminate between the two groups. A range of indicators can be used to define the two groups, such as the presence of a particular diagnosis, number of comorbidities, or level of healthcare utilization. It is worth emphasizing that a negative study may mean that (a) the construct is not valid, (b) the instrument is not valid applied under these circumstances, or (c) both. These possibilities underline the need to consider the validity of inferences from an instrument relative to the range of evidence in support of its use for a particular purpose.

Convergent Validity

Hypotheses are developed to test whether the instrument of interest correlates with other measures of the same construct. For example, in assessing the validity of a generic health index for cost-utility analysis in spine disorders, a study was conducted to test correlations between several of the most widely used instruments for this purpose, finding correlations ranging from .57 to .72. In this case, the research hypothesis was that the instrument of interest would demonstrate a positive correlation with other instruments designed for the same general purpose. Do correlations from .57 to .72 support the validity of inferences made for medical decision making using these instruments among persons with spinal disorders? It is important to consider and explicitly address the acceptable range of correlation necessary to support validity when designing construct validation studies. The possible range of correlation coefficients is from –1, indicating perfect inverse relationship, through 1, indicating perfect positive relationship and includes 0, which indicates no association. Due to the existence of measurement error, the coefficient estimate must be less than 1. Furthermore, when correlating two measurement instruments, the maximum value for the correlation coefficient is given by the square root of the product of their reliabilities. In other words, the maximal correlation between two measurement instruments is likely to be meaningfully lower than 1.0. For example, for two instrument reliabilities of .88 and .90, the maximum correlation possible between them would be .89. For instruments designed for the same general purpose, such as health indexes for cost-utility analyses, differences in construct definition or conceptual frameworks underpinning the design would contribute to diminution of correlation from this maximal value. With this in mind, the correlations noted above between health indexes could be considered moderate to strong evidence of construct validity. In the case of testing a new instrument, extremely high correlations may be evidence of redundancy and require revisiting the rationale for the creation of a new instrument. Under these circumstances, the new instrument should provide important practical advancements or meaningful improvements in face or content validity.

Experiments are also conducted to assess correlations of the instrument of interest with measures of other related constructs. For example, it may be hypothesized that a generic health index applied among persons with spine disorders would correlate with a disease-specific disability index. Both of these approaches are called convergent validity, and in practice, multiple measures are used for comparison. Depending on the construct used for comparison, the degree of correlation expected will vary.

Discriminant Validity

To be a valid measure of health status, a new instrument not only should correlate with measures of similar and related constructs, but should not correlate with unrelated variables. Investigators ask, "What variables or constructs should not be correlated with the measure in question?" and design experiments to assess the relationship between the instrument and a seemingly unrelated variable. An unanticipated association may guide instrument developers to areas of potential improvement in the instrument.

Assessing the Internal Structure of the Construct

Assessing the internal structure of an instrument in relation to the theoretical framework for the construct can make important contributions to evidence for construct validity. Factor analysis is a key analytic tool that is used to describe the

relationships between questions or items in an instrument. For example, factor analysis can be used to test whether a health status measure designed to represent five dimensions of health (e.g., physical function, symptoms, mental health, self-care, and usual activities) actually represents five separate dimensions. If the questions within the instrument are found to aggregate in three major groupings, this would call into question the five-dimension definition of the construct. Alternatively, for measurement instruments designed to tap only one dimension, factor analysis can be used to confirm that only one dimension is included in the construct.

Christine M. McDonough

See also Health Outcomes Assessment; Health Status Measurement, Floor and Ceiling Effects; Health Status Measurement, Generic Versus Condition-Specific Measures; Health Status Measurement, Minimal Clinically Significant Differences, and Anchor Versus Distribution Methods; Health Status Measurement, Reliability and Internal Consistency; Health Status Measurement, Responsiveness and Sensitivity to Change; Health Status Measurement Standards

Further Readings

Aday, L. A., & Cornelius, L. J. (2006). *Designing and conducting health surveys* (3rd ed.). San Francisco: Jossey-Bass.

American Psychological Association. (1985). *Standards for educational and psychological testing*. Washington, DC: Author.

Bowling, A. (1997). *Measuring health: A review of quality of life measurement scales* (2nd ed.). Philadelphia: Open University Press.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52,* 281–302.

Health Outcomes Methodology Symposium Proceedings. (2000). *Medical Care, 2000, 38*(9 Suppl. II).

McDowell, I. (2006). *Measuring health: A guide to rating scales and questionnaires* (3rd ed.). New York: Oxford University Press.

Nunnally, J. C. (1978). *Psychometric theory* (2nd ed.). New York: McGraw-Hill.

Streiner, D. L., & Norman, G. R. (2003). *Health measurement scales*. New York: Oxford University Press.

Health Status Measurement, Face and Content Validity

Health status measurement is a fundamental part of healthcare disciplines, for example, medicine, nursing, and clinical psychology. Health status refers to the perceptions of a person with respect to his or her health condition. Measuring health status is really a process in which a systematic and standardized attempt is made to observe an often complex clinical phenomenon. Health status measurement instruments are essential for this purpose. Instruments of health status measurement predominantly focus on overall well-being, functional status, symptom status, disease burden, health-related quality of life, psychological well-being, or satisfaction with care. Health status measurement has a significant impact on medical decision making. It provides important data and a platform for clinicians to monitor health conditions, predict clinical outcomes, assess the burden of a disease condition, and evaluate treatment effects. All health status measurements, from clinician-rated to patient-reported outcomes, should require convincing evidence that the clinical judgments or inferences drawn from scores on measurement instruments are valid and clinically useful.

Face and content validity are part of instrument development and validation, which provides theoretical support and some sorts of evidence about the validity of a health status measurement. *Content validity* refers to the degree to which the content and structural format of a health status measurement instrument are relevant to and representative of the intended construct (an abstract or a general idea, e.g., health-related quality of life) for particular characteristics of the client and purposes of the measurement. *Face validity* is a component of content validity that provides an additional attribute of the health status measurement instrument. It pertains to whether the content domains or items in a scale and their relation to the measurement purpose look valid to target respondents. Ensuring face validity is a minimum prerequisite for acceptance of a health status measurement instrument for target respondents.

Significance in Validity

The term *validity* refers to the degree to which a test or a measurement instrument measures what it purports to measure. *Test validity* encompasses reliability, validity, sensitivity, and responsiveness to change, which are interrelated and mutually inclusive to contribute to different aspects of evidence of the validity of a measurement instrument. Such lines of evidence include numerical analysis of internal structure of the instrument by correlating scores among items and with external criteria (other established instruments).

Face and content validity comprises a category of validity. The concept of face and content validity was first introduced into the literature of educational and psychological testing in the early 1940s. Face and content validity involves not only a qualitative process but also a numerical analysis to ensure that the measurement instrument as a whole has enough items and adequately covers the domain of content as well as having suitable structural format in the earliest stage of instrument development and validation. David Streiner and Geoffrey Norman, in 2003, pointed out the importance of face and content validity in measuring validity, in which the higher the content validity of an instrument, the broader are the inferences that can validly be drawn about the client under a variety of conditions and in different situations. Face and content validity is a fundamental requirement of all health status measurement instruments and is a prerequisite for establishing other types of validity. This initial stage of instrument development and validation is the most crucial, and no amount of psychometric analyses can transform an ill-conceived instrument into a good one.

Aspects of Face and Content Validity

There are several aspects of face and content validity: content domains, structural format, and target population. A content domain comprises the definition and dimension of the measurement of construct as well as the content items that are specific to the characteristics of the client and purposes of the measurement. The structural format includes the instructions to respondents, item wording, item format (question vs. statement) and item response form (ordinal vs. interval scale), temporal parameters of responses (timed vs. untimed), item weighting (equal vs. different weight in contributing to the total score), and scoring methods (summative score vs. transforming the raw score). *Target population* refers to the population for whom the instrument is to be applicable or to the patients who have a particular health condition or illness. All these aspects can affect the degree to which the observed data tap into the intended construct and the interpretation of the observed scores. Most important, they can influence the clinical judgments or inferences drawn from scores on the instrument and, thus, medical decisions.

Methods

Content validation occurs throughout the development of a health status measurement instrument. The ultimate goal of content validation is to maximize item coverage and relevancy so as to ensure that the health status measurement instrument comprises items that are relevant to and representative of the intended construct for the particular characteristics of the client and purposes of the measurement. It should be borne in mind that the content domains and items generated in content validation may change after other types of validity testing.

Item Coverage

Content coverage refers to the degree to which the content is adequate and representative for the intended construct and the purpose of the measurement. The extent of item coverage is not amenable to exploration by numerical analysis. It depends largely on the process of item generation. Subject-matter expert (e.g., clinician's) judgment, the patient-as-expert method, clinical observation, theoretical modeling, and literature review are the most commonly used approaches to item generation. Expert judgment is formed on the basis of a clinician's years of experience in the subject area. Clinicians who have extensive experience with the subject matter can explain the health status of particular salience to the intended construct from their perspective. The patient-as-expert method fulfills a basic requirement of patient-reported outcome instruments (e.g., condition-specific health-related quality-of-life instruments), in which the content should be generated from relevant patients. Patients can articulate what they feel and can explain the

areas of salience and concern associated with their health conditions. Clinical observation in a systematic manner helps suggest items. Theoretical modeling provides a conceptual foundation underlying the development of a measurement instrument and helps inform the hypothesized dimensions of the measurement construct so as to guide the development of the content domain. A review of published literature (subject-area research and previous measurement instruments) provides additional items to avoid possible omission of items that could be clinically significant. These methods are not mutually exclusive but depend on the nature of measurement instruments.

Exploratory in-depth qualitative interviews and focus group discussions with subject-matter experts and patients suffering from the illness and/or their families are the most efficient techniques to generate items. There are no hard rules governing the use of expert judgments and the patient-as-expert method, such as how many experts or patients to use, or how to handle differences among the experts or patients. The criterion often used in qualitative interviews is *sampling to redundancy*; that is, interviewing people until the point at which no new significant themes emerge. Normally, two or three focus groups with 6 to 12 informants and a facilitator in each group are needed. It should be borne in mind that the informants should represent the full diversity of subject-matter experts and patients with illness to minimize bias elicited from underrepresented or overrepresented samples.

Item writing and structural format planning are essential to content validation, which should include input from subject-matter experts, linguistic experts, and psychometricians. The characteristics of the target population, including age and reading comprehension level, are the major considerations in self-reported instruments. Focus group discussion with members of the target population can be used to assess the face validity of the instrument. In focus groups, participants can comment on the clarity, understandability, and appropriateness of all instructions and content items and check on the most appropriate wording.

Item Relevancy

Content relevance refers to the congruence between the measurement content and the purpose of the measurement. All the items that are included should be relevant to the construct being measured, and any irrelevant items should be excluded. Items that are not related to the construct could introduce errors in the measurement. A health status measurement instrument aiming to assess sore mouth, for example, should include items relating to all relevant issues associated with sore mouth, such as mouth pain and difficulty in eating. Irrelevant items, such as headache, should be excluded. Otherwise the instrument would discriminate among the patients on some dimension (headache) other than the one purportedly tapped by the instrument (sore mouth), and this has implications for medical decision making.

Item relevance is commonly approached by using several reviewers to critically evaluate whether individual items and the entire instrument are relevant to the construct being measured. Irrelevant, redundant, and ambiguous items should be excluded. Reviewers also comment on other aspects of content validity (e.g., item formats and response forms, item weighting, and scoring). Reviewers should be chosen to include subject-matter experts, psychometricians, and the target population (e.g., patients). A minimum of five reviewers are needed to provide a sufficient level of control of error variance resulting from chance agreement. The underlying measurement construct and the general goal for measurement should be provided to the reviewers. This information allows the reviewers to have the necessary theoretical background to provide a comprehensive review of the construct and to determine whether the proposed format and wording yield the appropriate level of validity. The interrater agreement (IR) and content validity index (CVI) are commonly used to check the relevancy of items by the degree of agreement among the reviewers in evaluations of the measurement content. IR and CVI calculations should apply to both individual items and the entire instrument. CVI is derived from the rating of the content relevance of the items on an instrument using a 4-point ordinal rating scale: 1, *not relevant;* 2, *somewhat relevant;* 3, *relevant;* and 4, *very relevant*. The actual CVI is the proportion of items rated 3 or 4 by the reviewers. IR is computed by adding the number of agreements among the reviewers (all items rated 1 or 2 by all reviewers, plus all items rated 3 or 4 by all reviewers) and dividing by the total number of items.

Alternatively, the entire instrument can be administered to a group from the target population as a pretest. All the participants are then interviewed to determine whether they find the items to be relevant and important.

Karis K. F. Cheng

See also Health Status Measurement, Construct Validity; Health Status Measurement, Reliability and Internal Consistency; Health Status Measurement Standards

Further Readings

Burns, W. C. (1996). *Content validity, face validity, and quantitative face validity*. Retrieved January 20, 2009, from http://www/burns.com/wcbcontval.htm

Davis, L. L. (1992). Instrument review: Getting the most from a panel of experts. *Applied Nursing Research, 5,* 194–197.

Fayers, P. M., & Machin, D. (2001). *Quality of life: Assessment, analysis and interpretation*. New York: Wiley.

Haynes, S. N., Richard, D. C. S., & Kubany, E. S. (1995). Content validity in psychological assessment: A functional approach to concepts and methods. *Psychological Assessment, 7,* 238–247.

Lynn, M. R. (1986). Determination and quantification of content validity. *Nursing Research, 35*(6), 382–385.

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (4th ed.). New York: McGraw-Hill.

Sireci, S. G. (1988). The construct of content validity. *Social Indicators Research, 45,* 83–117.

Streiner, D. L., & Norman, G. R. (2003). *Health measurement scales: A practical guide to their development and use* (2nd ed.). Oxford, UK: Oxford University Press.

Vogt, D. S., King, D. W., & King, L. A. (2004). Focus groups in psychological assessment: Enhancing content validity by consulting members of the target population. *Psychological Assessment, 16,* 2231–2243.

Health Status Measurement, Floor and Ceiling Effects

Floor and ceiling effects refer to specific limitations encountered when measuring health status scores. Floor effects occur when data cannot take on a value lower than some particular number; ceiling effects occur when data cannot take on a value higher than an upper limit. Health status instruments or surveys that are used to assess domains or attributes of health status use a rating scale. This is commonly a Likert scale with rating scales between 1 and 10, for example. There are limitations to the use of such instruments when measuring health status for either evaluative or discriminative purposes. An awareness of these limitations is important because of the problems that can occur in the interpretation of the results obtained when measuring health status, regardless of the domain being measured or the instrument that is being used. In interventional clinical trials, the degree to which health status changes is an important outcome; and the results of a study can be affected by floor and ceiling effects. In cost-effectiveness evaluations, the denominator of the ratio reported could be higher or lower than anticipated if there is a floor or ceiling effect. Therefore, recognizing ceiling and floor effects, and doing the best to minimize or eliminate these limitations, is important for studies that affect medical decision making. This entry further defines floor and ceiling effects, discusses how these effects are typically detected and potentially accounted for, provides examples of ways researchers try to minimize these scaling effects, and discusses the implications of floor and ceiling effects on randomized clinical trials and policy decisions. Finally, newer psychometric methods that are emerging to minimize such effects are briefly discussed.

Definitions

A ceiling effect occurs when the majority of scores are at or near the maximum possible score for the variable that the health status survey instrument is measuring. The survey instrument cannot measure scores above its ceiling. If a high percentage of people score at the top of a scale, it is impossible to detect an improvement in health for that group. Measures of activities of daily living (ADL) often have ceiling or floor effects in certain populations. For example, some individuals with specific chronic diseases such as stroke may exhibit high ceiling effects on more general surveys of health status, thus limiting the ability to distinguish certain aspects of health status between individuals scoring

at the ceiling. Ceiling effects are particularly important limitations when researchers are looking for the impact of treatment interventions on changes in health status. A floor effect occurs when the majority of scores are at or near the minimum possible score for the variable that the health status survey instrument is measuring. If a high percentage of people score at the bottom of a scale, it is impossible to detect a decline in health for that group. In clinical trials, for example, floor effects occur when outcomes are poor in the treatment and control conditions.

There are numerous health status measurement survey instruments that are generally divided into generic and disease-specific measures. Common examples of generic health status questionnaires for individuals with chronic diseases include several iterations of the Medical Outcomes Study survey and the EuroQol (EQ-5D). Some examples of disease-specific health status questionnaires include the Chronic Heart Failure Questionnaire (CHQ) and the Peripheral Artery Questionnaire.

Effects can vary by instrument. For example, comparative examinations of the SF-6D and EQ-5D across seven patient/population groups (chronic obstructive airways disease, osteoarthritis, irritable bowel syndrome, lower back pain, leg ulcers, postmenopausal women, and the elderly) revealed evidence of floor effects in the SF-6D and ceiling effects in the EQ-5D. This suggested that the SF-6D tended to discriminate better at higher levels of function and had heavy floor effects, while the EQ-5D performed in the opposite manner—it did well at lower levels of function, but had high ceiling effects. The choice of an instrument depends on what one wishes to measure. If the population has considerable morbidity, the EQ-5D may be a better choice. For a generally healthy population, the SF-6D may be the better choice. Another illustrative example is that of the problems encountered in the Veterans Health Study that used the MOS-VA. The VA had to extend the MOS SF-12/36 to include some instrumental activities of daily living (IADL)/ADL type times because of floor effects that occurred with the standard MOS. The pervasiveness of ceiling and floor effects has prompted the quest for a more appropriate approach to health status questions to accurately assess the health status of individuals and populations.

Detecting Ceiling and Floor Effects

Traditionally, classical test theory (CTT), a type of psychometric theory that analyzes measurement responses to questionnaires, has been used to evaluate the psychometric properties of health. Determining if a floor or ceiling effect exists requires an examination of the acceptability of the distribution of scores for the health domains obtained from the health status instrument. Measures of central tendency of the data, including mean and median, as well as the range, standard deviation, and skewness are used for such purposes. A score would generally be considered acceptable if the values are distributed in a normal or bell-shaped curve, with the mean near the midpoint of the scale. Floor effects can be determined by examining the proportion of subjects with the lowest possible scores. Similarly, ceiling effects are calculated by determining the proportion of subjects who achieved the highest possible score. Criteria for defining floor and ceiling effects are controversial. Some recommend a skewness statistic between −1 and +1 as acceptable for eliminating the possibility of a floor or ceiling effect.

Dealing with scales where the distribution is skewed, that is, where there is a ceiling or floor effect, is most problematic when comparing groups, as many statistical procedures rely on scores being evenly distributed. Making comparisons between groups in a clinical trial, or testing the effect of an intervention, may require additional advanced statistical techniques to adjust or account for the skewness of the data.

Minimizing Ceiling and Floor Effects

There are considerable conceptual and methodological challenges that confront users of health status instruments. Some individuals believe that ceiling and floor effects can be managed with statistical techniques. Others believe that these effects can be avoided or minimized by using disease-specific health surveys. Other options are to begin with a generic survey and use the disease-specific survey only if a ceiling or floor effect is observed. Still others believe that valuable information about the quality of life for individuals can be obtained by using both types of surveys.

Implications in Clinical Trials

Increasingly, researchers believe that measures of health status should be included in clinical trials. Historically, clinical research has focused on laboratory outcomes such as blood pressure, cholesterol, HgbA1C, morbidity, and/or mortality. These have been the outcomes measures of greatest interest to researchers, clinicians, and patients.

It is now necessary to employ health status measures to obtain a comprehensive assessment of practical health outcomes for individuals enrolled in clinical trials. The selection of the survey depends on the objectives of the evaluation, the targeted disease and population, and psychometric characteristics. Many of the disease-specific health status measures are sensitive to the occurrence of clinical symptoms or relatively small differences between treatment interventions—in particular, those studies examining the effect of medications—thus reducing the possibility of ceiling effects. Detecting worsening health among people who are already ill presents a different challenge. Low baseline scores make it difficult to detect health status decline, arguing again for disease-specific measures to avoid floor effects. In general, if one encounters a floor or ceiling effect in a study using a general health status measure, then a disease-specific measure, which is purposefully designed to be responsive to disease progression and/or treatment responsiveness issues, should be administered as well.

Disease-specific measures are believed to be more sensitive to treatment effects; however, a number of generic health status measurement scales have demonstrated the ability to discriminate between groups and clinical responsiveness. Thus, while many argue for the exclusive use of disease- and domain-specific measures for different disease conditions, the general recommended approach in randomized clinical trials of new medical therapies is to incorporate both generic and specific instruments to comprehensively assess health status. It may be worthwhile to pilot measures in the type of population to be studied, thus establishing that the measures adequately represent the health of the population before using them to establish the effectiveness of interventions. There is general agreement on the need for more comprehensive measures with multiple domains and multiple items to detect subtle changes in both healthy and severely ill populations.

Policy Implications

Because health status measures can provide comparisons across conditions and populations, they are of interest to policy and decision makers. Such information has the potential to improve the quality of care and establish reasonable reimbursement practices.

These measures are also of interest to clinicians because they help to determine the impact of therapeutic interventions and quality of life in their particular patient populations. Health status measures may provide clinicians with information not otherwise obtained from patient histories. Surveys can be self-administered, scanned, and used to provide rapid feedback of health status data—a phenomenon already occurring in many parts of the United States.

However, these measures must also be interpretable by policy and decision makers, and challenges exist in ensuring that decision and policy makers and clinicians understand these more complex scaling issues with health status measures. Without a full understanding of the concepts and methods, results could impart an incorrect message to a clinician or policy maker and ultimately discourage continued use of the measure. Strategies to make scores interpretable have been described. For an evaluative instrument, one might classify patients into those who experienced an important improvement, such as change in mobility, and those who did not and examine the changes in scores in the two groups. Data suggest that small, medium, and large effects correspond to changes of approximately 0.5, 1.0, and greater than 1.0 per question for instruments that present response options on 7-point scales.

Item Response Theory

CTT remains the dominant theory of measuring health status by researchers and clinicians. However, in the field of psychometrics, CTT is becoming outdated and replaced by more sophisticated, complex models. Item response theory (IRT) potentially provides information that enables a researcher to improve the reliability of an assessment beyond that obtained with CTT. Although both theories have the same aims, IRT is considered to be stronger in its ability to reliably assess health status. IRT allows

scaling of the level of difficulty of any item in a domain (e.g., physical function). Thus, theoretically, an item bank could have hundreds or thousands of survey questions covering a huge range of capabilities in a domain. Computerized adaptive testing (CAT) is a way of iteratively homing in on a person's level of ability in a particular domain by selectively asking questions across the broad domain and narrowing the estimate of ability by selecting new items to ask the person based on his or her responses to previous items. For example, if a person has told you that he or she can run a mile, there is no need to ask if he or she can walk one block. CAT could potentially eliminate floor and ceiling effects by having an item bank so broad that all meaningful levels of ability are covered.

However, the newer models are complex and spreading slowly in mainstream research. It is reasonable to assume that IRT will gradually overtake CTT, but CTT will likely remain the theory of choice for many researchers, clinicians, and decision makers until certain complexity issues associated with IRT can be resolved.

Barbara A. Bartman

See also Decision Making in Advanced Disease; Decisions Faced by Nongovernment Payers of Healthcare: Managed Care; EuroQoL (EQ-5D); Government Perspective, Informed Policy Choice; Health Outcomes Assessment; Health Status Measurement, Generic Versus Condition-Specific Measures; Health Status Measurement, Minimal Clinically Significant Differences, and Anchor Versus Distribution Methods; Health Status Measurement Standards; Measures of Central Tendency; Outcomes Research; Randomized Clinical Trials; Scaling; SF-6D; SF-36 and SF12 Health Surveys

Further Readings

Brazier, J., Roberts, J., Tsuchiya, A., & Busschbach, J. (2004). A comparison of the EQ-5D and SF-6D across seven patient groups. *Health Economics, 13*(9), 873–884.

Guyatt, G. H., Feeny, D. H., & Patrick, D. L. (1993). Measuring health-related quality of life, *Annals of Internal Medicine*, *118*, 622–629.

Kazis, L. E., Miller, D. R., Clark, J. A., Skinner, K. M., Lee, A., Ren, X. S., et al. (2004). Improving the response choices on the veterans SF-36 health survey role functioning scales: Results from the Veterans Health Study. *Journal of Ambulatory Care Management, 27*(3), 263–280.

Kind, P. (2005). *EQ-5D concepts and methods: A developmental history*. New York: Springer.

McDowell, I. (2006). *Measuring health: A guide to rating scales and questionnaires*. Oxford, UK: Oxford University Press.

Mesbath, M. (2002). *Statistical methods for quality of life studies: Design, measurements, and analysis*. New York: Springer.

Health Status Measurement, Generic Versus Condition-Specific Measures

Many measures of health status have been developed in the past decade for describing health outcomes and quantifying the changes. The term *health status measure* is often used interchangeably with other terms such as *health measure, health-related quality of life,* or *quality-of-life measure* even though the scope and definition of each term might not be the same. *Health status measurement* is defined as an instrument used to describe an individual's health state as defined by the descriptive system developed for each instrument.

Health status measurement can be classified into two main categories: generic measure and condition-specific measure. A generic measure is designed for use across a wide range of conditions, treatments, and populations. It is applicable to different subgroups in the general population or patient groups with various conditions or interventions. In contrast, a condition-specific measure is designed for measuring outcomes affected by a given condition only, for instance, lung cancer or arthritis. A condition-specific measure is tailor-made and is not supposed to be used for other conditions/diseases or the general population. A generic measure is designed to be applicable to any population; thus, it allows for meaningful comparisons between healthcare programs or interventions even if the involved patients or treatments may be different. In general, a generic measure has a descriptive system covering common domains of health so as to be relevant to everyone. Such core domain design, however, might

be inappropriate or insensitive for some specific conditions. On the other hand, a condition-specific measure is more sensitive to the degree of severity of condition and change over time, since the measure can focus on the most important domains affected by the condition. It can also include domains that are relevant to the condition but that are often missed by generic measures so that the relevant consequences of the condition can be captured. However, the condition-specific measure, which focuses on domains of interest or importance affected by the condition, does not allow comparison between different conditions.

For any measurement, there are two principle elements—the description and valuation. In a health status measure, the description is based on establishing a nominal descriptive system with which the defined health may be expressed in terms of key domains of interest. In other words, the descriptive system comprising chosen domains reflects the health definition adopted by the instrument developers. There exist discrepancies between measures in terms of their health domains of interest, and, as a consequence, the descriptive system varies among measures. For instance, some measures have taken a broader approach toward health, including the aspect of participation in society as one of the health domains even though it is arguable that social activities or role performance are not matters of health, per se. Others have chosen the "within skin" approach, focusing on impairment or inability of the individual only. A health status measure in which health itself is expressed as the domain of interest is also called a *multi-attribute health status measure*. Again, the definition of health varies from one measure to another and it should be borne in mind that there is no single measure with a descriptive system that captures all aspects of health.

Another key component for measurement is valuation. To perform valuation is to determine a set of weights associated with elements of a descriptive system. Thus, with this set of weights (commonly known as the scoring system for an instrument), scores can be calculated for domains or health states defined by the descriptive system. Various methods exist for eliciting weights, such as category scaling, visual analogue scale (VAS), time trade-off (TTO), standard gamble (SG), and paired comparison. Different eliciting methods generate different values. For instance, the VAS value of a given health state is generally lower than the TTO score of the same state. In most measures of health status, however, all items score equally, with equal weight for each response level in an item and with all items of equal importance.

Generic Measure

Generic measures can be further divided into two categories: one is the preference-based measure, also known as the index-based measure, and the other is the profile measure, also known as the non-preference-based measure. A preference-based measure offers a single summary numerical score for each health state defined by the instrument. This form of presenting a health outcome is particularly useful in economic evaluation, where a single index of health-related quality of life that summarizes health status utilities is needed. This is unlike a profile-based measure, which describes a health outcome by several different domains/dimensions in such a way that it is presented as a profile with several scores.

Preference-Based Measure

Due to the growth of economic evaluation, the popularity of the preference-based measure that provides a single summary score as a health-related quality of life for quality-adjusted life year (QALY) calculation has boomed. Its ease of use and its off-the-shelf service, providing a ready-made questionnaire and a set of weights, has led it to be widely adopted in cost-effectiveness studies. There are many index-based measures available, and the most commonly seen include the Quality of Well Being (QWB) scale, the Health Utility Index Mark 2 and 3 (HUI2/3), the EQ-5D, and the SF-6D. The following sections give a brief introduction to the EQ-5D and the HUI2/3. The interested reader can refer to additional sources, such as the books by McDowell and by Brazier and his colleagues, as listed in the Further Readings. It should be noted that each measure varies considerably in several aspects, such as the chosen domains in the descriptive system, the eliciting method, and the sample population for conducting the valuation. Therefore, the values obtained by each measure do not necessarily agree with one another.

EQ-5D

The EQ-5D is a generic preference-based measure of health status developed by the EuroQol Group. Established in 1987, this group of multidisciplinary researchers from Europe designed the EQ-5D as a simple and generic measure to be used alongside other generic or condition-specific measures. Nowadays, the EQ-5D is one of the most widely used generic measures and has more than 100 official translations available.

The EQ-5D has two main components—the EQ-5D descriptive system and the EQ visual analog scale (EQ VAS). The EQ-5D descriptive system comprises the following five dimensions: (1) mobility, (2) self-care, (3) usual activities, (4) pain/discomfort, and (5) anxiety/depression. Each dimension has three levels: (1) no problems, (2) some problems, and (3) severe problems. The respondent is asked to choose the most appropriate statement in each of the five dimensions. The EQ VAS is a vertical 20-centimeter-long thermometer with the lower and upper end points valued at 0 and 100 and labeled as "Worst imaginable health state" and "Best imaginable health state," respectively. The respondent rates his or her current health state on the EQ VAS.

A total of 243 (3^5) possible health states is defined by this five-dimensional, three-level descriptive system. Each health state can be assigned a single summary index score on what is known as the EQ-5D index by applying a scoring algorithm that essentially attaches values (also called weights) to each of the levels in each dimension. An algorithm normally is derived from the valuation of a set of EQ-5D health states in general population samples. The most widely used value set (EQ-5D index scores for each of 243 health states) is the TTO-based set of values obtained from the Measurement and Valuation of Health (MVH) study in the United Kingdom, for which a representative sample consisting of 3,395 subjects from the general population was interviewed. Many other country-specific value sets have been developed, including ones for the United States, the Netherlands, Japan, and so on. There are three ways of reporting EQ-5D results: (1) the EQ-5D health state, a profile reporting the problem level in each dimension; (2) the EQ-5D index score, representing social preference for the health state defined in the descriptive system; and (3) the EQ VAS score, a self-rated health score based on a VAS.

The EQ-5D is designed for self-completion by respondents and can also be interviewer administered in person or over the phone. More information can be obtained from the official EQ-5D Web site.

HUI2/3

The Health Utility Index Mark 2 and 3 (HUI2/3) are generic preference-based health status measures. HUI instruments are designed to provide utility scores for health outcomes evaluations. The first version of a HUI instrument (HUI Mark 1) was created in the 1970s to evaluate the outcomes of neonatal intensive care. HUI measures have continued to develop, and there are now two versions available: HUI2 and HUI3. The HUI2 was initially designed for measuring long-term outcomes of treatment for children with cancer and now can be used as a generic measure. The latest version, HUI3, was developed to address some issues of the HUI2 by extending and altering the attributes of its predecessor version.

Based on survey results from parent and child pairs and a literature review, the HUI2 consists of seven attributes (domains), such as (1) sensation, (2) mobility, (3) emotion, (4) cognition, (5) self-care, (6) pain, and (7) fertility. Each attribute has three to five levels of function, and therefore, the HUI2 defines up to 24,000 unique health states. Removing attributes of fertility, replacing self-care with dexterity, and adding distinct components of sensation to the HUI2, the attributes addressed in the HUI3 are vision, hearing, speech, ambulation, dexterity, emotion, cognition, and pain; each attribute has five to six levels. As stated above, the HUI2 is a generic measure for adults; it can also apply to child populations after removing the attribute of fertility. In total, the HUI3 defines 972,000 unique health states. The choice of attributes in the HUI has been based on the "within the skin" approach—focusing on the most fundamental and important attributes of physical and emotional health status, and excluding aspects of participation in society, such as social activity or role performance.

The HUI2 scoring algorithm was developed from the valuation of a random sample of 293 parents of schoolchildren in Hamilton, Ontario,

Canada. Both VAS and SG methods were adopted to elicit values from the sample. The scoring system was developed based on the multi-attribute utility theory, where utility function for each attribute was estimated separately and multiplicative function form was adopted in the final scoring formula. A power transformation was developed to convert values of health state measured by a VAS into utility as elicited by the SG method. The scoring formula of the HUI3 uses a similar approach based on responses from a representative sample of 504 adults from Hamilton, Ontario.

Further information on the HUI instruments can be found at the Health Utilities Inc. Web site.

Profile Measure

Several generic health profile measures have been developed and are available in the literature, such as the Sickness Impact Profile, Nottingham Health Profile, Short Form 36, Short Form 12, WHOQOL-BREF, and so on. Following is a brief introduction to the Sickness Impact Profile and Short Form 36. The interested reader can refer to additional sources, such as the book by McDowell or the one by Bowling, as listed in Further Readings.

Sickness Impact Profile

The Sickness Impact Profile (SIP) is a landmark instrument in the development of outcomes measurement. Its design had great influence on later measures such as the Nottingham Health Profile. The great care and thoroughness that went into its development are noteworthy. It was originally designed as a generic measure intended for use across different conditions and populations. The SIP can be either self- or interviewer-administered.

The SIP is a behaviorally based measure of dysfunction focused on assessing the way in which sickness changes daily activities and behavior. The development of its descriptive system took a bottom-up approach, collecting statements for change in behavior attributable to sickness both from patients and individuals and from the literature. A total of 312 unique statements were identified and sorted into 14 categories by the research team. The final version contains 136 items in 12 categories, including ambulation (12 items), mobility (10), body care and movement (23), communication (9), alertness behavior (10), emotional behavior (9), social interaction (20), sleep and rest (7), eating (9), work (9), home management (10), and recreation and pastime (8). Respondents choose/check the items in each category that describe and are related to their health.

The score can be presented by category, by physical and psychosocial dimensions, or by a single overall score within a range of 0 to 100. A lower score indicates better health. The overall score for the SIP is calculated as the sum of the weights of items checked across all categories divided by the sum of the weights for all items multiplied by 100. The same principle is used for calculating two-dimensional scores by limiting checked items to relevant categories only. Ambulation, mobility, and body care and movement form the physical dimension, while communication, alertness behavior, emotional behavior, and social interaction constitute the psychological dimension. The weights were developed using equal-appearing interval scaling procedures involving more than 100 judges.

More information can be obtained from the Medical Outcomes Trust Web site.

Short Form 36

One of the most widely used health profile measures is the Short Form 36 (SF-36) questionnaire. The SF-36 originated from the Medical Outcome Study (MOS), initially designed to evaluate health utilization of different health delivery systems in the United States. The 36 items comprising the SF-36 were derived from long-form measures of general health embodied in the MOS. Since its inception, the SF-36 has been continually developed by Ware and coworkers, being used for collecting data from several U.S. national surveys to develop social norms. There are several available versions of this 36-item questionnaire, and variation exists among their scoring systems. There are, for instance, the SF-36 by QualityMetric, the RAND 36-Item Health Survey 1.0 by RAND, and the RAND-36 HIS by the Psychological Corporation. There is also the SF-36v2, the latest version of the SF-36. This questionnaire, developed by QualityMetric, is demonstrated here as an example.

Like its predecessor, the SF-36, the SF-36v2 contains 36 items. Thirty-five items of the SF-36v2 cover eight health domains, such as (1) physical functioning (PF), (2) role-physical (RE), (3) bodily

pain (BP), (4) general health (GH), (5) vitality (VT), (6) social functioning (SF), (7) role-emotional (RE), and (8) mental health (MH) scales. These domains are constructed with 2 to 10 items each, and each item has response levels ranging from three to five categories. The only item that does not contribute to any domain is the one measuring perceived health change. There are two forms available with different recall periods: standard (past 4 weeks) and acute (past 1 week). The questionnaire can be either self-completed or interviewer-administered in person or over the phone. The differences between Versions 1 and 2 of the SF-36 include a layout improvement, the wording of the items, and an increase in the response categories of 7 items from either dichotomous or six-category to five-category.

Item responses for each domain are summed and transformed (using a scoring algorithm) into a scale of 0 to 100, with a higher score representing better health. Apart from the domains of body pain and general health, the scoring algorithm for the rest of the domains assumes equal weights for each response level in an item and between items. The score can be further standardized into a mean of 50 and a standard deviation of 10 based on the U.S. population norm. Thus, a score above or below 50 is interpreted variously as above or below average. The 8-domain score can be further summarized into a physical and a mental component summary (PCS and MCS) using the algorithm developed from the 1990 U.S. general population survey with factor analysis and the orthogonal rotation method.

Currently, there are several different lengths and versions of Short Form questionnaires available, including the SF-12 and SF-8. More information can be obtained from the QualityMetric Web site.

Condition-Specific Measure

As described earlier, the emphasis of condition-specific measures is on aspects of health affected by a condition. There might be some overlapping between generic and condition-specific measures, but the latter have domains not included in generic measures or domains with more detailed scopes. For instance, condition-specific measures might include domains measuring particular treatment effects or symptoms or focus greatly on some domains such as mobility or dexterity, depending on what is of interest in the condition. Recently, there has been growing attention placed on developing preference-based, condition-specific measures. The rationale is that the generic measure might not be appropriate for a given condition, and most condition-specific measures do not provide a summary score weighted by social preference for use in economic evaluation. Here, this entry introduces the Functional Assessment of Cancer Therapy (FACT) and briefly discusses its development into a preference-based measure as an example.

Functional Assessment of Cancer Therapy

The FACT is a cancer-specific measure designed for use in the evaluation of intervention in various types of cancers. The FACT consists of a core set of items applicable to all types of cancer and cancer-specific supplements. The instrument has evolved, and its applications have been expanded to different chronic illnesses and conditions. Since 1997, it has been renamed the Functional Assessment in Chronic Illness Therapy (FACIT). The FACT-L, for lung cancer, is explained here.

The core set of 27 items applicable to all types of cancer is known as the FACT-General and comprises four domains: physical well-being (7 items), social/family well-being (7), emotional well-being (6), and functional well-being (7). These domains were identified using factor analysis. Each item in the FACT-General has a five-level response. The 10 items specific to lung cancer are labeled as additional concerns, assessing coughing, breathing, smoking, and so on.

The Trial Outcome Index (TOI) can be computed for any FACT measure. It is the sum of physical well-being and functional well-being, plus additional concerns subscales. In the FACT-L, a total of 21 items is used to calculate the TOI score. Like most condition-specific measures, the FACT assumes that there is equal weight for each level in an item and equal importance among items. Such a scoring system might be sensitive enough for clinical purposes. However, it does not have the necessary properties required by economic evaluation.

The development of the preference-based FACT-L is aimed at addressing the above issue. A study conducted by Kind and Macran developed

a set of social preference weights for the FACT-L. The approach adopted consisted in first revising items in the FACT-L through both quantitative and qualitative methods to be amenable for valuation. The health states defined by the revised items were then valued by a sample of the United Kingdom's general population using a VAS through a postal survey. Econometric methods were used to develop weights for 10 items of the FACT-L based on the collected data. Thus, the derived utility weights of the FACT-L can be used in cost-utility analysis legitimately.

More information about the FACT can be obtained from the FACIT Web site.

Choosing a Measure

Generic and condition-specific measures can be seen as complementary measures to each other. One provides information for comparison across different populations, and the other offers the most relevant information on a given condition. However, when making the choice of measures—whether to use generic with condition-specific measures or choosing between a profile-based or a preference-based measure—care must be taken with regard to the purpose of the measurement as well as the burden that this would represent for respondents. It should be borne in mind that there is no single measure with a descriptive system that captures all aspects of health, and the exclusion does count—elements missing from the descriptive system have an arbitrary zero weight.

Ling-Hsiang Chuang

See also EuroQoL (EQ-5D); Health Status Measurement Standards; Health Utilities Index Mark 2 and 3 (HUI2, HUI3); SF-36 and SF12 Health Surveys; Sickness Impact Profile

Further Readings

Bergner, M., Bobbitt, R., Carter, W., & Gilson, B. S. (1981). The Sickness Impact Profile: Development and final revision of a health status measure. *Medical Care, 19,* 787–805.

Bergner, M., Bobbitt, R., & Kressel, S. (1976). The Sickness Impact Profile: Conceptual formulation and methodology for the development of a health status measure. *Journal of Health Service, 6,* 393–415.

Bowling, A. (2005). *Measuring health: A review of quality of life measurement scales* (3rd ed.). Maidenhead, UK: Open University Press.

Brazier, J., Ratcliffe, J., Salomon, J. A., & Tsuchiya, A. (2007). *Measuring and valuing health benefits for economics evaluation.* Oxford, UK: Oxford University Press.

Brooks, R. (1996). EuroQol: The current state of play. *Health Policy, 37,* 53–72.

Cella, D. F., Bonomi, A. E., Lloyd, S. R., Tulsky, D. S., Kaplan, E., & Bonomi, P. (1995). Reliability and validity of the Functional Assessment of Cancer Therapy–Lung (FACT-L) quality of life instrument. *Lung Cancer, 12,* 199–220.

Cella, D. F., Tulsky, D. S., Gray, G., Sarafian, B., Linn, E., Bonomi, A., et al. (1993). The Functional Assessment of Cancer Therapy scale: Development and validation of the general measure. *Journal of Clinical Oncology, 11,* 571–579.

EQ-5D: http://www.euroqol.org

The EuroQol Group. (1990). EuroQol: A new facility for the measurement of health-related quality of life. *Health Policy, 16,* 199–208.

Feeny, D. H., Furlong, W. J., Torrance, G. W., Goldsmith, C. H., Zenglong, Z., & Depauw, S. (2002). Multiattribute and single-attribute utility function: The Health Utility Index Mark 3 system. *Medical Care, 40,* 113–128.

Functional Assessment of Chronic Illness Therapy (FACIT): http://www.facit.org

Health Utilities Inc.: http://www.healthutilities.com

Kind, P., & Macran, S. (2005). Eliciting social preference weights for Functional Assessment of Cancer Therapy: Lung health states. *PharmacoEconomics, 22,* 1143–1153.

McDowell, I. (2006). *Measuring health: A guide to rating scales and questionnaires* (3rd ed.). New York: Oxford University Press.

Medical Outcomes Trust—Instruments: http://www.outcomes-trust.org/instruments.htm

QualityMetric: http://www.qualitymetric.com

Torrance, G. W., Feeny, D. H., Furlong, W. J., Barr, R. D., Zhang, Y., & Wang, Q. (1996). A multi-attribute utility function for a comprehensive health status classification system: Health Utilities Mark 2. *Medical Care, 34,* 702–722.

Ware, J., Kosinski, M., & Dewey, J. E. (2001). *How to score Version 2 of SF-36 Health Survey.* Lincoln, RI: QualityMetric.

Ware, J., Kosinski, M., & Keller, S. D. (1994). *SF-36 physical and mental health survey scale: A user's manual.* Boston: Health Institute, New England Medical Center.

Health Status Measurement, Minimal Clinically Significant Differences, and Anchor Versus Distribution Methods

When measuring quality of life, patient preferences, health status, or other types of patient reported outcomes (PROs), the term *minimal clinically significant difference* (MCSD) indicates the smallest amount of meaningful change or difference that can be assessed by a PRO measure. The term *meaningful change,* in this context, refers to the smallest difference that is perceived by patients (or other stakeholders) as beneficial or harmful and that would lead to a change in treatment.

From this perspective, the MCSD is a numerical value, and PRO score differences that exceed this value are considered indicative of important or meaningful change. MCSDs vary across different PRO measures (i.e., a difference of 10 points may be indicative of meaningful change for one measure but not another) and different populations (i.e., the same measure may have different MCSDs depending on the patient group being assessed). In practice, because of the difficulties inherent in establishing MCSDs, exact values are rarely identified. More frequently, investigators specify a range of values within which the MCSD is likely to fall.

Investigators frequently emphasize that the MCSD does not necessarily correspond to the smallest detectable difference. In other words, not all improvements or declines that are noticeable are necessarily noteworthy. Despite consensus on this point, investigators often disagree as to which methods allow determination of whether a difference is minimally important versus minimally detectable. This disagreement, in turn, may be linked to the noticeable variation in terminology that characterizes the MCSD literature. For example, although the current report employs the term *minimally significant clinical difference*, this concept is frequently referred to as the *minimal important difference, the clinically important difference, the minimal difference, the important difference,* and other similar combinations of words and phrases.

Despite occasional disagreement over terminology, investigators have made significant progress toward developing MCSD methods, primarily because without some means to assess meaningful change, PRO data cannot be used effectively. For example, suppose in the context of a clinical trial that a group receiving a new drug scores a statistically significant 8 points higher on a PRO measure of pain relief as compared with a placebo control group. Because statistical significance does not necessarily imply clinical significance, the investigators will be unable to conclude that the new drug provided a nontrivial benefit. In other words, without insight into the MCSD, there is no way to determine whether the drug's ability to reduce pain is large enough to make a meaningful difference in patients' lives. Similar issues arise when PRO data are employed in clinical, administrative, policy-making, or regulatory settings.

Given the complexity involved in identifying MCSDs, the current lack of a gold standard method for doing so is not surprising, although the health sciences appear to be converging toward a set of best practices. Currently, two different approaches are usually employed when identifying MCSDs: anchor-based and distribution-based methods. Anchor-based methods rely on some external criterion of known significance against which changes in PRO scores can be calibrated. Distribution-based methods rely primarily on the statistical properties of PRO sample values or the reliability of the PRO measure itself. The anchor- and distribution-based approaches are described individually in the following sections, but many authors recommend that both should be employed when identifying MCSDs, as each method approaches the task from a conceptually distinct perspective.

Anchor-Based Approaches

The anchor-based approach relies on identifying an external criterion (i.e., an anchor) that is relatively interpretable, and then examining how differences in PRO scores map onto that anchor. Anchors can take many forms, including patients' self-reports of change, clinical outcomes or conditions, or other events.

Jaeschke and his colleagues were one of the first groups to demonstrate the use of self-reported change using transition assessment items. Transition assessments are typically used in longitudinal investigations, such as clinical trials, in which

patients periodically complete a PRO measure as well as a self-report transition item assessing whether the patient has experienced no change, small but important change, moderate change, or large change on the PRO of interest since the last assessment. Investigators essentially derive an estimate of the MCSD by computing the average difference between consecutive PRO scores for the group reporting small but important change. This same approach also permits the identification of meaningful changes that are moderate to large in magnitude by computing the average difference score of the relevant groups. When transition assessment items are used, MCSDs for positive and negative change are sometimes derived separately, as some prior work has demonstrated asymmetries in the MCSD depending on whether a patient is improving or declining.

Although the anchor-based method is sometimes equated with the use of transition assessments, clinically based anchors are also frequently employed by MCSD researchers. For example, as part of a study intended to determine the MCSD of the Impact of Weight on Quality of Life–Lite (IWQL-L) instrument, the investigators categorized patients from several longitudinal studies into groups according to how much weight they had lost. To derive the MCSD, the investigators calculated the average difference score from the IWQL-L for the group that had lost from 5% to 9.9% of their original weight, the smallest amount deemed meaningful by the Food and Drug Administration.

Clinical anchors vary across studies, depending on the patient population. Investigators frequently advocate using multiple anchors, which are usually chosen because they are related conceptually and statistically to the PRO of interest. For example, when establishing MCSDs for some of the Functional Assessment of Cancer Therapy (FACT) scales, investigators used hemoglobin level, performance status, and response to treatment as anchors. All three predict and are clinically relevant to the cancer outcomes assessed by the FACT.

Anchors not only can be based on different criteria (e.g., patients' perceptions, clinical factors) but may also be derived from studies with different methodological designs. Although within-subject anchors are frequently used (e.g., transition assessments), anchors derived from between-subjects or cross-sectional comparisons are also possible. For example, the difference score between two groups that differ in clinically important ways (e.g., healthy individuals vs. hypertensives) has sometimes been used as an indicator of the MCSD for a given PRO measure.

Some authors have suggested that anchor-based methods are superior to distribution-based methods because only the former provide direct information about the importance or meaningfulness of a change. However, anchor-based methods have also been critiqued on several grounds. Because anchors are often themselves arbitrary, there is usually no way to verify empirically that the groups formed by anchors truly differ in important ways. Some authors have suggested that transition assessment anchors may be particularly problematic. Although self-reported change is the only way to directly incorporate the patient's perspective, transition assessments usually consist of a single item and are retrospective in nature, characteristics that could undermine their psychometric validity. Baseline status (e.g., poor health vs. good health) also appears to affect the size of MCSDs, a tendency that is likely to be more pronounced when anchor-based methods are used. For these reasons, and because different types of anchors tend to produce different MCSDs, most investigators recommend using multiple anchors. Typically, various types of anchors are explored over several studies to arrive at an MCSD or, more commonly, a range of values in which the MCSD is likely to fall.

Distribution-Based Approaches

As stated earlier, distribution-based approaches rely primarily on the statistical properties of sample data or the PRO measure itself. The chief value of distribution-based approaches is that, unlike anchor-based methods, they allow the identification of differences or changes that are essentially too large to have occurred by chance or from measurement error. Several of the more popular distribution-based methods are summarized in the following sections.

One of the most common distribution-based approaches relies on the effect size associated with a difference or change. Effect size can be computed by dividing the difference between two sample means (or the average difference in the case of a repeated measures design) by the sample standard

deviation or the pooled standard deviation. The resulting proportion, which essentially redefines the difference in standard deviation units, can be interpreted according to the well-known guidelines proposed for the behavioral sciences by Cohen, who suggested that effect sizes of .20 to .49, .50 to .79, and .80 and above should be considered small, moderate, and large, respectively. Thus, from this perspective, any difference or change associated with an effect size of .20 or greater would be considered clinically meaningful.

The reliable change index provides an alternative distribution-based method to identify MCSDs at the individual level and is computed by dividing the difference or change between two PRO scores by the standard error of the difference between the scores. The index, which depends on the standard error of measurement, contrasts an observed change with the change that would be expected from chance or measurement error. Some authors suggest that reliable change has occurred as long as the index value exceeds 1.96. This would indicate that the likelihood of obtaining the observed difference is only about 5% or less if there has been no actual change, suggesting that the change or difference is "real" and not the result of chance variation.

The standard error of measurement (SEM) can also be used in its own right to help derive MCSDs. The SEM is the standard deviation of an individual's scores on a specific measure. Because all measures contain some error, an individual's score would vary to some extent if the same measure were to be repeatedly administered to that individual. The SEM indicates how much variation would occur, with more precise or reliable measures having a lower SEM. Thus, the greater a difference or change relative to the SEM, the more likely that difference or change is likely to be "real" and not the result of chance or error. How much greater than the SEM a meaningful difference should be is somewhat controversial, with suggestions ranging from 1 SEM to 1.96 SEMs to 2.77 SEMs. The SEM can be calculated by multiplying the sample standard deviation by the square root of one minus the reliability of the PRO measure. Because of the inverse relationship between sample variance and measure reliability, a measure's SEM should remain fairly stable across different samples.

Distribution-based methods are relatively easy to implement because they do not require anchor data and, as previously noted, provide the additional advantage of identifying MCSDs that exceed variation due to chance or measurement error. However, as several investigators have noted, this property by itself does not guarantee that the change or difference is necessarily large enough to be important from the perspective of the patient or other stakeholder. Conversely, some distribution methods may result in MCSD estimates that are too large. For example, some investigators have noted that the 1.96 criterion commonly used in conjunction with the reliable change index is fairly strict, resulting in conservative (i.e., large) MCSD estimates relative to other methods. In general, it is often not clear which standards to apply when establishing MCSDs using distribution-based methods. For example, the number of SEMs that a difference or change has to exceed to be considered meaningful tends to vary across disciplines.

Distribution-based approaches, especially those involving effect size and the SEM, have received increasing attention over the past decade, largely due to the work of Wyrwich and her colleagues. Using data from a variety of patient samples, Wyrwich and others have found that 1 SEM is frequently, though not always, equivalent to an effect size of approximately .50 when PRO measures with appropriate levels of reliability are used. These findings suggest that differences or changes that exceed 1 SEM may generally be large enough to be meaningful.

Additionally, MCSDs identified using anchor-based approaches are often associated with an effect size approximating .50, although this phenomenon appears most robust in patients with chronic health conditions. Consequently, some authors have suggested that in the absence of other information, a difference or change associated with an effect size of .50 is likely to be clinically meaningful. However, as these and other authors have cautioned, some prior work has identified anchor-based MCSDs that are associated with effect sizes both smaller and larger than .50, thus highlighting the critical role that anchor-based methods can play.

Best Practices

Developing methods to identify MCSDs has proven a challenging and complex task. However, consensus is emerging over a set of best practices. Specifically,

most investigators recommend triangulating on MCSDs using a combination of both distribution- and anchor-based approaches across multiple samples. Whereas distribution-based methods help ensure that MCSDs are large enough to exceed chance variation, anchor-based methods help ensure that MCSDs are properly sized to reflect truly meaningful and important change.

R. Brian Giesler

See also Health Status Measurement, Assessing Meaningful Change; Health Status Measurement, Responsiveness and Sensitivity to Change

Further Readings

Crosby, R. D., Kolotkin, R. L., & Williams, G. R. (2003). Defining clinically meaningful change in health related quality of life. *Journal of Clinical Epidemiology, 56,* 395–407.

de Vet, H. C., Terwee, C. B., Ostelo, R. W., Beckerman, H., Knol, D. L., & Bouter, L. M. (2006). Minimal changes in health status questionnaires: Distinction between minimally detectable change and minimally important change. *Health and Quality of Life Outcomes, 4,* 54–59.

Guyatt, G. H., Osoba, D., Wu, A. W., Wyrwich, K. W., & Norman, G. R. (2002). Methods to explain the clinical significance of health status measures. *Mayo Clinic Proceedings, 77,* 371–383.

Jacobson, N. S., & Truax, P. (1991). Clinical significance: A statistical approach to defining meaningful change in psychotherapy research. *Journal of Consulting and Clinical Psychology, 59,* 12–19.

Norman, G. R., Sloan, J. A., & Wyrwich, K. W. (2003). Interpretation of changes in health-related quality of life: The remarkable universality of half a standard deviation. *Medical Care, 41,* 582–592.

Revicki, D., Hays, R. D., Cella, D., & Sloan, J. (2008). Recommended methods for determining responsiveness and minimally important differences for patient-reported outcomes. *Journal of Clinical Epidemiology, 61,* 102–109.

Sloan, J. A., Cella, D., & Hays, R. D. (2005). Clinical significance of patient-reported questionnaire data: Another step toward consensus. *Journal of Clinical Epidemiology, 58,* 1217–1219.

Wyrwich, K. W., Bullinger, M., Aaronson, N., Hays, R. D., Patrick, D. L., Symonds, T., et al. (2005). Estimating clinically significant differences in quality of life outcomes. *Quality of Life Research, 14,* 285–295.

Health Status Measurement, Reliability and Internal Consistency

Reliability refers to the concept of consistency of measurement. Medical decision making is based on observations and measurements taken from the patient. For decision making to be based on the best possible information, such measurements should be both reliable and valid. Reliability of measurement is a necessary condition for a measure to be valid but is not sufficient to ensure validity. A psychometric test, attitude scale, or observational measurement may be reliable in that it can be consistently assessed over time but not valid in that it does not accurately reflect the construct it is designed to measure. In terms of statistical theory, reliability is defined as the ratio of the variation of the true score and the variation of the observed score.

Assessments of the reliability of a measure can be broadly split into two groups: (1) methods of determining the reliability of a test by repeated administration (e.g., test-retest reliability and inter-rater reliability) and (2) methods that require a single administration of the test. The latter methods are often termed internal consistency measures.

Repeated Administration

The logic of repeated administration as a method for determining the reliability of a measure is simple. If a test or observation is reliable, then if it is measured twice in the same individual without any change in the individual occurring, the value of the measurement should be the same. Test-retest reliability involves the administration of the same test or observation to a group of individuals at two points separated by a period of time. There is no simple guide to the ideal time period for test-retest reliability—essentially, it should be short enough to ensure that the sample has not changed in the aspect being measured but long enough to prevent individuals recalling their previous answers and using their recall as the basis for their responses on the second occasion. Periods of 2 to 4 weeks are typical. Some researchers advocate asking respondents to indicate whether they feel that they have

changed in the construct under study between the two testing sessions, and then excluding those with a perceived change.

Interrater reliability refers to the consistency of measurement between two raters making independent observations of the same individuals. Interrater reliability is most commonly used in psychology where observational data are collected. Two individuals may observe either live behavior or a video recording of behavior and make ratings in terms of a standardized measurement. The independent ratings can then be observed. Simple measures of the percentage agreement and percentage disagreement can be calculated. The extent of agreement will depend to a large degree on how well the behaviors and the categories to be recorded are specified and the extent to which observers have been trained to use the scoring method. Extensive training and calibration in the measurement method prior to the research is an excellent way to ensure reliability of observational measurements.

In a large-scale study involving the observation of behavior, both interobserver and test-retest reliability measures may be taken. This is to avoid the potential for measurement error resulting from a change in the way raters make their observations over time. By reappraising data from earlier in the trial, it is possible to determine whether such measurement drift has occurred.

The statistical methods used to determine whether there is consistency across the two measurements for both test-retest reliability and interrater reliability depend on the nature of the data. For nominal data, Cohen's kappa statistic is calculated. This is the proportion of observations that agree across the two testing situations, corrected for chance levels of agreement. Values of Cohen's kappa range from 0 to 1.0, where 1.0 is perfect agreement. The following guide to the extent of agreement was produced by Douglas Altman:

0 to .6 poor agreement,

.6 to .8 satisfactory agreement,

.8 to 1.0 excellent agreement.

For ordinal data, the weighted kappa is used. This is an extension of Cohen's kappa but gives greater weight to disagreements far removed on the ordinal scale and smaller weight to disagreements falling on adjacent points. The values of weighted kappa are interpreted in the same manner as kappa.

Where the data tested for reliability are continuous, there are a number of statistical methods to determine the extent of consistency in the measurement. Some researchers have suggested the use of simple correlation statistics, but this is now considered inappropriate, since measurements could be perfectly correlated but differ in magnitude. To overcome this limitation, Bland and Altman suggested a simple method of plotting the data to explore the consistency of measurement. A Bland-Altman plot comprises a plot of the difference between two measurements on the same individual against the mean of the two measurements. Typically the overall mean and standard deviation of the two measurements are also placed on the graph. This allows the researcher to explore the range of magnitude of differences and, importantly, whether differences are larger at the extreme points of the measurement scale. This is important, since regression to the mean will affect extreme values most markedly. An alternative approach that is currently popular is to calculate the intraclass correlation (ICC) of the two sets of data. The ICC is a correlation coefficient where the intercept is forced to occur at the origin (0, 0) of the bivariate plot. The ICC is equivalent to a correlation coefficient and can be interpreted as such.

Single Administration

Single administration methods have the advantage that the consistency of measurement can be determined in a single assessment. They are therefore easier and more economic to undertake and especially useful where either the measurement changes the participant in some way, or there is a risk that the participant will change before a second observation can be arranged. These methods most commonly are used for questionnaires or observations with multiple related items measuring the same construct. The *split half* method works on the assumption that if all the items in a questionnaire are consistently measuring the same construct, then it should be possible to split the items into two halves that will correlate highly with each other. An obvious example would be to compare odd-numbered and even-numbered items. It is also

possible to calculate all the possible split-half combinations of a questionnaire (given by the number of items – 1) and then calculate the average correlation between all the possible combinations. Generally, longer tests will show higher internal consistency, but using a split-half method means that the two forms of the scale are effectively half as long as the original. It is possible to correct for this in the calculation of the internal consistency statistic by using the Spearman-Brown formula.

Cronbach's alpha is a widely used measure of the internal consistency of scales. It is the average of the correlations of each item with the total score of the scale (excluding that item). The logic of Cronbach's alpha is that if all the items in a scale are reliably measuring the same construct, then they should all correlate highly with each other. Coefficient alpha will range from 0 (no consistency) to 1.0 (perfect consistency). For psychometric tests, an alpha above .7 is recommended, whereas for clinical tests, a value of alpha greater than .9 should be sought, according to Bland and Altman. Cronbach's alpha can be used where items have multiple response categories or in the situation where items have binary responses. In the latter case (binary response categories), Cronbach's alpha is equivalent to the Kuder-Richardson Formula-20, which has been used as a measure of internal consistency. However, since the two formulae are equivalent for the case where response categories are binary but, in addition, Cronbach's alpha deals with the wider case, alpha is generally the preferred statistic. Certain statistical packages will calculate Cronbach's alpha for a scale along with statistics for individual items within the scale (such as how well the scale performs when individual items are deleted). These statistics are often used to derive scales that are internally consistent by selectively removing items; this procedure can be an effective way of developing internally consistent scales, but some caveats should be noted: First, Cronbach's alpha is not suitable for very low numbers of items (three or two items); second, the scale derived by removing items may be internally consistent, but such a correction does not necessarily reflect the construct that the scale was originally devised to measure—measures of internal consistency give information on reliability, not validity.

The reliability of observations can generally be improved by adopting a range of methods. As mentioned previously, adding items to make a scale longer will improve reliability, provided that the items are conceptually similar. Ensuring that the meaning of all items is clear to participants will ensure that their answers are consistent. Item analysis—that is, exploring the psychometric properties of each item in a scale—will help identify "rogue" items that are lowering the reliability estimates. Item analysis includes exploring the correlation of the item to all the other items; those with an average interitem correlation less than .3 should be excluded.

It should also be noted that the reliability statistics calculated are related to the scores of a measure rather than the measure itself, and therefore will vary across different samples. Reliability estimates from one sample might differ from those of a second sample if the second sample is drawn from a different population. This is particularly true when samples are drawn from clinical and nonclinical samples. For example, a measure of eating disorder symptomatology may be extremely reliable in a sample of males without an eating disorder, since the entire sample will have very low scores on all items. However, in a clinical sample of women with eating disorders, there will be greater variation in scores that introduces greater possibility for measurement error and hence a nonreliable measure.

J. Tim Newton and Koula Asimakopoulou

See also Factor Analysis and Principal Components Analysis; Health Status Measurement, Construct Validity; Health Status Measurement, Face and Content Validity; Intraclass Correlation Coefficient

Further Readings

Altman, D. G. (1999). *Practical statistics for medical research* (2nd ed.). New York: Chapman & Hall.

Bland, J. M., & Altman, D. G. (1986). Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet, 1,* 307–310.

Bland, J. M., & Altman, D. G. (1997). Statistics notes: Cronbach's alpha. *British Medical Journal, 314,* 572.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*(3), 297–334.

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin, 2,* 420–428.

Health Status Measurement, Responsiveness and Sensitivity to Change

For medical decision making, the purpose of measurement is often to measure change in health over time associated with treatment. With this in mind, there has been considerable attention paid to the ability of health status instruments to measure change, often called responsiveness or sensitivity to change. First, it should be noted that there is debate about whether or not to use change scores (the difference between scores at two points in time) at all, and although this discussion is beyond the scope of this entry, further readings have been provided as a starting point for those interested in delving more deeply into this topic. This entry defines responsiveness and sensitivity, provides context for the use of these terms, and explains the main approaches used in validation studies of health status measures.

Definitions

There is a significant body of literature about the appropriate methods to use when comparing measurement instruments over time in the absence of a gold standard. Unfortunately, there are several taxonomies in use, some with overlapping terms. Therefore, it is important to define key terms used in validation studies. Some have argued for using two terms, *sensitivity* and *responsiveness,* to describe the ability to measure change. Within this framework, *sensitivity* denotes the ability to detect any change at all and is assessed using calculations based on the variation within samples. These methods are often called *distribution-based*. *Responsiveness* refers to the ability to measure clinically important change and is calculated using external criteria, or *anchors*, to provide meaning for a specified magnitude of change in score; it is sometimes called an *anchor-based* approach. Another way of thinking about responsiveness is that it translates change on the new instrument into similar change on a familiar scale (the anchor), with the aim of enhancing the interpretability of results from the new measurement instrument. Commonly used external criteria or anchors include clinical tests, performance tests, and ratings of status by providers, patients, and caretakers. For example, a responsiveness study may calculate how much change on New Scale *X* would be associated with patient-reported "mild improvement" or a one-level change in self-report of symptom level (e.g., from moderate to mild). In this way, various anchors may be used to calculate *minimal important difference* (MID), representing the amount of change in score that could be considered clinically meaningful. Implicit in the selection of the anchor is the perspective of interest for the external criterion.

There is continued dialogue about conceptual frameworks and optimal naming conventions. Some argue against the distinction made between validity and responsiveness, noting that responsiveness is the ability to measure "what is intended" as it changes and is, therefore, more appropriately named *longitudinal validity*. Using this framework, the terms *responsiveness* and *sensitivity* refer to aspects of construct validity. However, in practice, *responsiveness* and *sensitivity* are often used more generally to describe the ability to measure change. In this way, the terms are not meant to be distinguished from validity, or to distinguish between measuring meaningful change and measuring any change at all. In this entry, the term *responsiveness* is used to describe anchor-based aids to interpretability, to denote the ability to measure clinically meaningful change. The term *sensitivity* refers to distribution-based methods to characterize the ability to detect change. Both are framed within the realm of longitudinal construct validity studies.

Methods

A wide range of methods is available for appraising the responsiveness and sensitivity of measurement instruments, enabling researchers to choose the method that best suits their measurement purposes. However, this poses challenges for comparisons of validation studies and for facilitating the interpretation of changes in health-related quality of life (HRQOL) measurement instruments. A review conducted in 2002 reported 25 definitions of responsiveness and 31 different statistical measures. Studies comparing approaches report a

range of agreement in results using different methods. While many debate the merits of various estimates, others call for consensus to minimize confusion and doubt about the validity of HRQOL measurement overall. Still others promote the benefits of the scope of information that comes from various approaches to validation studies.

Sensitivity to Change

Distribution-based methods for assessing sensitivity generally measure change relative to an estimate of variation expressed as the ratio of raw score change ($\text{Mean}_2 - \text{Mean}_1$) to measure of variance (standard deviation of the baseline score) and can vary depending on time points and patient groups selected. The types of variation considered can be generally categorized as reflecting one of three statistical characteristics. These are statistical significance (e.g., paired *t* test), sample variation (e.g., effect size and standardized response mean), and measurement precision (e.g., standard error of the measurement [SEM]). The effect size and standardized response mean (SRM) provide estimates of group change in terms of standard deviation units; so, for example, an effect size of 1.0 can be interpreted as change on the order of 1 standard deviation. The calculation for these statistics can be seen below. An effect size in the area of .2 is considered small, .5 is medium, and .8 is large. For the *t* test, effect size, and SRM, larger values indicate greater responsiveness. In contrast, SEM, as calculated below, estimates change for an individual, and smaller values represent better responsiveness.

$$\text{Effect size} = \frac{\text{Mean}_2 - \text{Mean}_1}{\text{Standard deviation}_{\text{baseline}}}.$$

$$SRM = \frac{\text{Mean}_2 \quad \text{Mean}_1}{\text{Standard deviation}_{\text{Change score}}}.$$

$$SEM = \text{Standard deviation}_{\text{baseline}} \sqrt{1 - \text{Reliability}_{\text{Test-Retest}}}.$$

Responsiveness

Various approaches have been developed to attach meaning to the magnitude of change in a measurement instrument. Characterization of MID requires comparison with an external criterion for health change. A common approach to MID is to calculate the mean change for the group within a study that fulfilled the criteria for important change. For example, in a validation study of health indexes, minimal important difference was defined as one level of change reported using a symptom satisfaction question and using a 10- to 19-point change in score using a disease-specific disability index. Approaches exist to elicit individual estimates of minimal important difference prospectively. Because MID estimates do not incorporate any information about the variability of the sample, methods have been developed to provide this information. For example, the responsiveness statistic (RS-MID) divides the MID by a measure of variation for those in the sample who were unchanged. The responsiveness statistic incorporates information about the distribution and judgment about meaningful change from an external criterion. It therefore is both distribution and anchor based. The criteria available for comparing MID estimates necessarily constrain the scope of interpretation of the responsiveness estimate. Therefore, the nature of the criterion—whether self-reported, performance based, clinician reported, or diagnosis based—must be considered when forming conclusions about the validity of the inferences that can be made from the instrument.

Interpretation

Differences between estimates of responsiveness and sensitivity using various approaches can be disconcerting. However, the interpretation of responsiveness and sensitivity estimates rests on the conceptual framework for the statistical procedures used. Below, a sample of the considerations related to a subset of measures is discussed.

The SRM uses a measure of variation in observed change of the sample in the denominator, while the effect size uses a measure of baseline variation of the sample at baseline. If the individuals all experience similar, large change, the SRM would be small relative to the effect size. Similarly, if there is little variation in the population at baseline, effect size will be relatively large. Some argue that statistics such as the SRM, using the standard

deviation of change, characterize statistical significance rather than sample variation and should not be used in longitudinal validity studies. Theoretically, distribution-based statistics may capture change beyond measurement error that is not necessarily clinically meaningful.

Anchor-based approaches to estimating responsiveness provide information about the ability to capture change relative to a relevant, external criterion, or anchor, explicitly incorporating judgments about important change. MIDs do not incorporate sample variation or the variability inherent in the measurement system. Incorporating variability into MID may make interpretation of the measure more complex. MIDs are reported in the units of the system of interest and can be understood within the relevant context, while the dimensionless RS-MID is less familiar to most audiences than effect size, SRM and SEM, and MID in most circumstances and may be challenging to interpret. It may be that advantages of accounting for variability in the measurement system could be outweighed by the challenges of interpreting and communicating the meaning of this statistic.

Although it is distribution-based, the SEM is conceptually different from the effect size and the SRM. The SEM incorporates the standard deviation of the sample at baseline and the reliability of the measurement instrument and is used to interpret individual change. According to classical test theory, the SEM is a property of the measurement instrument and can be applied across populations. There is debate about the interpretation of the SEM relative to the magnitude and meaning of change. Various authors support the use of 1 to 2 SEM as an indication of change beyond measurement error. Furthermore, these thresholds are applied as thresholds for minimal important change. Investigations into the relationship between sensitivity and responsiveness measures have reported that MID estimates center on one half of a standard deviation, suggesting that 1 SEM is a reasonable threshold for important change.

From a practical standpoint, when planning a study, effect size or SRM would be useful to inform sample size calculation and system selection relative to distributional characteristics. If the treatment and study goals address a particular dimension, MIDs based on highly relevant anchors (e.g., symptom satisfaction) may add valuable information. To enhance interpretability, the SEM provides estimates of the threshold for significant change. The MID enhances the interpretability of change by estimating the threshold for important change on the group level from a specific perspective. MIDs from various perspectives may provide important information about the orientation of the measurement systems under consideration and therefore guide system choice.

Finally, responsiveness and sensitivity estimates make up a portion of a larger array of approaches available to study the validity of health status measurement instruments. By choosing statistical procedures based on their design to address specific hypotheses, and interpreting their results within this context, investigators contribute to the larger process of accumulating evidence about the level of confidence with which decision makers can make inferences based on scores from health status measurement instruments.

Christine M. McDonough

See also Health Outcomes Assessment; Health Status Measurement, Construct Validity; Health Status Measurement, Floor and Ceiling Effects; Health Status Measurement, Generic Versus Condition-Specific Measures; Health Status Measurement, Minimal Clinically Significant Differences, and Anchor Versus Distribution Methods; Health Status Measurement, Reliability and Internal Consistency; Health Status Measurement Standards

Further Readings

Beaton, D. E., Boers, M., & Wells, G. (2002). Many faces of the minimal clinically important difference (MCID): A literature review and directions for future research. *Current Opinion in Rheumatology, 14,* 109–114.

Cronbach, L. J., & Furby, L. (1970). How should we measure "change"—or should we? *Psychological Bulletin, 74*(1), 68–80.

Guyatt, G. H., Osoba, D., Wu, A. W., Wyrwich, K. W., & Norman, G. R. (2002, April). Methods to explain the clinical significance of health status measures. *Mayo Clinic Proceedings, 77,* 371–383.

Guyatt, G. H., Walters, S., & Norman, G. R. (1987). Measuring change over time: Assessing the usefulness of evaluative instruments. *Journal of Chronic Diseases, 40,* 171–178.

Liang, M. H. (2000, September). Longitudinal construct validity: Establishment of clinical meaning in patient evaluative instruments [See comment]. *Medical Care, 38*(9 Suppl.), II84–II90.

McDowell, I. (2006). *Measuring health: A guide to rating scales and questionnaires* (3rd ed.). New York: Oxford University Press.

Streiner, D. L., & Norman, G. R. (2003). *Health measurement scales*. New York: Oxford University Press.

Terwee, C., Dekker, F., Wiersinga, W., Prummel, M., & Bossuyt, P. (2003). On assessing responsiveness of health related quality of life instruments: Guidelines for instrument evaluation. *Quality of Life Research, 12,* 349–362.

Health Status Measurement Standards

Health status measurement is important in determining the health of a population. The definition of health represents one of the contemporary challenges in health services research, as defining health is complex and measures of health vary. Defining health status measurement is relative, as health is a multidimensional, multiconstruct concept for which users of any metric of health must provide appropriate context if they are to understand how the outcome applies to their research needs. Before further discussing health status measurement standards, one must have a clear definition of health.

What Is Health?

The most widely accepted definition of health was adopted in 1948 by the World Health Organization (WHO) for an individual anywhere in the world. WHO defines health as "a state of complete physical, mental and social well-being and not merely the absence of disease or infirmity." The global community did not adopt a definition of health to measure either presence or absence of disease. Health is multidimensional, involving the physical health of the body, mental/emotional health, and social well-being. This definition, as discussed by Donald Barr in his book *Health Disparities in the United States*, while all-encompassing in its appeal, has a limited perspective for health policy due to the lofty, impractical expectation of health that few can truly achieve. By setting an unattainable standard of health, individuals are automatically positioned for failure to meet the expectation of health within this definition. WHO's widely accepted definition seems to provide a black-and-white view of health with few shades of gray—that is, one is either in good health or not. Thus, there is no way of tracking the degree to which health changes over time to determine improving or deteriorating health. Likewise, there are few mechanisms for comparison between individuals. The WHO definition provides an opportunity to understand the three approaches to health. These approaches address physical health, mental health, and social well-being. First, it is necessary to explore each approach and then discuss the potential for quantifying health.

Physical Health as the Absence of Disease: The Medical Approach

A definition of health provided by sociologist Andrew Twaddle in 1979 was used by the U.S. medical profession for much of the 20th century, according to Barr. Discussed in Twaddle's definition was the need for it to be understood that health first and foremost is a "biophysical state" and that illness is any state that has been diagnosed by a competent professional. These two components as identified by Twaddle according to the medical model include the following: (a) absence of symptoms (e.g., sensations noticed by the patient and interpreted as abnormal) and (b) absence of signs (e.g., objective criteria noted by a medical professional). This medical model approach tells us what health is *not*. If a person has abnormal signs or symptoms, according to this approach, the medical model does not define or discuss what health is. This is termed in medicine as a *rule-out* definition. Here, the health professional looks for the presence of abnormal signs or symptoms. And when both are absent, it is possible to rule out ill health. If one does not have ill health, then from this medical model perspective, the individual is healthy.

The medical approach in isolation creates problems. If the doctor and the patient disagree, which component takes precedence? Consider the scenario if a patient has the symptoms of a headache, yet the doctor, after running all available

diagnostic tests, finds no signs of illness? Through ruling out ill health, the doctor can reassure the patient that he or she is healthy. However, the patient may still feel the headache and thus be unhealthy and expect to be treated as such. The converse is true, too. What if the patient has no abnormal symptoms? Consider a patient with high cholesterol. A person with high cholesterol develops symptoms over a period of time. Should we consider a person with elevated cholesterol to be unhealthy, if we also take into account that this individual's cholesterol may eventually lead to future health problems? Despite this person's feeling "fine," what might be the consequence of stigmatizing this person as "unhealthy"?

Within the medical approach, concerns exist about the reliability of objective testing measures creating abnormal illness signs due to variability of tests (e.g., EKGs, CAT scans, laboratory tests) both in individual interpretation (e.g., physician, test administrator) and due to differences in test specificity and sensitivity. Across the country, variations in the interpretations of findings by doctors have been documented, including variations in considering what comprises both a "normal" and an "abnormal" finding.

The Psychological Approach

The psychological state of the health of an individual, measured on a survey, is purely determined by the self-assessment of that individual and not by an independent evaluator. The question of how you would rate your overall feeling of well-being on a scale of 1 to 10 represents an example of such an assessment. Likert scales and a variety of other measures have been developed to measure self-perception of health. Often these measures are time sensitive, as time-specific individual circumstances are likely to influence answer choices and apt to cause change within these self-assessments of health. The mental health scores of an individual facing a particular stressor, such as an event or challenge (e.g., a test), may reflect a lower sense of well-being. However, after the stressor is resolved, that same individual may report substantially improved well-being. Additionally, issues with mental health reports by proxy have created a potential selection or interpretation bias. An example is if parents answer a questionnaire about their child's mental state and their subjectivity does not reflect their child's true mental health state.

Social Health and Functioning Approach

The level of functioning within one's social context was an approach to health taken by sociologist Talcott Parsons in 1972. This health approach applies less to the actual physiology of the individual and more to what the person is able to do with his or her body. This approach assesses the ability to function despite any limitations. Therefore, this concept removes the dichotomous view of health as defined by WHO and places emphasis on an individual's own social circumstances and social roles to define normal functioning.

Health comparisons between two individuals may be problematic again if this social health and functioning definition is considered in isolation. Differing states of health may be assessed when two people have different social roles and tasks but the same physical functioning. Consider the example of a concert pianist and a person who packs fish, both of whom are afflicted with carpal tunnel syndrome. The concert pianist may be seen as unhealthy because the condition affects the ability to play the piano and may be more likely to have a medical intervention. On the other hand, a fish packer on the assembly line with carpal tunnel syndrome may warrant little consideration despite the great discomfort and numbness of his hands. Inequalities in social economic status, including educational opportunities and attainment, may influence and lead to clearly dissimilar health experiences due to defined roles and tasks. Therefore, supporting this model of health in isolation may perpetuate these inequalities.

Functioning is important for defining health in people with disabilities and the elderly, as discussed by health services researcher Lisa Iezzoni. Iezzoni discusses "function status" as the end result of a person's health. Specifically for the evaluation of health for people with disabilities, function status measures have been typically grouped into activities of daily living (ADLs; e.g., eating, walking, bathing, dressing, toileting) and instrumental ADLs (IADLs; e.g., cooking, housework, shopping, using public transportation, managing personal finances, answering the phone). As described by Iezzoni in the book *Risk Adjustment for Measuring*

Healthcare Outcomes, measuring function status includes several challenges. Function status is not operationalized by demographics or clinical characteristics alone. Numerous studies have shown that demographics (e.g., age, sex, gender) and pathology explain only part of function status variation. Additionally, function status measures do not apply similarly across all conditions and patient populations, as there have been floor and ceiling effects with respect to function status measurement—that is, instruments have failed to detect improvement in functioning because those with poor health tend to remain constantly low (floor effect), and those with higher functioning remain unnoticed by instruments. Also, the mode of administration creates differences in functional status measurement, as face-to-face administration results in a more optimistic measure than self-administration. Moreover, disease-specific measures may be more appropriate than a generic universal measure of functioning. The use of condition-specific measurement scales (e.g., Arthritis Impact Measurement Scale, Visual Analogue Pain Scale, Gait Evaluation) are more sensitive to function changes for individuals with these conditions than a generic universal measure. Last, there exists conflict between single, composite, and summary measures of function versus multiple scales capturing different dimensions of function status. Single measures of health may well result in misrepresentation of function. In instruments like the 36-Item Short-Form Health Survey (SF-36), two summary measures of health are created, the Physical Component Summary (PCS) and the Mental Component Summary (MCS) scales. The implications of these composite measures are still being explored.

Health Is Multidimensional

Each approach to health described above implies something about individual health states; however, the overall state of health is more ambiguous. Health represents multiple dimensions. One health services researcher, Frederick Wollinsky, suggests dichotomizing health for each of the above three approaches (i.e., psychological, social, and absence of disease) into "well" and "ill" health, measuring health by ratio of "well" to "ill" dimensions of health. The SF-36 developed by John Ware and colleagues represents a multidimensional measure of health, except that, instead of the dichotomous measures suggested above, the SF-36 measures each dimension of health as a continuous measure of health.

The SF-36 uses different combinations of the 36 items in the instrument to create eight distinct scales, each measuring a different dimension of health. As discussed earlier, four of the eight scales create the PCS, which is a summary measure of physical health, and the remaining four create the MCS, which is a summary measure of mental health. The four scales to create the PCS are (1) physical functioning, (2) role limitations due to physical problems, (3) bodily pain, and (4) general health perceptions. The four scales that make up the MCS are (1) vitality, (2) social functioning, (3) role limitations due to emotional problems, and (4) general mental health.

Using a multidimensional instrument, such as the SF-36, allows providers and health service researchers to assess health across medical, social, and psychological constructs. These three constructs of health are causally linked to each other, as physical health changes create changes in social roles, which ultimately affect mental health. Each aspect of these dimensions is unique to the characteristics of both the individual and the environment in which that individual lives. These three dimensions are not the final outcome, but rather the intermediate factors that will ultimately affect health-related quality of life. Individual and environmental characteristics will buffer or enhance the health of an individual. For example, for an individual with strong social, psychological, and environmental support, a specific symptomatology may result in a smaller impact on functional status, whereas for another with similar symptomatology, weaker support may result in greater impact on function status, leading to poorer health. Quality-of-life measures are greatly affected by symptoms of illness and functional limitations; however, the presence of these symptoms alone will not result in reduced quality of life. For example, in a *BMC Public Health* 2008 article, "Age at Disability Onset and Self-Reported Health Status," Eric Jamoom and colleagues were able to show that the age at which one acquires symptoms or activity limitation is associated with health status differences. Health status measurement is tied intimately to context of health, and the jury is still out on a standard measure of health. Because health

status is so complex, there is no single gold standard of health status measurement. Often a gold standard is considered a reference to determine whether a newer instrument adequately measures health as reflected in the older instrument. Therefore, an older survey is often used as a gold standard for a comparison to determine the criterion validity of a newer survey to ensure that the instrument still adequately represents the measures from the reference instrument. However, health status measures are subject to limitations with validity and reliability considerations, and definitions of health measurement are required to continue to understand the context for distinction.

Definitions of Health Status Measurement

A panel of 57 experts with backgrounds in medicine, biostatistics, psychology, and epidemiology participated in the COSMIN (COnsensus Standards for the selection of health Measurement INstruments) Delphi study to develop standards for selecting health measurement instruments. Definitions below are provided from the preliminary version of the COSMIN Checklist as designed in the protocol from Mokkink and colleagues in *BMC Medical Research Methodology*.

Reliability: The degree to which the measurement is free from measure error. The extent to which scores for patients who have not changed are the same for repeated measurement under several conditions: for example, using different sets of items from the same health-related patient-reported outcomes (HR-PRO) (internal consistency); over time (test-retest); by different persons on the same occasion (interrater); or by the same persons (i.e., raters or responders) on different occasions (intrarater).

Internal consistency: The degree of interrelatedness among the items.

Measurement error: The systematic and random error of a patient's score that is not attributed to the true changes in the construct to be measured.

Validity: The degree to which an instrument truly measures the construct it purports to measure.

Content validity: The degree to which the content of a HR-PRO instrument is an adequate reflection of the construct to be measured.

Construct validity: The degree to which the scores of a HR-PRO instrument are consistent with hypotheses based on the assumption that the HR-PRO instrument validity measures the construct to be measured.

Criterion validity: The degree to which scores of a HR-PRO instrument are an adequate reflection of a gold standard. A gold standard for HR-PRO instruments does not exist. When assessing criterion validity of a shortened questionnaire, the original long version may be considered as the gold standard.

Cross-cultural validity: The degree to which the performance of the items on a translated or culturally adapted HR-PRO instrument are an adequate reflection of the performance of the items of the original version of the HR-PRO instrument.

Face validity: The degree to which items of a HR-PRO indeed look as though they are an adequate reflection of the construct to be measured.

Eric W. Jamoom

See also Health Outcomes Assessment; Health Status Measurement, Construct Validity; Health Status Measurement, Face and Content Validity; Health Status Measurement, Reliability and Internal Consistency; SF-36 and SF-12 Health Surveys

Further Readings

Barr, D. A. (2008). *Health disparities in the United States: Social class, race, ethnicity, and health*. Baltimore: Johns Hopkins University Press.

Iezzoni, L. I. (1997). *Risk adjustment for measuring healthcare outcomes*. Chicago: Health Administration Press.

Jamoom, E. W., Horner-Johnson, W., Suzuki, R., et al. (2008). Age at disability onset and self-reported health status. *BMC Public Health*, *8*, 10.

Kane, R. L. (2006). *Understanding health care outcomes research*. Sudbury, MA: Jones & Bartlett.

Mokkink, L. B., Terwee, C. B., Knol, D. L., Stratford, P. W., Alonso, J., Patrick, D. L., et al. (2006). Protocol of the COSMIN study: COnsensus-based Standards for the selection of health Measurement INstruments. *BMC Medical Research Methodology, 6*, 2. Retrieved January 20, 2009, from http://www.biomedcentral.com/1471-2288/6/2

National Quality Measures Clearinghouse: http://www.qualitymeasures.ahrq.gov

Research at the Research Rehabilitation and Training Center: Health & Wellness: http://www.ohsu.edu/oidd/rrtc/research/index.cfm

SF-36.org: A community for measuring health outcomes using SF tools: http://www.sf-36.org

World Health Organization. (2003). *WHO definition of health*. Retrieved January 20, 2009, from http://www.who.int/about/definition/en/print.html

Health Utilities Index Mark 2 and 3 (HUI2, HUI3)

The Health Utility Index Mark 2 and 3 (HUI2 and HUI3, collectively referred to as HUI2/3) are generic preference-based health status measures designed to provide utility scores for health outcomes evaluation. The HUI2/3 measure is generic, applicable to a wide range of conditions, treatments, and populations. Preference-based design provides a single-summary numerical score for each health state defined by the instrument. This form of presenting health outcomes is particularly useful in economic evaluation, where a single index of health-related quality of life that summarizes health status utilities is needed. The first version of the HUI was created in the 1970s to evaluate the outcomes of neonatal intensive care for very-low-birth-weight infants. Since then the HUI measures have continued under development, and there are now two versions available: HUI Mark 2 and HUI Mark 3. Nowadays, the HUI1 is rarely applied, while the HUI2/3 is continually used. Both of these measures have been employed in various clinical studies, in economic evaluations, and in population health surveys.

The HUI Mark 2/3 is a multi-attribute health status measure. In such a measure, the defined health is expressed in terms of key attributes of interest. In other words, the descriptive system, which is known as a classification system, comprises chosen attributes and reflects the health definition adopted by the instrument developers. The term *attribute* used here is the same as the term *domain* or *dimension* defined in other measures.

Development

The first version of the HUI was developed based on the pioneering work on the Quality of Well-Being (QWB) instrument by Fanshel and Bush in 1970. The conceptual framework of the QWB provided a template for developing the HUI instrument. To evaluate the outcomes of neonatal intensive care for very-low-birth-weight infants, Torrance and his colleagues in the 1980s expanded the QWB's descriptive system into the classification system of the HUI Mark 1, which consisted of four attributes: (1) physical function, (2) role function, (3) social-emotional function, and (4) health problems, with six, five, four, and eight levels per attitude, respectively, thus defining a total of 960 unique health states.

The HUI Mark 1 was further extended for pediatric application. Aiming to measure the long-term outcome of childhood cancer, Cadman and his colleagues reviewed the literature and created a list of potentially important attributes for health-related quality of life. They invited 84 parent-and-child pairs to select the most important attributes from the list and, as a result, a core set with six attributes was created. These six attributes are (1) sensory and communication ability, (2) happiness, (3) self-care, (4) pain or discomfort, (5) earning and school ability, and (6) physical activity ability. The six attributes identified by Cadman and his colleagues, plus an attribute of fertility, for capturing the impact of child cancer treatment on fertility, became the HUI Mark 2 classification system. Each attribute has three to five levels of function and, therefore, defined 24,000 unique health states. Although the development of the HUI2 is for measuring treatment outcomes of children with cancer, the HUI2 soon was used as a generic measure for adults, and it was applied to different populations and conditions due to its generic-like attributes. It is suggested that the HUI2 can be used as a generic measure for the child population after removing the attribute of fertility.

The HUI3 was developed to tackle some concerns of the HUI2, to be structurally independent, and to be applicable in both clinical and general population studies. There are several changes in the HUI3's multi-attribute system as compared with its predecessor version. The changes included removing the attribute of self-care and replacing it with dexterity to achieve structural independency, adding distinct components of sensation such as vision, hearing, and speech, and excluding the attribute of fertility. Therefore, there are eight attributes in the HUI3 system: (1) vision, (2) hearing,

(3) speech, (4) ambulation, (5) dexterity, (6) emotion, (7) cognition, and (8) pain. Each attribute has five to six response levels that are combined in such a way that the HUI3 defines in total 972,000 unique health states.

As described by the instrument's developers, the choice of attributes in the HUI2/3 was based on a "within the skin" approach, focusing on the attributes of physical and emotional health status that are fundamentally most important to health status measurement and excluding the aspect of participation in society, as in social activity or role performance. Furthermore, the design of the descriptive system of the HUI2/3 was aimed to record functional capacity rather than performance. The reason for this was that a measure of performance reflects the level of capacity of an individual on a chosen function. Thus, people with the same underlying functional capacity could have a different level of performance.

Utility

As a preference-based measure, each health state defined by the HUI2/3 descriptive system can be assigned a utility score. It is calibrated by applying a formula (a scoring algorithm) to a health state, which essentially attaches values (also called weights) to each of the levels in each attribute. The HUI2 scoring algorithm was developed from the valuation of a random sample of 293 parents of schoolchildren in Hamilton, Ontario, Canada. Both visual analog scale (VAS) and standard gamble (SG) methods were adopted to elicit preferences from the sample. Each participant was asked to value 7 single-attribute states and 14 multi-attribute states using a VAS. Participants were also asked to value 4 multi-attribute states that overlapped with those 14 states in the previous task, using the SG method. Based on the data, a power transformation was developed to convert the value of the health state measured by the VAS into a utility value as elicited by the SG. The scoring algorithm was developed based on the multi-attribute utility theory, where utility function for each attribute was estimated separately and a multiplicative function form was adopted in the final scoring formula.

Multi-attribute utility theory (MAUT) is a method to estimate a mathematical function, which allows for calibration values for a large number of health states defined by a multi-attribute classification system, based on values of a small, carefully selected set of those states. The basic approach is to measure the utility function for each single attribute and to identify an equation that expresses the overall utility as a function of these single-attribute utilities (details can be found in the paper by Torrance and his colleagues published in 1996, as listed in Further Readings, below). MAUT can reduce the number of health states required for valuation to develop a scoring formula by assuming a function form in advance. The choice of functional form imposes a restriction in terms of how each attribute in the classification system is related to the others. There are three typical function forms available—additive, multiplicative, and multilinear. The evidence obtained from the HUI2 studies by Torrance and his colleagues supported the choice of multiplicative functional form (multilinear form was not considered because it requires the measurement of a large amount of multi-attribute health states for calculation).

One of the unique features of the MAUT method is the corner state. The corner state is a multi-attribute state, where one attribute is set at one extreme (the worst level of functioning) and the rest are set at the other extreme (the best level of functioning), and participants are asked to value several corner states. However, the structural independence of the classification system is a prerequisite for evaluating such corner states. For instance, participants could not imagine and consequently had difficulty valuing a corner state where a person was unable to control or use arms and legs (mobility attribute) but had no problem with self-care attributes such as eating, bathing, dressing, and using the toilet. Therefore, it is necessary to have no correlation between attributes—these must be structurally independent in such a way that each corner state, combining the worst level in one attribute with others at the best level, is possible. This issue was first found in the valuation study of the HUI2, and it has now been taken into account in the redesign of the HUI3. That is to say, the HUI3 is structurally independent.

The scoring formula of the HUI3 used a similar approach based on the responses from a representative sample of 504 adults from Hamilton, Ontario. Each participant was asked to value three anchor and three marker states and 22 to 24 multi-attribute health states using a VAS, plus 5 states using the SG method. The study also examined the possibility of

applying a simplified multilinear functional form, but it was concluded that the multiplicative form performed better in such a degree that the final scoring algorithm was based on multiplicative form.

Both the HUI2 and HUI3 offer a single numerical score for health states, anchored at 0 and 1, representing death and full health, respectively. A negative score indicating worse than dead is also allowed. The range of possible utility scores is from 1 to –.03 for the HUI2 and from 1 to –.36 for the HUI3.

There is extensive evidence supporting the reliability, validity, and responsiveness of the HUI2/3. The interested reader can refer to additional sources such as McDowell or Horsman and his colleagues, as listed in Further Readings, below. The minimal clinically important difference between HUI scores ranges from .02 to .04.

Current Versions

Currently, there are several versions of the HUI2/3 questionnaires available, depending on administration, whether they are self- or proxy assessed, and recall period. The HUI2/3 can be either self-completed or interviewer administered over the phone or in person. Both self- and proxy assessment are available. There are four different standard recall periods for each questionnaire available: 1 week, 2 weeks, 4 weeks, and "usual." The questionnaire with recall periods in weeks is usually applied in clinical studies or economics evaluation, while the questionnaire with the recall period, "usual," which does not specify the time, is mostly applied in population health surveys.

The HUI2 and HUI3 can be combined and applied together, which is known as the HUI. The HUI includes both classifications of the HUI2 and HUI3 and generates utility scores for both the HUI2 and HUI3. Further information on the HUI instruments can be found at the Health Utilities Inc. Web site.

Ling-Hsiang Chuang

See also Health Status Measurement, Generic Versus Condition-Specific Measures; Multi-Attribute Utility Theory

Further Readings

Feeny, D. H., Furlong, W. J., Boyle, M., & Torrance, G. W. (1995). Multi-attribute health status classification systems: Health Utility Index. *PharmacoEconomics, 7,* 490–502.

Feeny, D. H., Furlong, W. J., Torrance, G. W., Goldsmith, C. H., Zenglong, Z., & Depauw, S. (2002). Multiattribute and single-attribute utility function: The Health Utility Index Mark 3 system. *Medical Care, 40,* 113–128.

Health Utilities Inc.: http://www.healthutilities.com

Horsman, J., Furlong, W., Feeny, D., & Torrance, G. (2003). The Health Utilities Index (HUI): Concepts, measurement properties and applications. *Health and Quality of Life Outcome, 1,* 54–67.

McDowell, I. (2006). *Measuring health: A guide to rating scales and questionnaires* (3rd ed.). New York: Oxford University Press.

Torrance, G. W., Feeny, D. H., Furlong, W. J., Barr, R. D., Zhang, Y., & Wang, Q. (1996). A multi-attribute utility function for a comprehensive health status classification system: Health Utilities Mark 2. *Medical Care, 34,* 702–722.

Torrance, G. W., Furlong, W. J., Feeny, D., & Boyle, M. (1995). Multi-attribute preference functions: Health Utility Index. *PharmacoEconomics, 7,* 490–502.

Healthy Years Equivalents

The healthy years equivalent (HYE) provides a user-friendly metric that is needed for improved communication within and among researchers, decision makers, practitioners, and consumers. Unlike the quality-adjusted life year (QALY), which means different things to different people and often is not consistent with the underlying principles of cost-utility analysis (CUA), the HYE means only one thing—it is a utility-based concept, derived from the individual's utility function by measuring the number of years in full health, holding other arguments in the utility function constant, that produces the same level of utility to the individual as produced by the potential lifetime health profile following a given intervention. The measurement of HYE requires that individuals will be allowed to reveal their true preferences. This is because it seems reasonable, when asking the public to assist in the determination of healthcare priorities, to choose measurement techniques that allow the public to reveal their true preferences even if this requires the use of more complex techniques. If not, why do we bother asking them at all?

Concept

The underlying premise of cost utility/effectiveness analysis (CUA/CEA) is that for a given level of resources available, society or the decision maker wishes to maximize the total aggregate health benefits conferred by a proposed treatment. The principles underlying CUA/CEA are concerned with the simultaneous satisfaction of efficiency in both production (i.e., making sure that each level of outcome is produced with the minimum amount of resources) and product mix (i.e., making sure that the allocation of available resources between different "products" is optimal). In this way, CUA/CEA is consistent with the principles of welfare economics theory and a welfarist approach to economics. But to achieve the goal of maximizing health-related well-being (i.e., utility associated with health benefits) from available resources, the methods used to measure health-related well-being must be consistent with the theories on which the welfarist approach and principles of CUA are based. Under the welfarist approach, an individual's preferences are embodied in that individual's utility function. Thus, for a measure of outcome to be consistent with the welfarist approach, and hence CUA/CEA principles, it must be consistent with a theory of utility. Health (i.e., an expected profile of health over lifetime) is one argument in an individual's utility function.

From the perspective of the economist qua economist, "pure utility" is sufficient for comparing alternatives on the basis of individuals' preferences. Hence, for a von Neumann-Morgenstern (vNM)–type individual, for example, a single standard gamble (SG) question can provide the utility score (i.e., number of utils) for any potential lifetime health profile. But the utils measure, and the notion of cost per util, may not be very meaningful to individuals and organizations making choices among programs associated with different expected health profiles. Abraham Mehrez and Amiram Gafni, who were the first to introduce the approach, explained that the HYE responds to the need to improve communication within and among researchers, managers, practitioners, and consumers in a way that is consistent with the concept of utility and hence represents individuals' preferences. The HYE is *not* a direct measure of utility. It is an attempt to reflect individuals' preferences concerning uncertain health profiles using one argument in their utility function (i.e., duration), holding health status constant (i.e., full health). The intuitive appeal of years in full health has been established by the QALY measure, which was designed to be thought of as an equivalent number of years in full health—a number of quality-adjusted life years (QALYs). A different name was chosen to distinguish the HYE, which is a utility-based concept, from the QALY. Furthermore, it has been argued that length of life in full health—that is, healthy-years equivalent—represents a much simpler concept to explain to decision makers than the variable quality of life health status annuity, which the QALY represents.

The need for distinguishing HYEs from QALYs stems from the observation that QALYs mean different things to different people. In the health services research literature, most proponents and users of the QALY approach do not subscribe to the notion of an underlying utility model. For them, QALY is simply an index (i.e., the QALY measures years of life adjusted for their quality) with intuitive meaning. It is a measure of the individual's health status as distinct from the utility associated with this health state. There are others who subscribe to the concept of QALY as a measure of utility and who identify the utility model for which this would be the case. Those who subscribe to this concept face the problem of communication (i.e., the QALY is intended to measure the number of utils generated by a health profile, not adjusted years of life). Finally, there are those who view the QALY as an index, but one in which the weights attached to durations in different health states are calculated using utility theory, typically, vNM utility theory. It has been argued that this unit of output is therefore the utility-adjusted life year, which may or may not be as intuitively appealing as the quality-adjusted life year.

In terms of the conceptual limitations of the HYE, it has been noted that the HYE definition imposes the same restrictions as the QALY in terms of the (implicit) underlying assumptions of utility independence between health and other commodities in the individual's utility function. It has also been noted that the current definition of HYE is equally as restrictive as the QALY approach in terms of the exclusion of externalities (i.e., one person's health status may affect another person's

utility) from the individual's utility function. It has been suggested that the concept of external effects is much more applicable in the case of healthcare consumption than for most other commodities because of the special nature of the commodity, "health," that healthcare is expected to produce. Hence, such effects should be included when measuring outcomes. Finally, both the HYE and QALY use the same aggregation method to arrive at a social preference—an individual's health is measured in terms of QALYs or HYEs, and the community's health is measured as the sum of QALYs or HYEs (i.e., an additive model). This may not be consistent with the equity criteria adopted for the analysis.

Measurement

Can an algorithm be developed to measure HYEs that (a) does not require additional assumptions (i.e., in addition to the assumptions of the underlying utility theory) and (b) is feasible to use with the intended subjects (i.e., the number and complexity of questions asked is not too burdensome)? Proponents of the QALY as a direct measure of utility model have recognized that the additional assumptions of this model are restrictive but justify its use on the basis of measurement feasibility (i.e., price worth paying). While recognizing the importance of measurement feasibility issues, let's deal first with the question of a measurement algorithm and then the issue of whether the trade-off between feasibility and validity is necessary or appropriate.

The concept of HYE does not require that an individual subscribe to expected (i.e., vNM) utility theory. Any type of utility theory (i.e., non-vNM) can be used as a basis for generating algorithms to measure HYEs, and the choice of utility theory will determine the method of measurement. The only requirement is that preferences for health profiles are measured under conditions of uncertainty to reflect the nature of the commodity, health. For the case of a utility maximizer (i.e., vNM)–type individual and for the case of a decision tree (a typical case in medical decision making), HYEs are measured using the two-stage lottery-based method as follows: In Stage 1, SG is used to measure the utility of all potential lifetime health profiles. These are then used in association with the *ex ante* probabilities of each potential profile to calculate the expected utility of each treatment option, measured in utils. Note that, as explained above, this is sufficient to determine which treatment is preferred by the subject, but the outcomes measured have limited intuitive appeal for users. In Stage 2, the expected utils of each treatment option are converted to HYEs (i.e., more intuitively appealing years in full health equivalents) using again the SG method.

Does the algorithm described above provide scores for health profiles that accurately reflect an expected utility maximizer preference ordering? It has been shown that in the case of uncertainty, *ex ante* HYEs always rank risky health profiles the same way as expected utility. The assumptions needed for the other measures are risk neutrality with respect to healthy time for expected HYEs; risk neutrality with respect to time in all health states and additive independence of quality in different periods for risk-neutral QALYs; and constant proportional risk posture with respect to time in all health states and additive independence of quality in different periods for risk-averse QALYs. In other words, it is possible to develop algorithms to measure HYEs in a way that either does not require additional assumptions to those required by the chosen utility theory or requires fewer and weaker assumptions as compared with those required by the QALY model. Thus, for those interested in a utility-based measure that has intuitive appeal to users while preserving the individual's preference ordering, the HYE concept provides a measure superior to the QALY.

In terms of the feasibility of the HYE measure, "the jury is still out." Measuring HYEs is likely to involve greater respondent burden, mainly in terms of the number of questions being asked. That it may be more complex and time-consuming does not imply that it should not or cannot be used at all. This has resulted in a debate between those who are willing to add assumptions (typically invalid assumptions such as additive independence) to ease the measurement burden and those who would like to relax as many assumptions as possible even at a price of a more complex technique. The need to simplify the assessment task (i.e., reduce the number of questions asked to generate HYE scores) is most evident in the case of large decision trees. This is because the number of different potential lifetime health profiles is likely to be large. However, there are many assumptions

that researchers are making to populate large decision trees with numerical values. More empirical work is required to systematically test whether the use of more accurate measures of preference in the context of smaller and simpler decision trees provides more or less accurate ranking of societal preference as compared with the use of less accurate measures of preference in the context of large decision trees.

Finally, it has been suggested that, in principle, one can try to estimate the certainty equivalent number of HYEs that will always rank risky health profiles according to individual preferences using a time trade-off (TTO) question. In this case, the risky health profile to be assessed is framed as a probability distribution and is equated to the certainty equivalent number of healthy years. Note that this technique does not require that an individual be an expected utility maximizer. However, whether this could be done in practice is not known, as it is unclear whether that type of information can be processed in a meaningful way.

Amiram Gafni

See also Cost-Effectiveness Analysis; Cost-Utility Analysis; Expected Utility Theory; Quality-Adjusted Life Years (QALYs); Utility Assessment Techniques; Welfare, Welfarism, and Extrawelfarism

Further Readings

Ben Zion, U., & Gafni, A. (1983). Evaluation of public investment in health care: Is the risk irrelevant? *Journal of Health Economics, 2,* 161–165.

Bleichrodt, H. (1995). QALYs and HYEs: Under what conditions are they equivalent? *Journal of Health Economics, 14,* 17–37.

Gafni, A. (1994). The standard gamble method: What is being measured and how it is interpreted. *Health Services Research*, *29,* 207–224.

Gafni, A., & Birch, S. (1995). Preferences for outcomes in economic evaluation: An economic approach in addressing economic problems. *Social Science and Medicine, 40,* 767–776.

Gafni, A., & Birch, S. (1997). QALYs and HYEs: Spotting the differences. *Journal of Health Economics, 16,* 601–608.

Gafni, A., & Birch, S. (2006). Incremental cost-effectiveness ratios (ICERs): The silence of the lambda. *Social Science and Medicine, 62,* 2091–2100.

Johannesson, M. (1995). The ranking properties of healthy years equivalents and quality adjusted life years under certainty and uncertainty. *International Journal of Health Technology Assessment in Health Care, 11,* 40–48.

Mehrez, A., & Gafni, A. (1989). Quality-adjusted life-years (QALYs), utility theory and Healthy Years Equivalent (HYE). *Medical Decision Making, 9,* 142–149.

Ried, W. (1998). QALYs versus HYEs: What's right and what's wrong. A review of the controversy. *Journal of Health Economics, 17,* 607–625.

Hedonic Prediction and Relativism

Standard decision theory assumes that when choosing between options that have the same costs, decision makers evaluate which option will deliver the highest expected outcome utility and choose that option. This is known as a consequentialist utility analysis method. In reality, people rarely base their decisions strictly on this approach. In recent years, behavioral decision theorists have proposed that choices are often driven by decision makers' affect, or predicted experience, toward the choice options, and that such affect-driven decisions often lead to choices different from those that the standard utility analysis would prescribe. For example, before making a decision, they tend to think about the emotions that the outcomes of their choices are likely to trigger (i.e., decision makers predict their hedonic experiences). Evidence from behavioral decision research suggests that the emotions people expect to experience in the future are important determinants of their behavior. As a result of this development, decision theorists now make a distinction among three types of utilities—decision utility (as revealed by one's choice), experienced utility (feelings with the chosen option), and predicted utility (prediction of experienced utility). The last few decades have witnessed a large amount of research on the inconsistency between predicted and actual experience.

Hedonic Prediction

Hedonic prediction is a term denoting people's current judgments about what their emotions

(e.g., happiness, distress, pain, fear) or preferences (e.g., for different health states or treatments) will be in the future. A substantial body of empirical research from a range of medical and nonmedical domains demonstrates that people typically exaggerate their emotional reactions (positive or negative) to future events. The emotions that have been investigated include pain, fear, and subjective well-being (happiness). For example, people tend to overpredict different types of acute pain (e.g., menstruation pain, headache, postoperative pain, dental pain) and chronic pain (e.g., arthritis pain and low back pain). Overprediction has also been observed when people forecast emotions such as fear and anxiety. For example, people overpredict their fear of dental treatments, confined spaces, snakes, and spiders.

Researchers have also investigated people's forecasts of the impact of specific positive and negative events that affect their well-being (such as significant life events, medical results, and treatments). In general, people overpredict the hedonic impact of negative events. For example, patients about to undergo surgically necessary amputations delay or opt out of the operations because they anticipate that their lives will be ruined without a limb. Similarly, women were found to overpredict their distress after receiving positive test results for unwanted pregnancies. There is also evidence that dieters overpredict their distress after being unable to achieve their weight-loss targets. One study demonstrated that people also overpredict the level of distress experienced by other people, for example, after positive HIV test results. People also tend to overpredict the impact of positive events. Existing evidence suggests that patients who decide to undergo cosmetic surgery are not necessarily happier after it. People are also found to overpredict the relief in distress that people with negative results experienced. Other studies have shown that people exaggerate the positive effect of a lottery win on their life, the pleasure that they will derive from a future holiday trip, and the happiness that they will experience if their favorite sports team wins.

Other related research suggests that people often have poor intuitions about the hedonic impact of gains and losses. For example, people overestimate how much hedonic benefit they will derive from small gains. People also believe that they will return to their hedonic baselines more quickly after a small loss than after a large loss even when the opposite is true. People also expect that the hedonic cost of a loss will be greater than the hedonic benefit of an equal-sized gain even when this is not so. People often think that they would be willing to pay the same amount to gain an item as to avoid losing it, while, in reality, they are willing to pay less.

In summary, the anticipation of unpleasant life events such as illnesses may be different from the actual experience of the event. In light of this difference, health economists have outlined a dual model for the evaluation of patients' preferences, which assesses patients' anticipation of an illness separately from their experience of it. Psychological accounts for the documented errors in hedonic prediction include the *projection bias*, according to which people underestimate or even completely ignore their ability to adapt to new circumstances and, as a result, tend to exaggerate the impact of positive and negative events on their well-being. Another psychological account is based on the *focusing illusion*, which states that people focus too much on the event in question (e.g., illness or treatment) and neglect other life events that will occur simultaneously with the event at the center of the attentional focus. As a result of this neglect of future events that will be competing for attention with the key event, people produce exaggerated predictions of the hedonic impact of the latter on their subjective well-being.

Relativism

When people make a choice, they tend to contemplate how they will feel if the alternative that they choose turns out not to be the best one. Such counterfactuals, between the expected outcome and those that would occur if a different choice was made, shape many decisions because they tend to trigger anticipated regret. According to regret-based models of decision making, the utility of a choice option (and, hence, the likelihood of selecting it) should depend on both anticipated regret and the subjective value of the option. Empirical research has documented that regret has a powerful effect on choice. For example, people feel regret after both action and inaction due to anticipated counterfactual regret (after receiving the outcomes arising from a choice, people experience emotions as a result of these outcomes and also as a result of

the counterfactual comparisons of what the outcomes would have been, had they chosen differently). Some argue that all behavioral choices necessarily involve potential regret. This shows that the alternate option in a choice set influences the evaluation of each option, that is, that judgment and choices are relative, because the utility of an option is not independent of alternative options in a choice set. Such relativity has been demonstrated in studies showing that anticipated regret is exacerbated when people expect to receive feedback on the outcome of the foregone alternatives.

This relativistic paradigm is in line with behavioral evidence that people generate more accurate affective forecasts when they see an event within its context of other events. The explanation for this finding is that, by eliciting a context (i.e., the full set of outcomes), people realize that the specific event (or decision in question) is only one among many determinants of their well-being and often not the most important one. For example, some authors suggest that being exposed to other patients' posttreatment experience would allow patients to put a very unpleasant treatment within the context of the rest of their lives. The patients could then realize that what appears to be the focus of their lives at the time of the decision (i.e., the treatment and its consequences) may not be the focus of their lives later on.

Relativistic comparisons can also create certain biases. For example, when people miss a good bargain, they are less likely to take a subsequent one that is not as good. This phenomenon is termed *inaction inertia,* and according to regret-based explanations of it, people anticipate that buying the item will lead to regret, because it will remind them that they missed a better opportunity to buy it. This relativistic strategy is used when the difference between the previous and subsequent bargains is large. Studies have investigated the intensity of emotions caused by relativistic forecasts and have found poor accuracy in the prediction of the intensity of emotions. For example, in a negotiation task, subjects who made high offers overrated the regret that they would experience after they failed at a negotiation in which they had expected to succeed. Similar findings were obtained for disappointment. In another study, participants overrated the rejoicing that they would experience when they received marks for their coursework that were better than what they had expected. Thus, this line of research demonstrates systematic prediction errors in both negative and positive decision-related hedonic forecasts. This evidence corroborates and complements previous work, in which accuracy was assessed by comparing judgments of forecasted with experienced emotions. One such study compared forecasted regret if a contest was lost with the experienced regret reported when respondents were led to believe that they had lost that contest. The forecasted predictions were overrated relative to the experienced regret. Similar findings emerged from two additional studies, in which commuters making only forecasting judgments overrated the regret that experiencing commuters reported after missing a train.

Note that some studies show the opposite effect of relativistic biases—that people underweight their expected emotional experiences. For example, if people are asked to analyze reasons before making a decision, then they are less likely to choose the option they will like later on (as compared with people not asked to analyze such reasons). Some researchers suggest that analyzing reasons focuses the decision maker's attention on more tangible attributes along which to compare the choice options (such as cost and benefits) and away from less perceptible feelings. Other research shows that if people are not explicitly asked to analyze reasons, they may still choose options that are rationalistic but inconsistent with predicted preferences; it suggests that people automatically seek rationalism in decision making (i.e., they spontaneously focus on rationalistic attributes such as economic values, quantitative specifications, and functions).

Shared Decision Making

Clinicians, healthcare experts, and policy makers have argued for shared decision making between patients and doctors regarding choice of medical and surgical treatments. According to this framework, the patient is a key medical decision maker in the care plan. At the center of such care plans are patient preferences, which are usually defined as positive or negative attitudes toward bundles of outcomes such as disease or treatment. However, since preferences can be predicted or experienced, a more precise definition should define them in terms of the experienced and predicted (positive or negative) feelings and emotions that patients associate

with a disease and with the outcomes of its possible treatments. However, behavioral evidence suggests that people are poor at predicting the impact of an illness and its treatment on their subjective well-being, which suggests that they have little understanding of their own future feelings and preferences. Future research should aim to reveal how patients' self-forecasts affect their choice of treatments and whether such biased forecasts could be made more accurate.

Ivo Vlaev and Ray Dolan

See also Decision Making and Affect; Decisions Faced by Patients: Primary Care; Emotion and Choice; Experience and Evaluations; Health Outcomes Assessment; Models of Physician-Patient Relationship; Patient Decision Aids; Regret; Shared Decision Making; Utility Assessment Techniques

Further Readings

Dolan, P. (1999). Whose preferences count? *Medical Decision Making, 19,* 482–486.

Dolan, P., & Kahneman, D. (2008). Interpretations of utility and their implications for the valuation of health. *Economic Journal, 118,* 215–234.

Hsee, C. K., Zhang, J., Yu, F., & Xi, Y. (2003). Lay rationalism and inconsistency between predicted experience and decision. *Journal of Behavioral Decision Making, 16,* 257–272.

Sevdalis, N., & Harvey, N. (2006). Predicting preferences: A neglected aspect of shared decision-making. *Health Expectations, 9,* 245–251.

Sevdalis, N., & Harvey, N. (2007). Biased forecasting of post-decisional affect. *Psychological Science, 18,* 678–681.

Wilson, T. D., Wheatley, T., Meyers, J. M., Gilbert, D. T., & Axsom, D. (2000). Focalism: A source of durability bias in affective forecasting. *Journal of Personality and Social Psychology, 78,* 821–836.

Zeelenberg, M. (1999). Anticipated regret, expected feedback and behavioral decision-making. *Journal of Behavioral Decision Making, 12,* 93–106.

Heuristics

The term *heuristic* is of Greek origin and means serving to assist in finding out or discovering something. Imagine that you feel sick and decide to visit a doctor. He diagnoses heart disease and prescribes a new treatment. You might want to hear the opinion of a second expert before starting the treatment. You visit another doctor, who recommends a different medication. Now you must make up your mind rather quickly about which doctor you should trust. Inferential accuracy is also crucial: An error in judgment might lead to becoming more ill or even dying. How do doctors and patients solve the challenging task of making treatment decisions under time pressure and with limited information? One way to do it is to rely on heuristics.

There are two views on heuristics in the psychological literature. According to one line of thought, they are error-prone reasoning strategies that can lead to a number of cognitive illusions and biases. In another view, heuristics are cognitive shortcuts that can lead to as good or even better judgments than more complex decision strategies.

Error-Prone Reasoning Strategies

The first view of heuristics is the result of measuring human decision making against various normative standards, such as probability theory and logic. This research program was sparked by the seminal work of Daniel Kahneman and Amos Tversky in the 1970s. By comparing human reasoning and intuitive judgment with ideal standards of rationality, researchers within this program hoped to gain insight into the underlying psychological processes. Often, though, the program is charged with supporting the view that people are inherently faulty decision makers who use cognitive shortcuts that can lead to systematic errors.

Two of the more well-known heuristics studied in this program are *representativeness* and *availability*. Daniel Kahneman and Amos Tversky proposed that when using the representativeness heuristic, people judge the likelihood that an event belongs to a certain class, or is generated by a certain process, on the basis of its similarity to that class or process, neglecting its prior probability of occurrence. For example, most people will judge that the sequence of coin tosses head-tail-head-tail-tail-head is more likely than head-head-head-tail-tail-tail, because the former is perceived to be more representative of a random sample of coin tosses.

When people use the availability heuristic, they are estimating the likelihood of an event on the basis of how easily instances of the event come to mind. For example, people may overestimate the likelihood of certain causes of death, such as tornado or flood, because they are vivid and more likely to be talked about. In contrast, the likelihood of some more frequent but less "exciting" causes of death, such as heart attack or stroke, is underestimated.

This view of heuristics has spread to medical decision making. For instance, in a seminal study of how physicians process information about the results of mammography, David Eddy gave 100 physicians information about the prior probability that a patient has breast cancer, the hit rate or sensitivity of mammography, and the false-positive rate and asked them to estimate the probability that a patient with a positive mammogram actually has breast cancer. Eddy reported that most of the physicians had difficulties with probabilities and concluded that the physicians' judgments systematically deviated from statistical rules such as Bayes's rule, emphasizing cognitive illusions. Similar results were reported with physicians and students. From these studies, many researchers have concluded that the human mind does not appear to follow the calculus of chance or the statistical theory of prediction. If these conclusions are right, there is little hope for physicians and their patients.

Researchers tried to solve this problem by training the physicians to use decision-support tools, which weight and combine the relevant information by using regression, instead of relying on their intuitive judgment. For instance, physicians at the University of Michigan Hospital are trained to use the Heart Disease Predictive Instrument, which consists of a chart listing approximately 50 probabilities. The physicians have to check for the presence or absence of seven symptoms (evidence of which is routinely obtained during the patient's admission process) and can then find the probability that the patient has heart disease. The probability scores are generated from a logistic regression formula that combines and weights the dichotomous information on the seven symptoms. When using the Heart Disease Predictive Instrument, physicians achieve more accurate decisions than when they rely on their intuitive judgment. Many doctors, however, are not happy using this and similar systems, typically because they do not understand logistic regression. Even though this understanding is not necessary to use the prediction systems, the lack of transparency and the dependence on probability charts leaves them uncomfortable.

There is, however, an alternative: Another view of heuristics is that they are cognitive strategies that can provide good solutions to complex problems under restrictions of time and cognitive capacity.

Cognitive Shortcuts

The concept of cognitive shortcuts providing good solutions is the opposite of the traditional view that human decision making should be evaluated in comparison with models of unbounded rationality, such as Bayesian or subjective expected utility models or logistic regression. These models must often assume—unrealistically—that people can predict all consequences of their choices, are able to assign them a joint probability distribution, and can order them using a single utility function. But in real life, people rarely have the time or cognitive capacity to think of all the possible scenarios for the future, their likelihood, and their subjective utilities. Real life often involves so many possible choices and so many possible outcomes that the optimal solution to a problem rarely exists, or if it does, the solution requires prohibitively long and complex computations. Instead of trying to find the best solution, people may *satisfice*—that is, look for solutions that are good enough for their current purposes. The father of this *bounded rationality* view, Herbert Simon, argued that people rely on simple strategies that can successfully deal with situations of sparse resources.

A recent representative of the bounded rationality approach is the simple heuristics research program. This approach, championed by Gerd Gigerenzer, Peter M. Todd, and the ABC Research Group, proposes that heuristics may be the only available approach to decision making for the many problems for which optimal solutions do not exist. Moreover, even when exact solutions do exist, domain-specific decision heuristics may be more effective than domain-general approaches, which are often computationally unfeasible. This research program focuses on precisely specified

computational models of fast and frugal heuristics and how they are matched to the ecological structure of particular decision environments. It also explores the ways that evolution may have achieved this match in human behavior.

In line with this approach, Lee Green and David R. Mehr constructed a simple heuristic for the patient admission process in a coronary unit. This heuristic, a fast and frugal decision tree, relies on simple building blocks for searching for information, stopping the information search, and finally making a decision. Specifically, it first ranks the predictors according to a simple criterion (the predictor with the highest sensitivity first, the predictor with the highest specificity second, and so on), and information search follows this order. Second, the search can stop after each predictor; the rest are ignored. Third, the strategy does not combine—weight and add—the predictors. Only one predictor determines each decision. This decision rule is an example of one-reason decision making.

The fast and frugal decision tree proposed by Green and Mehr works as follows: If a patient has a certain anomaly in his electrocardiogram, he is immediately admitted to the coronary care unit. No other information is searched for. If this is not the case, a second variable is considered: whether the patient's primary complaint is chest pain. If this is not the case, he is immediately classified as low risk and assigned to a regular nursing bed. No further information is considered. If the answer is yes, then a third and final question is asked: whether he has had a heart attack before. The fast and frugal decision tree thus ignores all 50 probabilities of the original Heart Disease Predictive Instrument and asks only a few yes-or-no questions.

The fast and frugal tree, just like the Heart Disease Predictive Instrument, can be evaluated on multiple performance criteria. Accuracy is one criterion, and it turns out that the tree is more accurate in classifying heart attack patients than both physicians' intuition and the Heart Disease Predictive Instrument. Specifically, it assigned correctly the largest proportion of patients who subsequently had a myocardial infarction to the coronary care unit. At the same time, it had a comparatively low false alarm rate. Being able to make a decision fast with only limited information is a second criterion, which is essential in situations where slow decision making can cost a life. The fast and frugal decision tree uses less information than the expert system and uses less sophisticated statistical calculations. A third criterion is the transparency of a decision system. Unlike logistic regression, the steps of the fast and frugal tree are transparent and easy to teach. Therefore, in complex situations such as the patient admission process in a coronary unit, less is more. Simplicity can pay off.

Another well-studied cognitive strategy within the simple heuristics program is take-the-best, which is a domain-specific rather than a general problem-solving strategy, meaning that it is useful in some environments and for some problems but not for all. When using this heuristic, people infer which of two objects has a higher value on some criterion based on just one reason, or cue. An example would be inferring which of the two cities has a higher mortality rate based on the average January temperature, the relative pollution potential, or the average percentage of relative humidity. Like the fast and frugal decision tree of Green and Mehr, this heuristic considers cues sequentially in the order of how indicative they are of the objects' values and makes a decision based on the first cue that discriminates between objects. This heuristic is particularly successful when one cue is much more important than other cues, but it has been shown to be as good as computationally more demanding procedures in other environments, as well.

Empirical evidence suggests that take-the-best is a plausible behavioral model, especially when searching for information in the environment is costly or when decisions have to be made under time pressure. By "betting on one good reason" and disregarding the surplus information, this fast and frugal heuristic may be particularly useful for patient populations whose computational resources are limited due to aging or illness.

Future Research

Medical situations promote the use of heuristics because decisions in such contexts often need to be made under pressure and with limited information. Fast and frugal heuristics for medical decision making have the potential to be powerful

alternatives to the prescriptions of classical decision theory for patient care. A strategy that ignores information and forgoes computation can be not only faster, more frugal, and transparent but also more accurate. Simple tools for making accurate decisions under time pressure in the medical arena should be a major research topic for future investigation.

Rocio Garcia-Retamero and Mirta Galesic

See also Bias; Bounded Rationality and Emotions; Decision Making and Affect; Motivation; Trust in Healthcare

Further Readings

Garcia-Retamero, R., Hoffrage, U., & Dieckmann, A. (2007). When one cue is not enough: Combining fast and frugal heuristics with compound cue processing. *Quarterly Journal of Experimental Psychology, 60,* 1197–1215.

Garcia-Retamero, R., Hoffrage, U., Dieckmann, A., & Ramos, M. (2007). Compound cue processing within the fast and frugal heuristic approach in non-linearly separable environments. *Learning & Motivation, 38,* 16–34.

Gigerenzer, G., & Goldstein, D. G. (1996). Reasoning the fast and frugal way: Models of bounded rationality. *Psychological Review, 103,* 650–669.

Gigerenzer, G., & Kurzenhäuser, S. (2005). Fast and frugal heuristics in medical decision making. In R. Bibace, J. D. Laird, K. L. Noller, & J. Valsiner (Eds.), *Science and medicine in dialogue: Thinking through particulars and universals* (pp. 3–15). Westport, CT: Praeger.

Gigerenzer, G., Todd, P. M., & the ABC Research Group. (1999). *Simple heuristics that make us smart*. New York: Oxford University Press.

Green, L., & Mehr, D. R. (1997). What alters physicians' decisions to admit to the coronary care unit? *Journal of Family Practice, 45,* 219–226.

Kahneman, D., Slovic, P., & Tversky, A. (1982). *Judgment under uncertainty: Heuristics and biases*. Cambridge, UK: Cambridge University Press.

Kahneman, D., & Tversky, A. (1973). On the psychology of prediction. *Psychological Review, 80,* 237–251.

Pozen, M. W., D'Agostino, R. B., Selker, H. P., Sytkowski, P. A., & Hood, W. B. (1984). A predictive instrument to improve coronary-care-unit admission practices in acute ischemic heart disease. *New England Journal of Medicine, 310,* 1273–1278.

Simon, H. A. (1983). Alternative visions of rationality. In H. A. Simon (Ed.), *Reason in human affairs* (pp. 7–35). Stanford, CA: Stanford University Press.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science, 185,* 1124–1131.

Holistic Measurement

Holistic measurement is an approach to the measurement of preferences for health states or treatments in which a rater assigns values to each possible health state or treatment, where a state or treatment represents a combination of many attributes. During the assessment, the rater thus considers all the relevant attributes simultaneously.

Valuing Health States or Treatments

Basically, there are two different approaches to measuring preferences for health states, services, or treatments: the holistic and the decomposed. The decomposed approach expresses the overall value as a decomposed function of the attributes. It enables the investigator to obtain values for all health states or treatments without requiring the rater to assign values to every state or treatment; the rater is asked to value the attributes only. Holistic measurement is mostly used for health state valuation, but in some instances, it is used for the valuation of treatments or services as well, for example, in the willingness-to-pay method and the treatment trade-off method.

Holistic valuations of health states encompass valuations of the quality of life of those states, and the valuations are therefore sometimes called preference-based measures of quality of life, as distinct from descriptive measures of quality of life. Descriptive measures of quality of life generally generate quality-of-life profiles, that is, a combination of scores on different dimensions of quality of life, such as physical functioning, emotional functioning, and social functioning. A well-known example of such a descriptive instrument is the Medical Outcomes Study SF-36. These descriptive approaches to quality-of-life evaluation are not suitable for the purpose of decision making. In decision making, different attributes of treatment

outcomes have to be weighed. On the one hand, different aspects of quality of life may have to be balanced against each other. Does, for example, the better pain relief from a new neuralgia medication outweigh the side effects, such as sedation and confusion? On the other hand, quality of life and length of life may have to be weighed against each other. Does the increased survival from chemotherapy outweigh the side effects, or, on the contrary, are patients willing to trade off survival for improved quality of life? For such decisional purposes, a valuation of the health outcome is needed.

Holistic Methods

Several holistic methods exist to assess preference-based measures of quality of life. The standard gamble and the time trade-off measure the utility of a health state, a cardinal measure of the strength of an individual's preference for particular outcomes when faced with uncertainty.

Standard Gamble

In the standard gamble method, a subject is offered the hypothetical choice between the sure outcome *A* (living his remaining life expectancy in the health state to be valued) and the gamble *B*. The gamble has a probability p of the best possible outcome (usually optimal health, defined as 1) and a probability $(1 - p)$ of the worst possible outcome (usually immediate death, defined as 0). By varying p, the value at which the subject is indifferent to the choice between the sure outcome and the gamble is obtained. The utility for the sure outcome, the state to be valued, is equal to the value of p at the point of indifference ($U = p \times 1 + (1 - p) \times 0 = p$).

Time Trade-Off

In the time trade-off method, a subject is asked to choose between his remaining life expectancy in the state to be valued and a shorter life span in normal health. In other words, he is asked whether he would be willing to trade years of his remaining life expectancy to avoid the state to be valued. As an example, a 65-year-old man is asked how many years x in a state of optimal health he considers equivalent to a period of 15 years (his remaining life expectancy) in a disability state. By varying the duration of x, the point is found where he is indifferent to the choice between the two options. The simplest and most common way to transform this optimal health equivalent x into a utility (ranging from 0 to 1) is to divide x by 15.

Visual Analog Scale

A visual analog scale is a rating scale, a simple method that can be self-administered and, therefore, is often used to obtain evaluations of health states. Subjects are asked to rate the state by placing a mark on a 100-mm horizontal or vertical line, anchored by optimal health and death (or sometimes best possible health and worst possible health). The score is the number of millimeters from the "death" anchor to the mark, divided by 100.

The visual analog scale does not reflect any trade-off that a subject may be willing to make in order to obtain better health, either in terms of risk or in years of life. It can therefore not be considered a preference-based method, and transformations have been proposed to approximate standard gamble or time trade-off utilities. The choice of the method is still a matter of an ongoing debate. All three methods have been shown to be subject to biases in the elicitation process, but many of these biases can be explained by prospect theory.

Magnitude Estimation

Magnitude estimation is a scaling method that was developed by psychophysicists to overcome the limitations of the rating scales, that is, the lack of ratio-level measurement and the tendency of respondents to use categories equally often (verbal scale) or not to use the upper and lower ends of the scale (visual analog scale). The respondent is given a standard health state and asked to provide a number or ratio indicating how much better or worse each of the other states is as compared with the standard. For example, the research participants are instructed to assign the number 10 to the first case, the standard. Then a case that is half as desirable receives the number 5, and a case that is regarded as twice as desirable is given the number 20. Magnitude estimation is seldom used, since it is not based on any theory of measurement and since the scores have no obvious meaning in the context of decision making. They do not reflect

utility and as such cannot be used in decision analyses.

Person Trade-Off

A different variant called the person trade-off has gained popularity among health economists and policy makers. It was formerly known as the equivalence method, and the task is to determine how many people in health state *X* are equivalent to a specified number of people in health state *Y*. From a policy perspective, the person trade-off seeks information similar to that required by policy makers. It has been used in the elicitation of disability weights for the DALYs (disability-adjusted life years), a measure used by the World Health Organization as a summary measure of population health.

Willingness-to-Pay

The willingness-to-pay is a method used primarily by health economists. To value health states, it asks the respondents what amount, or what percentage of their household income, they would be willing to pay to move from a less desirable state to a state of optimal health. More frequently, it is used to assess respondents' willingness to pay for treatments and services. It is most commonly used in cost-benefit analyses, in which all outcomes are expressed in monetary terms, in contrast to cost-effectiveness analyses, in which health outcomes are expressed in (quality-adjusted) life years. As is the case for magnitude estimation and the person trade-off method, this method does not result in a utility.

Probability Trade-Off

The probability trade-off or treatment trade-off method assesses, in a holistic manner, respondents' strength of preference for a treatment (relative to another treatment). In these methods, preferences for combined process and outcome paths are elicited in the following way. The patient is presented with two clinical options, for example, Treatments A and B, which are described with respect to (chances of) benefits and side effects, and is asked to state a preference for a treatment. If Treatment A is preferred, the interviewer systematically either increases the probability of benefit from Treatment B, or reduces the probability of benefit from Treatment A (and vice versa if Treatment B is preferred). The particular aspects of the treatments that are altered in this way, and the direction in which they are changed, are decided on beforehand, according to the clinical characteristics of the problem and the nature of the research question. For example, these may include the probability of side effects of treatment, risk of recurrence, or chance of survival. The relative strength of preference for a treatment is assessed by determining the patient's willingness to accept side effects of that treatment or forego benefits of the alternative treatment. This general approach has been adapted specifically to a variety of treatment decisions. Examples are decisions about adjuvant chemotherapy in breast cancer, benign prostatic hypertrophy, treatment of lupus nephritis, and radiotherapy for breast cancer.

The resulting preference scores are idiosyncratic to the original decision problem, and only the strength of preference for Treatment A relative to Treatment B is obtained, not a utility. For formal decision analysis they are therefore not suitable. However, for decision support they seem appropriate as they are tailored to the clinical problem at hand and will reflect the real-life situation more than does utility assessment. These methods have indeed been used "at the bedside," using decision boards as visual aids. They seem a promising way to help patients who wish to engage in decision making to clarify and communicate their values.

Anne M. Stiggelbout

See also Contingent Valuation; Decomposed Measurement; Person Trade-Off; Prospect Theory; Utility Assessment Techniques; Willingness to Pay

Further Readings

Bleichrodt, H. (2002). A new explanation for the difference between time trade-off utilities and standard gamble utilities. *Health Economics, 11,* 447–456.

Llewellyn-Thomas, H. A. (1997). Investigating patients' preferences for different treatment options. *Canadian Journal of Nursing Research, 29,* 45–64.

Stiggelbout, A. M., & De Haes, J. C. J. M. (2001). Patient preference for cancer therapy: An overview of measurement approaches. *Journal of Clinical Oncology, 19,* 220–230.

Human Capital Approach

The human capital approach to economic evaluation places a monetary value on loss of health as the lost value of economic productivity due to ill health, disability, or premature mortality. More specifically, the human capital approach uses the present value of expected future earnings, often adjusted for nonmarket productivity, to estimate the potential loss to society if an individual dies or becomes permanently disabled. It is commonly employed in cost-of-illness (COI) analyses that distinguish between direct costs, chiefly medical care, and the indirect costs of lost productivity. It is also employed in certain cost-effectiveness and cost-benefit analyses, particularly in older publications.

The idea that a human life can be valued by capitalizing the value of future earnings goes back to Sir William Petty in England in the late 1600s. The application of human capital to economic evaluation of health interventions can be traced to Burton Weisbrod in the 1960s. Under this approach, productivity is calculated as the present value of the sum of expected labor market earnings in future years, adjusted for life table survival probabilities and discounting. It is standard practice to take the current pattern of average earnings stratified by age and sex and assume that an individual's earnings trajectory will trace the same pattern, adjusted for expected increases in future labor productivity and inflation-adjusted earnings. For example, in the United States, it is conventional to assume that future labor productivity will increase at 1% per year. If one combines this with a 3% discount rate as is recommended in the United States, one gets estimates roughly equivalent to use of a 2% discount rate without assuming future productivity increases.

It is standard in health economic evaluations to include the imputed value of household production as well as paid earnings in human capital estimates, although cost-benefit analyses in environmental policy typically do not do so. The inclusion of household productivity is particularly important for older people and women, who tend to have high values of household productivity relative to paid compensation. Time spent in household production can be valued using either the individual's own wage or imputed wage (opportunity cost method) or the average wage paid to workers performing similar services (replacement cost method); the latter is more commonly employed. The original justification for the inclusion of household services was to reduce the lower valuation placed on women's lives because of lower labor force participation. Although in principle one could put a monetary value on other uses of time, such as volunteer service and leisure, this is rarely done.

Earnings are typically calculated as gross earnings, including payroll taxes and employee benefits, and are intended to capture the full cost of employee compensation. The rationale for the human capital approach is that the marginal productivity of labor is equal to the compensation paid to the average employee and that the withdrawal of that individual's labor due to premature death or permanent disability would result in a loss to society of that individual's future production. In most applications, average earnings are estimated for everyone within a given age-group, stratified only by gender. This avoids the ethical problems that can result from using different earnings for individuals of different socioeconomic or ethnic groups, which can have the effect of causing diseases affecting disadvantaged groups to appear less costly. The same argument can be applied to the use of sex-specific earnings estimates, given that in almost all countries average earnings are lower for women than for men, even after taking household services into account.

The chief alternative to the human capital approach is the friction cost approach developed by Dutch economists in the 1990s who objected that the presence of unemployed labor made human capital estimates of productivity losses too high. This approach presumes that replacement workers are readily available and that the only loss in productivity due to a worker's death or disability is the short-term cost of recruiting and training a replacement worker. Human capital estimates of productivity losses are many times higher than those calculated using the friction cost method.

Older cost-benefit analyses typically used the human capital approach to put a monetary value on lost life. In recent decades, it has become standard to use estimates of willingness-to-pay (WTP) to value health, particularly in environmental and transportation policy analyses. WTP estimates of the value of a statistical life based on occupational mortality and compensating wage differentials are

typically several times higher than human capital estimates. Unlike WTP estimates, human capital estimates do not place a monetary value on pain and suffering or the grief experienced by family members and friends at the loss of a loved one. Because of the difficulty in putting a monetary value on such intangible costs, cost-benefit analyses published in medical journals often use the human capital approach for monetary valuations of health, which results in relatively conservative estimates of benefits as compared with cost-benefit analyses using WTP estimates.

Older cost-effectiveness analyses also often included productivity costs. However, the U.S. Panel on Cost-Effectiveness in Health and Medicine in 1996 recommended that reference case cost-effectiveness analyses conducted from the societal perspective only include direct costs of care and exclude productivity costs, a term that was suggested to supplant the term *indirect costs*. The rationale offered was that quality-adjusted life-years, or QALYs, recommended as a measure of health outcomes could entail double counting with economic productivity. The National Institute of Health and Clinical Excellence (NICE) in the United Kingdom likewise recommends that only direct costs be included in cost-effectiveness analyses.

The leading use of human capital estimates is in COI studies used to call the attention of stakeholders to the economic impact of diseases or injuries and the potential gains from allocating funds to research and prevention, but they can also be used in the economic evaluation of programs or interventions. For example, the economic benefit of folic acid fortification policies for the prevention of certain types of birth defects, spina bifida and anencephaly, has been calculated as the present value of lifetime earnings for averted cases of anencephaly, which is uniformly fatal in the neonatal period, and the present value of averted medical and educational costs and gains in economic productivity from the prevention of lifelong disability and early mortality resulting from averted cases of spina bifida.

Scott D. Grosse

Disclaimer: The findings and conclusions in this report are those of the author and do not necessarily represent the official position of the Centers for Disease Control and Prevention.

See also Cost-Benefit Analysis; Costs, Direct Versus Indirect

Further Readings

Grosse, S. D., Waitzman, N. J., Romano, P. S., & Mulinare, J. (2005). Reevaluating the benefits of folic acid fortification in the United States: Economic analysis, regulation, and public health. *American Journal of Public Health, 95,* 1917–1922.

Luce, B. R., Manning, W. G., Siegel, J. E., & Lipscomb, J. (1996). Estimating costs in cost-effectiveness analysis. In M. R. Gold, J. E. Siegel, L. B. Russell, & M. C. Weinstein (Eds.), *Cost-effectiveness in health and medicine* (pp. 176–213). New York: Oxford University Press.

Max, W., Rice, D. P., & MacKenzie, E. J. (1990). The lifetime cost of injury. *Inquiry, 27,* 332–343.

Verstappen, S. M., Boonen, A., Verkleij, H., Bijlsma, J. W., Buskens, E., & Jacobs, J. W. (2005). Productivity costs among patients with rheumatoid arthritis: The influence of methods and sources to value loss of productivity. *Annals of the Rheumatic Diseases, 64,* 1754–1760.

Waitzman, N. J., Scheffler, R. M., & Romano, P. S. (1996). *The cost of birth defects*. Lanham, MD: University Press of America.

Human Cognitive Systems

Human cognitive systems are the systems in the human mind that involve the conscious processing of information of various types and that help individuals deal with self, others, and the world. Human cognitive systems as they reside in the human mind are typically described as being mental processes that are typically accessible by only the individual. As the individual behaves in the world and as that individual communicates with others, the individual can share to some degree (but does not have to) what is going on in his or her own mind.

The processes underlying human cognitive systems, such as thinking and deciding, for the most part, are not held in consciousness but remain as unconscious or subconscious processes.

Referent of the Term *Cognition*

The term *cognition* can be used to refer to the *processes* of thought, the process of thinking, the

applying of rules, the development of plans in humans, the weighing of risk and benefit, or the performances of operations such as mathematical operations, or to the *results of such processes* in humans, animals, and machines. The term can apply to the processes of thinking, deciding, and perceiving, or to the results of such cognitive activity. The term *cognition* can apply to beliefs of as well as knowledge of individuals, groups, or populations.

The term may also apply to some views of perceiving (perception) but not necessarily to certain views of sensing (sensation). The construction of sensations into perceptions may be considered by some to be a cognitive process even though such construction or processing can occur unconsciously or subconsciously. And these perceptions often occur instantaneously, as when we see a tree and we see (perceive) this tree as having a back side, even though we do not see that back side in our perception. Yet if we walk around that tree, the argument continues, we would be surprised to find that it did not have a back side and was just, for example, an elaborately constructed stage prop and not a tree at all.

Preferences and Decision Making

Cognitive science in the late 1970s to early 2000s focused on the notion of mind and intelligence in terms of representation and computational procedures of human intelligence versus machine intelligence (artificial intelligence), with cognitive research on medical artificial intelligence, artificial intelligence in choice and explanation, artificial neural networks, prediction of medical conditions and disease processes such as community-acquired pneumonia, and computer-based explanations of decision theoretic advice, among others. Yet cognitive science of the late 1970s to early 2000s also needs to be recast in terms of its definition and needs and seen as taking on different dimensions than the cognitive science of the 1950s. Yet both domains still share crucial similarities.

Today, the concept of cognitive science goes beyond this notion. The cognitive sciences today, particularly as they apply to medical decision making, also explore the notions of patient preferences, how patients make decisions on their own (descriptive decision making), and how this descriptive decision making compares with other models of how decisions should be made (normative decision making). In terms of normative decision making, human preference in medical choice situations is compared with normative models of decision making, such as expected value theory. Today, it can be argued that emotion (and emotive theory) also has a role to play in the cognitive sciences.

However, today, the ways in which humans think, problem solve, and weigh decisions in terms of output of human mental processes (human intelligence) are not compared with machine intelligence (artificial intelligence) but rather are contrasted to the outputs of alternative approaches of how decisions should be made (normative theory, such as expected value theory) or alternative theories regarding how decisions are actually made by humans (psychological theory, such as prospect theory).

In such later comparisons a different form of cognitive science arises, one that—in essential qualities of comparison—is not that dissimilar in terms of methodology from the attempt of the earlier view of cognitive science to capture human mental output and problem-solving skills in terms of representations and computational procedures and then compare how human intelligence compares and contrasts with artificial intelligence.

Framing and Choice

Amos Tversky and Daniel Kahneman describe their work in choice and decision making in relation to the basic model of mind given above. When the authors talk about decision making, they use the term *decision frame* to refer to the decision maker's conception of what he or she considers—consciously, unconsciously, or subconsciously—as the acts, outcomes, and contingencies associated with a particular choice. They further note that this frame adopted by the decision maker is influenced by the way the problem is formulated and partly by the decision maker's own norms, habits, and personal characteristics.

Tversky and Kahneman compare their perspective on alternative frames for a decision problem to perception, particularly, the alternative perspectives on a visual scene. The term *veridical* means coinciding with reality. Tversky and Kahneman note that veridical perception requires that the

"perceived relative heights" of two neighboring mountains should not reverse with any change of vantage point of an observer. In a similar vein, the authors argue that rational choice requires that the preference between options should not reverse with changes of frame or perspective. They then link the imperfections of perception in the human being with the imperfections of human decision making. For the authors, changes of perspective, in fact, often reverse (or at least in some way influence) the relative apparent size of objects, and changes of perspective, in fact, often reverse (or at least influence in some way) the relative desirability of options.

Tversky and Kahneman thus characterize their own research in framing as an attempt to describe and thus represent the human cognitive system in terms of the systematic reversals of preferences. Their own research discovered that by varying the framing of acts, their contingencies, and their outcomes, human cognitive systems of study volunteers respond to such changes (descriptive decision making) in ways that are not predicted by normative models of decision making, such as expected utility theory.

Language Learning and Problem Solving

Just as Tversky and Kahneman consider their framing effects as almost akin to human perceptual effects, there are many basic questions raised by the application of framing effects into medical decision making that raise issues in other areas of human cognitive systems: namely, the capacity for language and problem solving.

Tversky and Kahneman's research, based on their concept of framing, focused on one type of study design methodology: choices between "a simple gamble" or "a sure thing," each with an objectively specified probability and at most two nonzero outcomes. Typical choice situations were based on monetary gains and losses or survival (health-related) gains and losses.

Tversky and Kahneman were interested in "gain" and "loss" situations. In a gain situation, a typical monetary choice given by Tversky and Kahneman to study participants was a sure gain of $250 versus a 25% chance to gain $1,000 and a 75% chance of gaining nothing. A typical monetary choice in a loss situation involved consideration of a sure loss of $250 versus a 25% chance of losing $1,000 and a 75% chance of losing nothing.

A typical survival-mortality choice situation given by Tversky and Kahneman to study participants for consideration involved a disease outbreak where the overall baseline situation is a disease outbreak with an overall expectation of 600 people being killed, but there is a gain scenario and a loss scenario. The gain scenario is illustrated by the choice given between two programs, Vaccine Program A, which if adopted would allow 200 people to be saved (+200 being saved), versus Vaccine Program B, which if adopted would result in a one-third chance of 600 people being saved (+200 being saved) and a two-thirds chance of no one being saved. In terms of loss, a typical survival-mortality choice situation that could have been given by Tversky and Kahneman to study participants for consideration could have involved a new vaccine first to be used in a human population that has an as yet unknown defect. This flawed vaccine—instead of saving lives—will cause more people to be killed until the fact that the vaccine is flawed is identified. The choice given is between two programs using the flawed vaccine, Flawed Vaccine Program C, which if adopted would allow 200 additional people to be killed of an at-risk group that the vaccine was intended to save (200 individuals being killed), versus Flawed Vaccine Program D, which if adopted would result in a one-third chance of 600 at-risk people being killed (200 individuals being killed) and a two-thirds chance of no one being killed.

Part of the challenges faced by the study volunteers in each of the above thought experiments is the extent to which the study participants could imagine the gains and losses to be real considerations in their own lives while answering the questionnaires in a study setting. This challenge also related to Tversky and Kahneman's methodology, specifically relating to the fact that it may be difficult to invent choice scenarios in survival-mortality contexts that involve gains and losses, as in the flawed vaccine example, to make the scenario believable enough that the study participants can place themselves in the scenario as the individual targeted to make the decision.

In general, it is also important for study investigators to recognize that some study participants are unwilling to agree to participation in questionnaire

studies that require them to consider gambles in relation to human life. Such individuals hold personal beliefs that will not allow them to participate in such studies, and thus, these individuals' opinions will not be reflected in the study results of such research endeavors in decision making.

The internal constraints—study of one type of choice between "a simple gamble" or "a sure thing," each with an objectively specified probability and at most two nonzero outcomes—explicitly placed by these two psychologists on their rigorous study methodology (with the aim of understanding what types of choices humans make in specifically defined choice situations involving gains and losses and considerations of gambles vs. sure things) allowed them to construct a Nobel prize–winning theory that was highly dependent on at least two capacities of human cognitive systems in addition to perception: language learning and problem solving.

Literacy and Numeracy

Here, language learning has two components: the ability to work with words (*literacy*) and the ability to work with numbers (*numeracy*). The problem that arises for Tversky and Kahneman is what happens when the citizens of a population have difficulties with literacy, numeracy, or both. An additional problem arises when the citizens of a population prefer to discuss issues such as are found in medical decision making in terms of quality expressions of chance (probability), that is, in terms of words and not in terms of numbers. David H. Hickam and Dennis Mazur have found that in discussion of risk in medicine, patients prefer to discuss risk with their physicians in terms of qualitative expressions of chance (probability), such as "rare," "possible," and "probable," and not in terms of numerical expressions of chance (probability), such as "percents."

The very understanding of medical decision making has been argued to require a high propensity for verbal and numerical abilities for those patients and consumers interested in participating in considering scientific evidence derived from research studies as part of their own decision making in relation to shared decision making between patients and their providers, or in terms of physician-based decision making where the patient may not want to make a decision on care but does want to track those decisions.

Yet there are many reasons to believe that solutions to these basic issues of literacy and numeracy in a population are not easy to move in positive directions beyond the research finding that a majority of patients do not want to discuss risk with their physicians in terms of numbers (quantitative expressions of probability) and prefer to discuss risk in terms of words (qualitative expressions of probability). This entry now considers an alternative to simple numbers: graphical data displays, for example, pie charts.

To date, graphs have been forbidden in the United States by federal regulation in direct-to-consumer advertisements used to sell medical products (prescription medicine and medical devices) to consumers through broadcast advertising over television because of the ways data can be unfairly manipulated in graphic data displays.

Jessica S. Ancker and colleagues note both the positive and negative impacts of graphs used to display data in medical and public health decision making. First, an example of a positive aspect of a graph is that it may allow patients and consumers to understand more clearly through a visual graphical display part-to-whole relationships. In helping patients understand data and scientific evidence in medicine, graphical data displays can help patients visually attend to key components of chance (probability) that can be expressed as a ratio of two numbers (e.g., a fraction with a numerator over a denominator). This fraction can also be expressed by a graph visually expressing the relationship between a numerator (the number of people sustaining adverse outcomes) and the denominator (the entire population studied), for example, in a pie diagram.

Second, graphs may be used to manipulate the very same numbers, for example, when the developer of the graph elects to display only the numerator in a graph used in a direct-to-consumer advertisement. A graph that displays only the numerator can be intentionally used to appear to inflate the perceived risk and, thus, induce risk-averse behavior on the part of the consumer. The *U.S. Code of Federal Regulations* attempts to guard against such manipulation by product manufacturers and their advertisers in the federal

regulation of direct-to-consumer advertising of prescription medicines.

Challenges facing direct-to-consumer advertising of prescription medicines and medical devices in cases where consumers and patients are not attending to numbers or graphs include the attention paid by (and the weight given to information by) consumers and patients to nonnumerical and nongraphical information such as *who* is endorsing the medical product. Directing consumers away from a full understanding of the numbers and scientific evidence that surrounds a medical product is often a key goal of a financially successful direct-to-consumer advertising program.

Dennis J. Mazur

See also Decision Psychology; Prospect Theory; Unreliability of Memory

Further Readings

Ancker, J. S., Senathirajah, Y., Kukafka, R., & Starren, J. B. (2006). Design features of graphs in health risk communication: A systematic review. *Journal of the American Medical Informatics Association, 13,* 608–618.

Baron, J. (2000). *Thinking and deciding*. New York: Cambridge University Press.

Fagerlin, A., Zikmund-Fisher, B. J., Ubel, P. A., Jankovic, A., Derry, H. A., & Smith, D. M. (2007). Measuring numeracy without a math test: Development of the subjective numeracy scale. *Medical Decision Making, 27,* 672–680.

Føllesdal, D. (1969). Husserl's notion of noema. *Journal of Philosophy, 66,* 680–687.

Lipkus, I. M., Samsa, G., & Rimer, B. K. (2001). General performance on a numeracy scale among highly educated samples. *Medical Decision Making, 21,* 37–44.

Mazur, D. J., & Hickam, D. H. (1997). Patients' preferences for risk disclosure and role in decision making for invasive medical procedures. *Journal of General Internal Medicine, 12,* 114–117.

Seebohm, T. M., Føllesdal, D., & Mohanty, J. N. (Eds.). (1991). *Phenomenology and the formal sciences*. Dordrecht, The Netherlands: Kluwer Academic.

Smith, D. W., & McIntyre, R. (1982). *Husserl and intentionality: A study of mind, meaning and language*. Dordrecht, the Netherlands: D. Reidel.

Tversky, A., & Kahneman, D. (1981). The framing of decisions and the psychology of choice. *Science, 211,* 453–458.

Woloshin, S., Schwartz, L. M., Moncur, M., Gabriel, S., & Tosteson, A. N. A. (2001). Assessing values for health: Numeracy matters. *Medical Decision Making, 21,* 380–388.

Hypothesis Testing

A scientific hypothesis is tested by evaluating the logical consistency of its implications and/or the accuracy of its predictions. Other grounds for assessing hypotheses include breadth of prediction, scientific fertility, simplicity, and aesthetic appeal; however, the term *hypothesis testing* refers only to accuracy. Statistical hypothesis testing, a form of inductive inference, is used extensively in medical research and described here as a form of proof by contradiction.

A hypothesis is rejected by a test if the hypothesis logically implies something false or strongly predicts something contradicted by data. The 2,500-year-old proof by Hippasus of Metapontum that $\sqrt{2}$ is not a ratio of whole numbers exemplifies the former. Hypothesizing the opposite, that $\sqrt{2} = a/b$ for whole numbers a and b, Hippasus deduced the impossible: that both numerator and denominator must remain divisible by 2, even after all possible cancellations of 2 from both a and b. Unable to deny the logic of this contradiction, a rational mind instead rejects the hypothesis, concluding that $\sqrt{2}$ *cannot be* such a ratio. This is proof by contradiction or, from the Latin, *reductio ad absurdum*.

Data may also contradict hypotheses. In deterministic settings, that is, when predictions are made with certainty because all relevant influences are presumed known, one valid incompatible datum overturns a hypothesis. The hypothesis "Elixir A cures all cancer" is overturned by a single treatment failure, demonstrating conclusively that other treatment is sometimes required. This is an empirical analog of proof by contradiction.

In medical sciences, though, knowledge is incomplete and biological variability the rule. Hence, determinism is rare. Medical hypotheses describe tendencies that are exhibited variably, in

complex systems governed by probabilities rather than individually predictable fates. The hypothesis "Elixir A increases the fraction of cases alive two months postdiagnosis" does not imply that a particular individual will live for 2 months. Unless 2-month survival is already extremely high, this hypothesis cannot be overturned by one or even several early deaths.

But suppose, in a trial of Elixir A, that all 10 clinically similar but otherwise unrelated patients who receive it die before 2 months postdiagnosis. If extensive data show that only half of similar untreated cases die this quickly, most would reconsider further use of Elixir A. Although 10 patients on any new treatment may all be unlucky, the chance of this happening in a specified study is below 2^{-10} = .098% if Elixir A is beneficial. Logically, either luck has been extraordinarily poor, or Elixir A doesn't work as hypothesized. A longer consecutive run of deaths would be even less likely, for example, 2^{-15} = .003% for 15 deaths, and hence more difficult to attribute to bad luck. With "enough" accumulated evidence, most persons bow to its weight and reject the initial hypothesis, because the hypothesis made a strong prediction that failed. Specifically, the hypothesis had predicted strongly, that is, with very high probability, that data would show a 2-month case fatality more similar to what the hypothesis describes (< 50%) than to what was actually seen (100%).

Such probabilistic proof by contradiction exemplifies statistical hypothesis testing, the focus of this entry. Statistical hypothesis tests influence most research on which medical decisions are based. Their general use is to select, from among statistical associations in data, those hardest to explain by play of chance in a particular data sample. The selected associations, unless explicable by study design problems, receive preferential evaluation for causal involvement in disease initiation, promotion, progression to disability, and therapeutic benefit.

Statistical Hypotheses, Distributions, and Evidentiary Standards

Statistical hypothesis testing presupposes a scientific hypothesis of interest, H, and a source of relevant data, for example, clinical or laboratory experiment, observational epidemiological study, or clinical database. The notation H_C is used for the complement, or negation, of H. Testing constitutes a formal confrontation of a prediction from either H or H_C with the data. This involves several steps, starting with a choice to test either H or H_C and selection of a basic probability model for the data-generating process. The probability model consists of a collection of probability laws assumed to include one which accurately portrays this process. These laws are usually constructed from component probability distribution functions, for example, binomial, Poisson, normal (Gaussian), or lognormal distributions, thought to describe the origins of individual observations or sets of observations from a patient. This class is then partitioned into two subsets of probability laws, respectively, consistent with H and with H_C. From this partition follows a *statistical hypothesis* H_0, postulating that the data arise from a member of the subset associated with whichever scientific hypothesis, H or H_C, was chosen for testing.

Predictions about data can then be based, when the scientific hypothesis selected for testing is correct, on one or more probability laws from the subset associated with H_0. Along the lines of "the enemy of my enemy is my friend," data discrepant with predictions from H_0 contradict the scientific hypothesis H or H_C on which H_0 is based, supporting the other. Note that the sometimes important distinction between testing a scientific hypothesis H using predictions from H_0, and testing whether H_0 contains an accurate probability model for the data, is often dropped in application. For simplicity and brevity, we must also sometimes drop it below.

In the 10-patient trial above, one hypothesized H, that a cancer patient treated rapidly with Elixir A stood more than a 50% chance of surviving at least 2 months after diagnosis, rather than H_C, a 50% chance or less. Based on the clinical similarity and presumably independent results of otherwise unrelated patients, the probability model consists of all binomial distributions Bin(10, π), with π the chance a patient survives 2 months, and Bin(n, π) the mathematically proven probability law describing a count of accumulated events from n independent tries of the same process, each with chance π of producing the event. These distributions with $\pi > 50\%$ reflect H and with $\pi \leq 50\%$ reflect H_C. H_0, based on H, hypothesizes that the data arise from

one of the former group. H_0 is said to be a "composite" versus "simple" hypothesis because it contains more than one distribution. Any Bin(10, π) in H_0 predicts some 2-month survivors with probability exceeding 99.9%. Total absence of survivors disconfirms this strong prediction based on H and supports H_C, that Elixir A is ineffective or even harmful.

Scientifically, one usually hopes to demonstrate presence rather than absence of a relationship. Thus, when H posits a relationship, H_0 is usually based on H_C, hypothesizing its absence or opposite. H_0 is then called the *statistical null hypothesis*, motivating the conventional subscripted 0. The researcher wishes to assemble enough data to contradict H_0, discredit H_C, and hence confirm H.

Suppose H_0 has been chosen based on some scientific hypothesis H. A method is then selected for locating the observed data on a scale of discrepancy from H toward H_C, in relation to other possible study results that might have but did not occur. The scale is defined by the value of a summary statistic, for example, a count (as above), proportion, mean, difference, or ratio of these or the maximum probability of the observed data for a probability law from H_0. Any such scale defines a possible hypothesis test. The scale ordinarily incorporates whether testing is "one sided" or "two sided." For instance, in the one-sided example above, 2/10 = 20% early deaths are less discrepant from H_0 ($\pi > 50\%$ for survival) than are 6/10 = 60% early deaths, but would be more discrepant on a reasonable scale for two-sided testing of H_0^*: Elixir A has no effect (H_0: $\pi = 50\%$).

Research results sufficiently discrepant to reject H_0 are determined, using the selected discrepancy scale, by designating a maximum allowable chance of erroneous rejections when H_0 applies. This probability, symbolized by α, is called the *significance level* or simply *level* of the test. The collection of results discrepant enough to reject is then formed by successively including possible results, from most discrepant toward less discrepant from H, as ordered by the summary statistic. The process stops when rejections based on the next value of the summary statistic would raise the accumulated rejection probability above α for some distribution in H_0.

The level α serves as a probabilistic standard for strength of evidence required to reject H_0. For tests based on continuous probability distributions, α is the chance that a true hypothesis will be erroneously rejected and otherwise is an upper bound. In practice, α is often chosen from among 10%, 5%, 1%, and 0.1%; lower values require stronger evidence to reject H_0. The Neyman-Pearson approach to statistical hypothesis testing, invoked explicitly by considerations of statistical power, also requires the specification of an *alternative hypothesis* H_A that comprises probability laws possibly applying when H_0 is false.

The steps above prescribe how to test H_0 using any data that occurs. The summary statistic and its location on the discrepancy scale are determined. This location indicates whether the evidentiary standard for rejection has been met and hence whether the test is passed or failed. If the latter, H_0 is rejected, and the effect, trend, or "signal" in the data is labeled "statistically significant." If the former, some say H_0 is accepted, others that it is retained, that the test fails to reject, or that the signal is not statistically significant. *Accepted* is a technical term in this context; literal acceptance is rarely if ever justified in medical hypothesis testing.

Technical Aspects and Examples

Most generally, a statistical hypothesis test may be viewed as partitioning the universe U of possible data sets that might be observed in a study into a *rejection region* U_R and its complement U_C. U_R is chosen so that, when H_0 is true, the maximum probability of observing data within U_R, called the *size* of the test, equals or is minimally below α. Among possible such choices of U_R, a region is selected with high chance of containing data likely to occur from the type of relationship one expects, and hopes to detect, if H_0 is false.

Implementation involves rejecting H_0, or not, based on the location of a summary *test statistic* within a reference probability distribution. The test of a genuinely null hypothesis, that is, one representing no difference between quantities such as averages or proportions over time or between groups, is then analogous to a clinical diagnostic test. In diagnosis, the detection target is disease in a patient; in statistical hypothesis testing, the target is a systematic statistical relationship in the population generating the data. The hypothesis test may mistakenly reject H_0 when true, a Type I error, or

mistakenly retain H_0 when false, a Type II error. These errors are respectively analogous to false-positive and false-negative diagnoses. The Type I error probability is the size of the test, which for simplicity we now assume equals the stipulated level α. Its complement, the chance $1 - \alpha$ that H_0 passes the test when true, is analogous to diagnostic specificity. Type II error probability is represented by β. Its complement $1 - \beta$, the chance of rejecting a false H_0, is the test's *power,* analogous to diagnostic sensitivity. Power and β are functions of the extent to which reality departs from H_0. Tables 1 and 2 show the analogy. In diagnosis, one desires highest possible sensitivity for given false-positive rate. In hypothesis testing, one desires highest possible power for given test level α.

Suppose one conducts a parallel group randomized trial comparing a new drug with placebo and wishes to demonstrate differences in mean diastolic blood pressures (DBP) and proportions of patients who experienced a myocardial infarction (MI) after 1 year. For DBP, blood pressures might be assumed normally distributed, and H_0 might state that DBPs within each group have identical distributions. The ubiquitous Student's t statistic is the ratio of the difference between mean DBPs among patients receiving the new drug and receiving placebo (essentially, the signal in the data), to its estimated standard error, a measure of sensitivity to random variation among patients, that is, statistical noise. U_R contains data sets for which this ratio differs from 0 more than would occur $100(1 - (\alpha/2))\%$ of the time if H_0 were true, as calculated from a Student's t distribution, the relevant reference probability law.

For MI, H_0 states that probability of MI within a year is unaffected by drug. An equivalent version of the well-known Pearson chi-square test uses the ratio of the difference between proportions of patients who experienced an MI to its estimated standard error when H_0 is true. U_R is determined as above, but using a different reference distribution.

Table 1 Diagnostic test probabilities

Disease	*Test Result*	
	Negative	*Positive*
Absent	Specificity	False-positive rate
Present	False-negative rate	Sensitivity

Table 2 Hypothesis test probabilities

Null Hypothesis	*Test Result*	
	Retain	*Reject*
True	$1 - \alpha$	Type I error: α
False	Type II error: β	Power: $1 - \beta$

Use of unsigned differences and $\alpha/2$, as above, pertain to two-sided tests. In one-sided testing, U_R contains only data sets reflecting the anticipated direction, up to accumulated probability $100(1 - \alpha)\%$. In the examples, only differences of prespecified sign would justify rejection. Since rejection in one direction is precluded, a one-sided test allows easier rejection of H_0 in the other, anticipated direction, than does a two-sided test if both are at level α.

Extensive theory guides selection of a hypothesis test to use information in data from a given scientific setting most efficiently, stemming from work of J. Neyman and E. S. Pearson on testing with stipulated or optimum power against a specified alternative hypothesis H_A. Many methods for constructing U_R have been developed. Likelihood ratio testing orders data sets for placement into U_R by the ratio, lowest to highest, of the highest probability of the data under a distribution in H_0 to their highest probability under a stipulated broader class of distributions (e.g., H_0 or H_A, if an H_A has been specified). Other methods, such as score and Wald tests, also use test statistics calculated from assumed or approximated underlying probability laws. Reference probability distributions are also derived from a priori mathematical models for data generation processes or from theoretical approximations to the random behavior of summary statistics from large samples.

Sometimes a reference distribution may be developed based on a verifiable property of the data collection process or on symmetry considerations in a related "thought experiment." Thus, when treatment assignments are randomized, a reference distribution may be obtained by considering how a test statistic would vary across all possible randomized assignments. Such randomization tests have

high credibility due to conceptual simplicity and freedom from mathematical assumptions.

Information is lost when tests are reported as statistically significant, or not, at a fixed level. If statistical significance is reported at 5%, researchers with a more stringent evidentiary standard, say 1%, are left ignorant as to whether their standard has been achieved. If nonsignificance at 5% is reported, other researchers with a less stringent evidentiary standard, say 10%, are left similarly ignorant. Most researchers therefore report the *p* value of a test, defined as the lowest level for which the test is statistically significant, allowing each reader to assess statistical significance relative to an individual standard. Modern software automates reporting of *p* values.

Some take this further, by omitting prespecification of α and U_R altogether and using the *p* value as an index of compatibility of the data with H_0, on the discrepancy scale that underlies the hypothesis test. Values close to 0 reflect discrepancy; values above 0.1 reflect compatibility, increasingly with the value. Such use reasonably accords with the views of founders of biometrics, including K. Pearson, W. S. Gosset, and R. A. Fisher, but is in some respects incompatible with the Neyman-Pearson theory that followed. Thus, somewhat different testing philosophies coexist in general scientific practice.

This discussion has focused on the frequentist-based hypothesis tests that dominate the current biomedical literature. A Bayesian inferential perspective offers useful alternatives by treating both data and hypotheses as subject to probability distributions and incorporating a priori probabilities of hypotheses. Due to their subjectivity, Bayesian hypothesis tests have not been widely accepted in scientific practice. The increasing capabilities of "objective Bayes" methods, emphasizing prior distributions that limit effects of subjectivity, may overcome resistance.

Peter B. Imrey

See also Bayesian Analysis; Coincidence; Confidence Intervals; Effect Size; Frequentist Approach; Likelihood Ratio; Managing Variability and Uncertainty; Probability; Sample Size and Power; Statistical Testing: Overview

Further Readings

Barnett, V. (1999). *Comparative statistical inference* (3rd ed.). New York: Wiley.

Carlin, B. P., & Louis, T. A. (2008). *Bayesian methods for data analysis* (3rd ed.). New York: Chapman & Hall.

Cox, D. R. (1977). The role of significance tests. *Scandinavian Journal of Statistics, 4,* 49–70.

Edgington, E. S. (1995). *Randomization tests* (3rd ed.). New York: Dekker.

Fisher, R. A. (1959). *Statistical methods and scientific inference*. Edinburgh, UK: Oliver & Boyd.

Fisher, R. A. (1970). *Statistical methods for research workers* (14th ed.). New York: Hafner.

Lehmann, E. L., & Romano, J. P. (2008). *Testing statistical hypotheses*. New York: Springer.

Mayo, D. G., & Cox, D. R. (2006). Frequentist statistics as a theory of inductive inference. In J. Rojo (Ed.), *Optimality: The second Erich L. Lehman symposium* (pp. 77–97), May 19–22, 2004, Rice University, Houston, TX; Beachwood, OH: Institute of Mathematical Statistics.

Neyman, J., & Pearson, E. S. (1966). *Joint statistical papers of J. Neyman and E. S. Pearson*. Berkeley: University of California Press.

Pearson, K. (1948). On a criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can reasonably be supposed to have arisen from random sampling. In E. S. Pearson (Ed.), *Karl Pearson's early papers*. Cambridge, UK: Cambridge University Press. (Reprinted from *Philosophical Magazine, Series 5, 50,* 157–175, 1900)

Wald, A. (1955). *Selected papers in probability and statistics*. New York: McGraw-Hill.

I

Index Test

A variety of methods have been proposed to examine the accuracy of diagnostic (or prognostic) tests when data are presented in a 2 × 2 format (see Table 1). Rule-in (true positive, or TP) and rule-out (true negative, or TN) accuracy are usually best considered separately, but there are circumstances in which a summary of the overall performance is needed. There are four common methods to calculate overall performance. These are the Youden index, the predictive summary index (PSI), the receiver operator curve (ROC), and the overall accuracy. The ROC, which can be considered a special case where overall performance of a test across a range of cut-off scores is needed, is not discussed in this entry. Each of these major methods has an associated reciprocal of absolute benefit, that is, the number to diagnose one additional case using the stated method (see Table 2).

Youden Index

The Youden index (Youden's *J*) is based on the characteristics of sensitivity and specificity. Sensitivity and specificity are essentially measures of occurrence rather than gain or clinical value. For example, an 80% sensitivity simply describes that a result occurs in 8 out of 10 of those with the index condition. Yet a test that was positive in 80% of those with a condition might or might not be valuable depending on the prevalence of that condition and also the number of times the test was positive in those without the condition. Sensitivity and specificity are often considered a hypothetical rather than clinical measure because their calculation requires application of a reference (or criterion) standard. In clinical practice, a reference standard is not usually calculated for all patients, hence the need for the test itself. In 1950, William John Youden (1900–1971) proposed the Youden index. It is calculated as follows: [$J = 1 - (\alpha + \beta)$ or *sensitivity* + *specificity* − 1]. If a test has no diagnostic value, sensitivity and specificity would be 0, and hence $J = -1$; a test with modest value where sensitivity and specificity = .5 would give a $J = 0$. If the test is perfect, then $J = +1$. The Youden index is probably most useful where sensitivity and specificity are equally important and where prevalence is close to .5. As these conditions often do not apply, other methods of assessing the value of diagnostic tests have been developed.

The Predictive Summary Index

In most clinical situations when a diagnostic test is applied, the total number of positive test results (TP + FP) (true positive + false positive) and negative test results (TN + FN) (true negative + false negative) is known although the absolute number of TP and TN is not. In this situation, the accuracy of such a test may then be calculated from the positive predictive value (PPV) and negative predictive value (NPV). Several authors have suggested that PPV and NPV are preferable to sensitivity and specificity in clinical practice. Unlike sensitivity and specificity, PPV and NPV are measures of discrimination

Table 1 Generic 2 × 2 table

	Reference Standard Disorder Present	*Reference Standard No Disorder*	
Test +ve	A	B	A / A + B PPV
Test –ve	C	D	D / C + D NPV
Total	A / A + C Sn	D / B + D Sp	

Note: Positive predictive value (PPV), negative predictive value (NPV), sensitivity (Sn), and specificity (Sp).

Table 2 Summary measures of diagnostic accuracy

Measure	*Basic Formula*	*Strength*	*Weakness*	*Reciprocal Absolute Benefit*	*Reciprocal Absolute Benefit Formula*
Youden index	Sensitivity + Specificity – 1	Relatively independent of prevalence Not clinically interpretable	Requires application of criterion (gold standard) Does not assess ratio of false positives to negatives	Number needed to diagnose	NND = 1 / Youden
Predictive summary index	PPV + NPV – 1	Measures gain Clinically applicable	Dependent of prevalence Places equal weight on rule-in and rule-out accuracy	Number needed to predict	NNP = 1 / PSI
Overall accuracy (fraction correct)	(TP + TN) / (TP + FP + TN + FN)	Measures real number of correct identifications versus misidentifications Can be easily converted into a percentage	Requires application of criterion (gold standard)	Number needed to screen	NNS = 1 / Identification index

Note: Positive predictive value (PPV), negative predictive value (NPV), predictive summary index (PSI), true positive (TP), true negative (TN), false positive (FP), false negative (FN).

(or gain). The gain in the certainty that a condition is present is the difference between the posttest probability (the PPV) and the prior probability (the prevalence) when the test is positive. The gain in certainty that there is no disease is the difference between posttest probability of no disease (the NPV) and the prior probability of no disease (1 – Prevalence). This is best illustrated in a Bayesian plot. In the Bayesian plot shown in Figure 1, the pretest probability is plotted in dark shading, and the posttest probability is plotted in gray shading where a test is positive and without shading where the test is negative. Thus, using the example of 80% sensitivity and specificity, the thick black line illustrates the posttest probability given a pretest probability of .5 and thus the overall gain in probability of an accurate diagnosis compared with the baseline probability. In this case, it is pre-post gain +ve (.8 – .5) plus pre-post gain –ve (.5 – .2) = .6.

Considering the overall benefit of a test from positive to negative, then, the *net gain* in certainty is a summation of [PPV – Prevalence] + [NPV – (1 – Prevalence)] = PPV + NPV – 1. This is the PSI. The PSI is usually a better measure of applied test performance than the Youden score. However, its strength is also its limitation as it is dependant on the underlying prevalence, reflecting real-world probabilities. This may be an advantage where the performance of a test must be calculated for particular settings, but occasionally it can be a disadvantage where test performances need to be compared across varying settings.

Overall Accuracy (Fraction Correct)

A third approach to calculating accuracy is to measure the overall fraction correct (FC). The overall FC is given by (TP + TN)/(TP + FP + TN + FN) or (A + D)/(A + B + C + D) from Table 1. 1 – FC is the fraction incorrect (or [FP + FN]/[TP + FP + TN + FN]). Arguably, the FC is not as clinically applicable as the PSI because the actual number of TP and TN must be known to calculate overall accuracy. However, if known, FC can be useful because it reveals the real number of correct versus incorrect identifications. It places equal weight on TP and TN, which may be misleading in some circumstances; for example, where an FP leads to retesting but an FN leads to no treatment. Recently, Alex Mitchell proposed a method to aid interpretation of the FC. The fraction correct minus the fraction incorrect might act as a useful "identification index," which can be converted into a number needed to screen. Thus,

$$\text{Identification index} = \text{FC} - (\text{Fraction incorrect})$$

$$\text{Identification index} = \text{FC} - (1 - \text{FC})$$

$$\text{Identification index} = 2 \times \text{FC} - 1.$$

Reciprocal Measures of Accuracy

Number Needed to Diagnose

The reciprocal of Youden's *J* was suggested as a method to calculate the number of patients who need to be examined in order to correctly detect one person with the disease. This has been called the *number needed to diagnose* (NND) originally suggested by Bandolier. Thus, NND = 1/[sensitivity – (1 – specificity)]. However, the NND statistic is hampered by the same issues that concern the Youden score, namely, that it is insensitive to variations in prevalence and subject to confusion in cases where sensitivity is high but specificity low (or vice versa). Additionally, the NND becomes artificially inflated as the Youden score approaches 0, and this is misleading because the Youden varies between –1 and +1, not +1 and 0. In short, the reciprocal of Youden's *J* is not a clinically meaningful number.

Number Needed to Predict

An improvement on the NND is to take the reciprocal of the PSI. This was proposed by Linn and Grunau and called the *number needed to predict* (NNP), which is the reciprocal of the PSI. Unlike the NND, this does reflect the local conditions of the test, that is, the current prevalence. However, it assumes equal importance of the PPV and NPV and may be prone to error when the sum of the PPV and NPV equals 1.0.

Number Needed to Screen

Mitchell recently suggested a new method called *the number needed to screen* (NNS) based on the difference between the real number of correctly diagnosed and incorrectly diagnosed patients. The

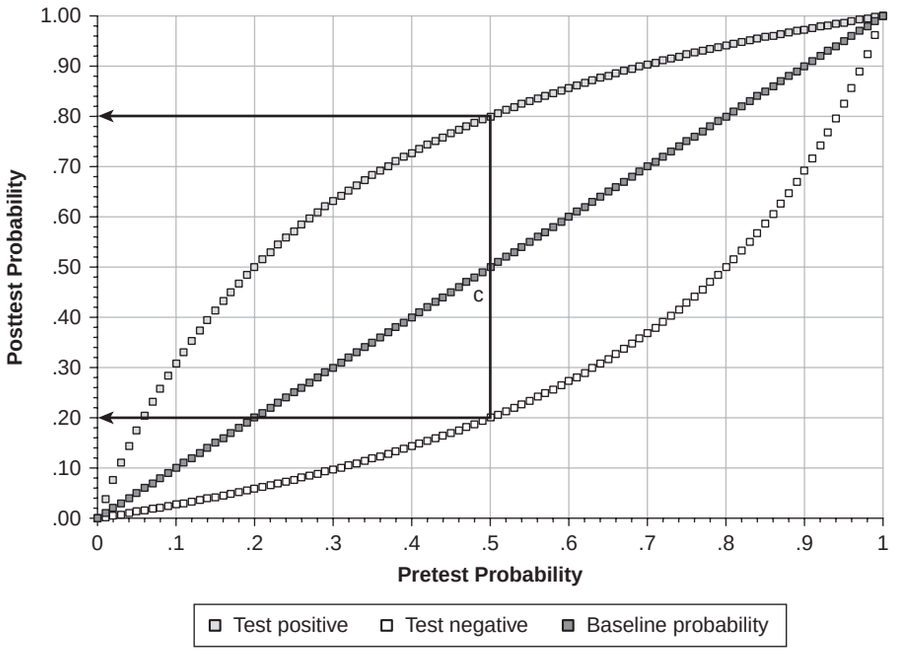


Figure 1 Pretest and posttest probability, given 80% sensitivity and 80% specificity

number needed to screen = 1/FC – (Fraction incorrect) or 1/Identification index.

Take a hypothetical example of a new screening test for Alzheimer's disease tested in 100 with the condition and 1,000 without which yields a sensitivity of .90 and a specificity of .50. The Youden score is thus .4 and the NND 2.5, suggesting 2.5 individuals are needed to diagnose one person with Alzheimer's disease. In fact, out of every 100 applications of the test, there would be 9 people with Alzheimer's disease (Prevalence × 100) of whom 90% would be true positives (= 8.2) and 81 without Alzheimer's disease (1 – Prevalence × 100) of whom 50% would be negatives (= 45.5). In this example, there would be 53.6 true cases per 100 screened (FC per 100 cases) but at the expense of 46.4 errors (Fraction incorrect) per 100 screened, a net gain of 7.3 identified cases per 100 screened. Thus, the NNS would be 13.75 applications of the test to yield one true case *without error*.

Unlike the Youden score or the NND, the clinical interpretation of the NNS is meaningful. It is the actual number of cases that need to be screened to yield one additional correct identification (case or noncases) beyond those misidentified. Unlike the Youden score and NND, which equally favor sensitivity and specificity regardless of baseline prevalence, the NNS emphasizes minimum errors taking into account the prevalence (or study sample). The NNS of a test will approach 1 as it reaches perfect accuracy. The unusual but not impossible situation in which a test that yields more errors than correct identifications will have a negative identification index, in which case the magnitude of the NNS can be interpreted as the actual number of cases that need to be screened to yield one additional mistaken identification (case or noncases) beyond those correctly identified. This is akin to the concept of number needed to harm (NNH) and requires no additional calculation in this case.

Discussion

Various methods have been suggested to examine the accuracy of diagnostic (or prognostic) tests when data are presented in a 2 × 2 format. Although the sensitivity, specificity, PPV, and NPV are often used by default, their strengths and weaknesses should be considered. Summary methods are most appropriate where one test must be used for both making and excluding a diagnosis. In many situations, the optimal test for diagnosis (case finding) may be different from the optimal test for exclusion. Therefore, where possible, clinicians should

examine rule-in and rule-out accuracy separately and thus not rely on Youden (NND), PSI (NNP), or even the NNS.

Summary measures such as the Youden score and its derivative, the NND, appear to be also limited by scores that are not clinically interpretable. Youden is best confined to pooled and meta-analytic comparison of multiple tests under differing conditions where prevalence varies. The number needed to treat is a clinically interpretable statistic although it is not without problems. Mitchell proposes that the optimal equivalent statistic for diagnosis is the NNS and not the NND. In clinical practice, choice of the optimal screening (diagnostic) or predictive method will also rely on issues of cost, acceptability, and practicality.

Alex J Mitchell

See also Diagnostic Tests; Number Needed to Treat; Receiver Operating Characteristic (ROC) Curve

Further Readings

Altman, D. G. (1998). Confidence intervals for the number needed to treat. *British Medical Journal, 317,* 1309–1312. Retrieved January 16, 2009, from http://bmj.com/cgi/content/full/317/7168/1309

Bandolier. (1996). How good is that test II. *BNET Healthcare, 27,* 2. Retrieved December 29, 2008, from http://findarticles.com/p/articles/mi_m5LWU/is_5_3/ai_n25024124/pg_1?tag=artBody;col1

Connell, F. A., & Koepsell, T. D. (1985). Measures of gain in certainty from diagnostic test. *American Journal of Epidemiology, 121,* 744–753.

Hutton, J. L. (2000). Number needed to treat: Properties and problems. *Journal of the Royal Statistical Society: Series A* (Statistics in Society), *163*(3), 381–402.

Laupacis, A., Sackett, D. L., & Roberts, R. S. (1988). An assessment of clinically useful measures of the consequences of treatment. *New England Journal of Medicine, 318,* 1728–1733.

Linn, S., & Grunau, P. D. (2006). New patient-oriented summary measure of net total gain in certainty for dichotomous diagnostic tests. *Epidemiologic Perspectives & Innovations, 3,* 11.

Moons, K. G. M., & Harrell, F. E. (2003). Sensitivity and specificity should be de-emphasized in diagnostic accuracy studies. *Academic Radiology, 10,* 670–672.

Pepe, M. S. (2003). The statistical evaluation of medical tests for classification and prediction. In *Oxford Statistical Science Series 28*. Oxford, UK: Oxford University Press.

Salmi, L. R. (1986). Re: Measures of gain in certainty from a diagnostic test. *American Journal of Epidemiology, 123,* 1121–1122.

Youden, W. J. (1950). Index for rating diagnostic tests. *Cancer, 3,* 32–35.

Influence Diagrams

The medical profession leads physicians and, in general, healthcare personnel to frequently face delicate situations that require accurate and careful decision making. The utilization of decision-support models is necessary to improve the decision-making process and to identify the optimal policy. The models allow unveiling of the decisions to be taken, their consequences, and the uncertain events that are involved in the problem. Influence diagrams are graphical tools for the representation and solution of decision-making problems. By representation, one means the identification of the decision-making problem elements. In particular, an influence diagram reveals the probabilistic dependences among the uncertain quantities and the state of information at each decision stage. By solution, one means the determination of the preferred alternative (best strategy selection) given the state of information. Influence diagrams grant decision makers the possibility of representing complex decision-making problems in a thorough albeit compact fashion. It is this strength over other representation techniques that has made the use of influence diagrams widespread in medical applications. This entry is organized as follows. It provides a description of influence diagrams by means of a sample medical example. The analysis of nodes and arcs of influence diagram follows. The discussion of the properties and levels of influence diagrams is offered next. The discussion is a prelude to a synthetic description of influence diagram solution algorithms. The relationship between influence diagrams and decision trees, and a brief mention about other graphical representation techniques, concludes the entry.

Description

An influence diagram is a directed graph composed of nodes and arcs (Figure 1).

The graph is acyclic. There are three types of nodes: decision, chance, and value. A decision node is represented by a rectangular box. A chance node is represented by a circular box. There is one unique value node displayed by a diamond or rhombus; occasionally an octagon is used. Arrows joining nodes are called arcs. The value node ends the diagram, and an influence diagram containing a value node is said to be oriented. In case there are no value nodes and decision nodes, the influence diagram coincides with a Bayesian network.

Figure 1 displays an influence diagram representing the following decision-making problem. A physician must select the treatment for a patient. The first stage of the treatment foresees a choice between Cures A or B. The two cures have a different efficacy and a different cost, with their overall effect strongly dependent on the patient's response. After 1 week, the physician reevaluates the patient's conditions. Depending on the evaluation results, the physician has to decide between continuing with Cure A, switching to B, or resorting to a third cure, C. The problem contains two (sequential) decisions.

Elements

Nodes

An influence diagram contains three types of nodes. Decision nodes display the decisions to be taken at different stages of the decision analysis problem at hand. A variable contained in a decision node is an alternative. The choice among alternatives is under the control of the decision maker, who selects the alternative that maximizes the decision maker's utility. In Figure 1, decision node "Decision 1: A or B?" represents the first selection between Cures A and B; the node "Decision 2: A, B, or C?" represents the selection between Cures A, B, and C. The second selection is made after reevaluation of the patient's conditions.

Chance nodes represent variables or events whose knowledge or realization is out of the control of the decision maker. Chance nodes are sometimes referred to as uncertainty or event nodes. Each chance node contains all possible realizations of the corresponding uncertain variable. Realizations are called outcomes. In Figure 1, the chance node "Patient conditions 1" represents the conditions of the patient after selection of A or B. If the analyst/decision maker considers that three possible states, namely, "fully recovered," "partially recovered," and "worsened," are possible, then the chance node will have three outcomes. The decision maker's state of belief of the likelihood of the outcomes is characterized by a corresponding conditional probability distribution.

Value nodes contain the decision maker's utility for each consequence. A consequence is the end state of the world resulting as a combination of selected alternatives and outcomes of uncertain events. Utility is a quantification of preference and must be assessed consistently with the axioms of decision theory. Value nodes are occasionally referred to as utility nodes.

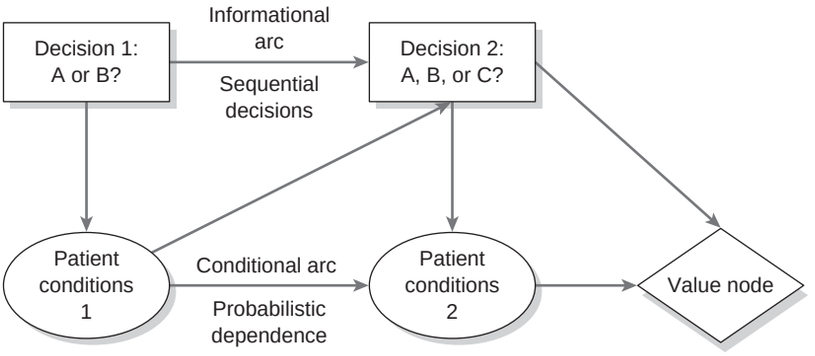


Figure 1 An influence diagram with two chance nodes, two decision nodes, and the value node. The order of nodes is Decision 1, Chance Node 1, Decision 2, Chance Node 2, and Value Node.

Arcs

Arrows joining nodes in an influence diagram are called arcs. Arcs are grouped into the categories of informational and conditional.

Arcs ending in decision nodes are informational. If the informational arc stems from a chance node, then it indicates that the decision maker is aware of the outcome of the chance node at the moment he or she has to make the decision. If an arc connects two decision nodes, then the decision maker is aware of the previously selected alternatives. Arcs connecting decision nodes are also called no-forgetting arcs. Informational arcs imply time precedence. For this reason, it is not possible to reverse the direction of an informational arc without altering the structure of the decision-making problem.

Arrows leading to chance nodes are conditional arcs. Conditional arcs indicate a probabilistic dependence among the distribution of the random variables contained in the chance nodes that the arcs link. Recall that probabilistic dependence indicates a form of relationship weaker than causal dependence. Let X and Y be the two random variables represented by the two chance nodes. Two cases are possible: X is probabilistically dependent on Y or it is not (Figure 2).

If there is a probabilistic dependence, this is displayed by the presence of an arc. The direction of the arrow shows the state of information of the decision maker, that is, whether the decision maker is capable of expressing the probabilities as $P(X|Y)$ or as $P(Y|X)$. A conditional arc does not necessarily correspond to time precedence. In fact, it is always possible to reverse a conditional arc using Bayes's theorem, provided that this operation is performed in a consistent fashion. The lack of a conditional arc between two nodes is a strong assertion about independence.

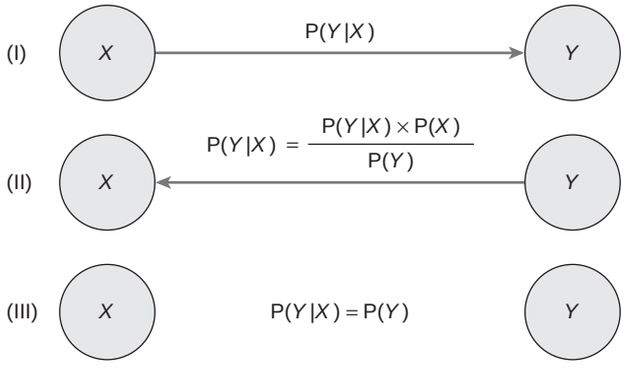


Figure 2 Arcs connecting chance nodes: I and II denote probabilistic dependence, III denotes independence

The influence diagram in Figure 1 contains six arcs. The arc between decision nodes "Decision 1: A or B?" and "Decision 2: A, B, or C?" is an informational arc. It denotes the fact that the physician, at the moment of the second decision, is aware of whether Cure A or B has been previously selected. The arc from chance node "Patient Conditions 1" to decision node "Decision 2: A, B, or C?" is also an informational arc, denoting that the physician, at the moment of the second decision, is informed of the patient's conditions after adoption of Cure A or B. The arc between chance nodes "Patient Conditions 1" and "Patient Conditions 2" is a conditional arc representing the fact that the decision maker considers the outcomes of the second chance node dependent on the conditions of the patient after the selection of the first cure.

Predecessors, Successors, Paths, and Barren Nodes

Let (i) and (j) denote any two nodes. A path from (i) to (j) is a set of arcs that starts from (i) and forms a directed line leading to (j). Thus, influence diagrams are directed graphs. The set of nodes that have a path leading to (j) is called the set of predecessors or parents of (j). Direct predecessors are defined as the subset of predecessors with an arc leading directly to j. Similarly, a successor is any node in the diagram that lies in a path emanating from node j. A direct successor is any node connected to (j) by arcs stemming from (j).

If a chance or decision node does not possess any successor, then it is said to be barren. If a chance node is barren, then it can be removed from the diagram without altering the evaluation. If a decision node is barren, it can be removed as well. In addition, all the alternatives in a barren decision node are optimal.

A diagram is acyclic if there are no paths starting at a node and ending at the same node. Acyclicity ensures that a decision maker does not infer information from a decision he or she has not yet made. An influence diagram is regular if it is acyclic, if the value node ends the diagram, and if

there is at least one directed path connecting (and therefore ordering) all the decision nodes.

Influence Diagrams Levels

One distinguishes three levels of influence diagrams: graphical, functional, and numerical. The graphical level displays nodes and arcs, evidencing probabilistic dependence and the flow of information, that is, the information available before each decision (Figure 1). At the functional level, one introduces the outcomes, the conditional distributions, and the alternatives of each chance and decision node, respectively. The numerical level is the level at which the values of the conditional probabilities and utilities are inserted. The insertion of the numerical values is necessary for the solution of the decision-making problem.

Algorithms for Solving Influence Diagrams

Solving an influence diagram means to determine the preferred strategy, that is, to identify the optimal choice at each decision node. The mathematical function associated with a decision node is the max(·) function, as a decision maker selects the alternative that maximizes expected utility. The operation associated with a chance node is conditional expectation.

In the original work of Ronald A. Howard and James E. Matheson, the solution of influence diagrams is envisioned in a two-step approach, through conversion of the influence diagram into the corresponding decision tree. A few years later, Ross D. Shachter proposes a complete algorithm for direct solution of influence diagrams. The direct solution algorithm proceeds through arc reversals and node elimination. These operations follow strict rules so as not to allow distortion of the calculation of the expected utilities and the flow of information. Such operations are called value preserving. The four main types can be listed as follows, in accordance with the taxonomy of Joseph A. Tatman and Ross D. Shachter: (1) arc reversal, (2) chance node removal through summation, (3) chance node removal by conditional expectation, and (4) decision node removal by maximization. The procedure foresees first the removal of barren nodes, followed by the iterative application of the four operations, until the best strategy is identified.

Significant research on the solution of influence diagrams has been undertaken in the fields of computer science and operations research, and numerous algorithms have been developed. This has resulted in the availability of commercial software that allows decision makers to implement and solve decision analysis problems directly on a personal computer by representing them in the form of influence diagrams.

Decision Trees and Other Graphical Techniques

Influence diagrams are often used in conjunction with decision trees. Some commercial and open-source software allows users to first structure the model in the form of an influence diagram and then to obtain the corresponding decision trees. Decision trees have the advantage of displaying the combinations of choices and outcomes that lead to each consequence, thus providing a detailed description of the decision-making problem. However, their size increases exponentially with the number of nodes. Not all influence diagrams can be directly converted into a decision tree, and to one influence diagram there can correspond more than one decision tree. The conditions that ensure the possibility of transforming an influence diagram into a decision tree are the single decision-maker condition and the no-forgetting condition. These names follow the taxonomy of Howard and Matheson. An influence diagram sharing these two requirements is called a decision network.

Besides decision trees and influence diagrams, related graphical techniques for the representation of decision-making problems are valuation networks and sequential decision diagrams. Concha Bielza and Prakash Shenoy have compared the four techniques, illustrating their merits and shortcomings.

As an example of application in the medical discipline, Manuel Gómez, Concha Bielza, Juan A. Fernández del Pozo, and Sixto Ríos-Insua use influence diagrams for decision aid in neonatal jaundice problems.

Emanuele Borgonovo

See also Applied Decision Analysis; Bayesian Networks; Bayes's Theorem; Decision Trees, Advanced Techniques in Constructing; Probability; Tree Structure, Advanced Techniques; Utility Assessment Techniques

Further Readings

Bielza, C., & Shenoy, P. P. (1999). A comparison of graphical techniques for asymmetric decision problems. *Management Science, 45*(11), 1552–1569.

Gómez, M., Bielza, C., Fernández del Pozo, J. A., & Ríos-Insua, S. (2007). A graphical decision-theoretic model for neonatal jaundice. *Medical Decision Making, 27,* 250–265.

Howard, R. A., & Matheson, J. E. (1984). Influence diagrams. In R. A. Howard & J. E. Matheson (Eds.), *Readings on the principles and applications of decision analysis* (Vol. II). Menlo Park, CA: Strategic Decisions Group. (Reprinted in *Decision Analysis*, Vol. 2, No. 3, September 2005, pp. 127–143)

Lee, R. C., Ekaette, E., Kelly, K.-L., Craighead, P., Newcomb, C., & Dunscombe, P. (2006). Implications of cancer staging uncertainties in radiation therapy decisions. *Medical Decision Making, 26,* 226–237.

Qi, R., & Poole, D. (1995). A new method for influence diagram evaluation. *Computational Intelligence, 11*(1), 498–528.

Shachter, R. D. (1986). Evaluating influence diagrams. *Operations Research, 34*(6), 871–882.

Shachter, R. D. (1988). Probabilistic inference and influence diagrams. *Operations Research, 36*(4), 589–604.

Smith, C. L., & Borgonovo, E. (2007). Decision making during nuclear power plant incidents: A new approach to the evaluation of precursor events. *Risk Analysis, 27*(4), 1027–1042.

Tatman, J. A., & Shachter, R. D. (1990). Dynamic programming and influence diagrams. *IEEE Transactions on Systems, Man and Cybernetics, 20*(2), 365–379.

von Neumann, J., & Morgenstern, O. (1944). *Theory of games and economic behavior* (1st ed.). Princeton, NJ: Princeton University Press.

Zhang, N. L. (1998). Probabilistic inference in influence diagrams. *Computational Intelligence, 14*(4), 475–497.

Information Integration Theory

Information integration theory (IIT) is an approach to the mathematical modeling of judgment developed by Norman Anderson. Although its original application was intended for the measurement of descriptions of personalities, it has since then been used to model many different kinds of judgments and decisions, including medical and health decisions. IIT is a metatheory or framework for studying judgments, along with an associated set of modeling methods. (See Figure 1.)

The IIT model posits four types of constructs associated with the judgmental response to a stimulus. The physical attributes of the stimulus, conventionally denoted by Φ_i, are the components of a stimulus that are observable by the judge. Each physical attribute is mapped to a psychological scale value, conventionally denoted by s_i, through psychophysical functions. The scale values are combined to form an overall psychological impression of the stimulus, conventionally denoted by Ψ, through a combination function, such as addition or averaging. Finally, the overall impression is mapped to an overt response by the judge to the judgment task (e.g., a category rating) by means of a response transformation function. For example, if the stimuli are health states, each attribute of the health state (pain, functioning, etc.) would be mapped onto a scale value and then combined to form the overall impression of the health state, which might then be reported by the judge as a utility through a 0 to 100 rating or used by the judge as a basis for preferences in a choice-based utility assessment.

IIT experiments typically present judges with a set of stimuli in which the attributes are systematically manipulated and responses are collected. By plotting responses against levels of stimulus attributes, the experimenter can make inferences about the form of the combination function in terms of cognitive algebra. For example, assuming that the response transformation is linear, a multiplicative combination function is characterized by a fan-shaped (convergent or divergent) set of curves, while an additive function is characterized by a set of parallel curves. Additional experiments can distinguish between additive and averaging functions, as well as other functional forms. For example, judgments of overall satisfaction with a group practice might be an additive function of the satisfaction with the individual physicians in the practice (where each additional moderately satisfactory physician increases the perception of the practice), an averaging function (where each additional moderately satisfactory physician pulls the practice toward "moderate

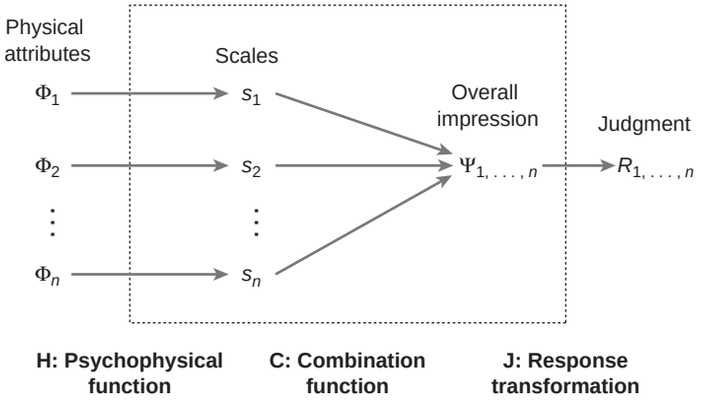


Figure 1 The information integration theory framework

satisfaction"), or a configural weighted function (where the most or least satisfying physician is the sole determinant of the satisfaction with the practice).

Given the basic form of the combination function and response function, the experimenter can fit models to derive the scale values associated with each stimulus, the overall psychological impression of the stimulus, and the parameters of the three functions.

IIT is particularly useful when modeling task or context effects on judgment because it separates the processes of perception, integration, and response. For example, judgments of the same stimuli made using two different response tasks (e.g., a category-rating scale of the importance of a health benefit and a statement of willingness-to-pay to achieve the health benefit) can be modeled with the same psychophysical and combination functions. The underlying overall impression of the health benefit would be assumed not to change as a result of the task, and modeling would focus on fitting response functions to each task that best capture the responses under that assumption. These response functions may be of interest in themselves or may be important in the interpretation of responses to novel health benefits. On the other hand, differences in stimulus context within a single response task (e.g., differences in the perception of stimuli depending on the overall range of stimuli or other factors at the level of the stimulus set) can be modeled in the psychophysical or combination functions to gain insight into how attributes of choices are perceived and how these perceptions relate to the overall judgment.

Alan Schwartz

See also Conjoint Analysis; Context Effects; Judgment; Lens Model

Further Readings

Anderson, N. H. (1981). *Foundations of information integration theory*. New York: Academic Press.
Anderson, N. H. (1982). *Methods of information integration theory*. New York: Academic Press.
Anderson, N. H. (1991). *Contributions to information integration theory*. Hillsdale, NJ: Lawrence Erlbaum.

Informed Consent

Informed consent within court decision making is a complex linkage of information and a patient's assent to (approval of) a medical intervention that the patient's physician is recommending in the patient's care, based on that physician's disclosure of the information. The complexity of linkage of consent and assent (approval) relates to the nature and depth of the information disclosure required of physicians by the courts.

The term *informed consent* did not enter the judicial lexicon until 1957 when the term appeared in a California appellate decision, *Salgo v. Leland*

Stanford Junior Board of Trustees. The judge in the *Salgo* decision simply used the term *informed consent* without defining it.

Consent

Consent is a concept that can be traced back to 1767 in the British court decision *Slater v. Baker and Stapleton*. In this court case, a patient alleged that a physician had not obtained his consent prior to intervening medically. The patient had broken his femur, and the physician set the femur. The patient was seen by the physician's father, who was also a physician, and the father judged the femoral fracture to be healing well with callus formation. When the original physician saw the patient again in follow-up, the physician re-broke the healing fracture and set the fracture in a mechanical device with teeth.

Today, this intervention by the physician would be viewed as an experiment. The British court, however, treated the issue as an issue of consent in the patient-physician relationship and asked whether there was a custom among physicians to secure their patient's consent prior to intervening on a patient. The physicians summoned by the court to testify argued that there was such a custom among physicians. And the court decided the case in favor of the patient.

Basis of Doctrine of Consent

In the United States, Judge Benjamin Cardoso in 1914 grounded consent on the patient's right of self-determination, and in the United States subsequent cases continued this view that physicians had an obligation to obtain their patients' consent on the basis of the patient's right to self-determination. Great Britain and Australia consider consent as grounded on a physician's duty of care, not the patient's right of self-determination.

Battery Versus Negligence

In the United States, the failure of a physician to secure a patient's consent prior to a medical intervention was viewed in terms of a "battery" or intentional harm to the patient. Courts in the United States continued to view cases of lack of consent as battery until 1960, at which time the Kansas Supreme Court in *Natanson v. Kline* argued that consent should be considered on a theory of negligence on the part of the physician. In the United States, the State Supreme Court of Pennsylvania continues to hear informed consent cases brought against physicians in terms of battery. In Great Britain, consent has been considered in terms of negligence on the part of physicians.

Standards

Professional Standard

From 1767 to 1972, there was only one standard of consent and informed consent, the professional standard. Under a professional standard, a physician is judged in terms of whether or not he or she secured a patient's consent or informed consent in terms of whether physicians in the physician's community of peers would have obtained consent or informed consent. And the information that a physician discloses (or does not disclose) to a patient is judged in terms of whether a physician in that physician's community of peers would have disclosed that information to his or her patient. Thus, from 1767 to 1972, physician testimony in the courtroom determined how a jury was to decide a case. The jury would be asked to decide a question such as the following: Did the physician obtain consent from the patient in terms of how physicians in his or her community of peers would have obtained consent? And in terms of the information provided, did the physician provide that information that physicians in his or her community of peers would provide to their patients, based on the physicians' testimony as given in the courtroom.

Reasonable Person Standard

Judge Spottswood Robinson, in the U.S. landmark federal case *Canterbury v. Spence* in the District of Columbia in 1972, argued that there never was such a standard as the professional standard of consent or informed consent. He argued that because there was no agreement among physicians as to what information should be disclosed to patients in informed consent, a new standard needed to be introduced. Robinson argued for the reasonable person standard where a physician was to be judged in terms of whether that physician provided to the patient that information that a

reasonable person in that patient's position would want to know. He did not argue for a standard that the patient should be told what a *reasonable person* would want to know. Rather Robinson argued for a standard where the patient should be told what *a reasonable person in the position of the patient* would want to know.

Robinson used the concept of "hindsight bias" to argue against a subjective patient standard in informed consent. Here, he argued that a patient's testifying about what he or she would have wanted prior to a procedure in which the patient had actually sustained the adverse outcome would be influenced by hindsight bias. Robinson thus argued that a jury needed to consider not what the patient testifies he or she would want to have known in informed consent but what a reasonable person in the patient's position would want to know.

Today, only Great Britain continues to hold a professional standard in consent. Both Canada and Australia have adopted a reasonable person standard of informed consent (Canada) and consent (Australia).

Reasonable Volunteer Standard

In the area of research on humans, the United States holds a much more stringent standard in informed consent than provided by either a professional standard or a reasonable person standard. This standard is the reasonable volunteer standard. The reasonable volunteer standard was developed by the National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research and described in the Belmont Report published in 1978. In the Belmont Report, this National Commission rejected the professional standard in research because the professional standard assumed a common understanding among physicians on what was to be done medically and how a patient should be informed, whereas research on humans is conducted when such common knowledge does not exist. The National Commission rejected the reasonable person standard because the research subject is a volunteer who may well want to know much more about the risks of research than does a patient who delivers himself or herself to a physician for needed care. The reasonable volunteer standard of the Belmont Report argues that a study participant must be given that information that a reasonable volunteer would want to know. Only the United States holds the reasonable volunteer standard in research on humans.

PARQ

In clinical care, Robinson argued that the patient was to have disclosed information about the nature of the procedure (P), the alternatives to the procedure (A), and the risks of the procedure (R). And the physician must answer truthfully and honestly any questions that a patient has (Q).

Robinson argued that the physician has the obligation to disclose information to the patient in informed consent. It is not the obligation of the patient to ask questions to start the disclosure. However, after the physician makes his or her disclosure in informed consent, the patient has the right to seek truthful and clear answers of his or her physician, and the physician must answer to the best of the physician's abilities the patient's questions.

A problem here is that the physician may not have answers to the patient's questions. For example, in the case of Cox-2 inhibitors used to treat the discomfort of arthritis, at the present time it is known scientifically that Cox-2 inhibitors increase the risk of heart attack and stroke in some patients. If the patient asks the question "*How* do Cox-2 inhibitors increase the risk of heart attack and stroke?" at the time of this writing, the physician cannot answer this question because the issue of the causal mechanism of how Cox-2 inhibitors cause the increased risk of heart attack and stroke is not known. As in many questions in medicine and surgery, the fact of the matter is that a full causal explanation may be unavailable at the time the patient asks the causal questions "Why does the reaction occur?" or "How mechanistically does the prescription medicine cause the effect in the patient?"

Risk Disclosure

In the case of the high court decisions noted above in Great Britain, the United States, Canada, and Australia, each case involved the alleged failure of the physician caring for the patient to disclose to the patient—in the consent or informed consent session surrounding the physician-recommended medical intervention—the severe adverse outcomes

that were known risks of the medical intervention and that severe risk materialized in the patient during that medical intervention. Thus, consent and informed consent as focused in our high courts have risk disclosure as their primary focus.

Historically, in consent and informed consent cases, the type of information that was not disclosed to a patient was information about severe adverse outcomes that were estimated to have a low chance of occurrence. Robinson argued that patients were interested in severe adverse outcomes such as death or stroke no matter what the chance (probability) of their occurrence.

Judge-Made Law

Every physician and every physician-in-training must understand that consent or informed consent had its origins in judge-made law in Great Britain, the United States, Canada, and Australia. Today, state legislatures also influence informed consent within particular states in the United States. But the point is that physicians must always understand the laws of consent and informed consent by which they will be judged should a patient bring a case against them as physicians.

Nature of Risk

Robinson argued that the physician must translate medical terminology into language the patient can understand. Yet there are problems in understanding what certain terms mean scientifically in common language. For example, a physician talking about the risks of carotid endarterectomy translates the medical term *cerebral vascular accident* into *stroke* and informs the patient that there is a risk of stroke during and after carotid endarterectomy. Yet the patient unbeknownst to the physician does not realize that stroke means more than motor paralysis and that stroke can cause sensory, motor, cognitive, and other types of damage. This is a problem in understanding what a term means medically or scientifically when the term is a common language word such as *stroke*.

Chance

High courts have considered risk disclosure primarily as the disclosure of adverse outcomes and have not focused on the disclosure of the chance (probability) of that adverse outcome. When the Supreme Court of Canada in *Reibl v. Hughes* in 1980 examined where the risk numbers come from in informed consent disclosures, this court argued that that may be reason not to disclose chance when there may be a great debate about what that chance actually is.

Nature of Medical Decision Making

After recognizing the importance of the judge-made law of consent and informed consent, one must recognize that the risk disclosure that dominates court decision making is only a part of medical decision making and only a part of the tasks that physicians face and must consider as they help patients reach a decision. The minimal information that influences a patient's decision is risk information, benefit information, and the physician's opinion in his or her own care. Risk and benefit are typically obtained through the peer-reviewed medical literature. This literature itself is geared more toward benefit information and less toward the exposition of risk information. Indeed, the publication of research studies in the peer-reviewed medical literature depends on the interest in the research, and the discovery of new benefits is often better reading even in the physician community than is the discovery of new risks related to a medical intervention or a medical product such as a medical device or a prescription medicine. Indeed, recent research on Cox-2 inhibitors has shown the lack of systematic scientific understanding of the mechanism of action of the risks of increasing the chance of heart attack and stroke before medicines of that class of drug were approved and marketed, as such risks were identified only after the drug's approval for marketing.

Yet it must also be recognized that a physician who is attending only to optimum risk disclosure to his or her patient is doing an injustice to that patient in terms of failing to deliberate about and fully consider what the medical decision the patient is facing is actually about. Here, the full medical decision depends on balancing the risks and benefits across the range of diagnostic and therapeutic interventions that are available and scientifically understood by their exposition in the peer-reviewed medical literature and then attempting to see to what extent

that literature applies to the individual patient who must come to a decision in his or her own care. For example, the peer-reviewed surgical literature often involves studies in major medical institutions and surgeons who care for and operate on the sickest of the sick patients and who are national or international experts in the care of these patients. In reality, in day-to-day practice, the range of skill of surgeons may fluctuate mildly or markedly from the skill of the expert surgeons. All these points may need to be represented in the decision that a patient and a physician must make in a patient–physician relationship about whether to intervene medically now, whether to delay medical intervention until a later time (watchful waiting), or whether to not intervene medically at all and to allow Nature to take its course.

Dennis J. Mazur

See also Bias; Cognitive Psychology and Processes; Numeracy

Further Readings

Canterbury v. Spence, 464 F.2d 772 (D.C. Cir. 1972).

Faden, R. R., & Beauchamp, T. L. (1986). *A history and theory of informed consent*. New York: Oxford University Press.

Mazur, D. J. (1986). Informed consent: Court viewpoints and medical decision making. *Medical Decision Making, 6,* 224–230.

Mazur, D. J. (1990). Judicial and legislative viewpoints on physician misestimation of patient dysutilities: A problem for decision analysts. *Medical Decision Making, 10,* 172–180.

Mazur, D. J. (2003). Influence of the law on risk and informed consent. *British Medical Journal, 27,* 731–734.

Mazur, D. J., & Hickam, D. H. (1997). Patients' preferences for risk disclosure and role in decision making for invasive medical procedures. *Journal of General Internal Medicine, 12,* 114–117.

Natanson v. Kline, 350 P.2d 1093, 1104 (1960).

Powers, M. (2003). Communicating risk in the United Kingdom. *British Medical Journal, 327,* 735–736.

Reibl v. Hughes, 2 RCS 880 (1980).

Robertson, G. (1981). Informed consent to medical treatment. *Law Quarterly Review, 97,* 102–126.

Rogers v. Whitaker, 175 CLR 479 (1992).

Salgo v. Leland Stanford Junior University Board of Trustees, 154 Cal. App. 2d 560, 317 P.2d 170 (1st Dist. 1957).

Schloendorff v. Society of New York Hospital, 211 N.Y. 125, 105 N.E. 92 (1914).

Sidaway v. Board of Governors of the Bethlem Royal Hospital and the Maudsley Hospital and Others, A.C. 871 (H.L.) (1985).

Slater v. Baker and Stapleton, 95 Eng. Rep. 860, 2 Wils. K.B 359 (1767).

Informed Decision Making

Over the past half century, rapid and demanding modes of change throughout healthcare have converted the physician into a "provider" and the patient into a "consumer." Among the results of this transformation is a pervasive belief that primary authority over patient healthcare decisions, which had long been the domain of the physician, began to shift to the patient and, more recently, to proprietary healthcare organizations. Regardless of who has authority over certain decisions at any point throughout a given case, it is important that all relevant parties engage in informed decision making if they are to achieve a successful outcome.

Informed decision making is a process through which a person uses available information as a means to settling on how to accomplish a certain objective. In the context of medicine and healthcare, the process has traditionally been one in which the physician provides the patient (or a surrogate decision maker) with information about the benefits and risks of various treatment options based on medical tests and a collection of the patient's personal information. Today, while private and government sector organs of managed care are likely to be invested in, provide input to, and dictate treatment alternatives for a patient case, the bulk of information about care and treatment options has historically been communicated within the doctor-patient relationship. But the order of this relationship has for the past several decades begun to undergo enormous change.

One simple reason for such an upheaval in this relationship is that decisions about one's healthcare can be scientifically and emotionally complex. Healthcare is an intention that fundamentally requires the practitioner and the patient to reach a consensus on processes, outcomes, and measurements of care—and to do so with a focus

that balances the preferences of the patient and the dimensions of disease. As such, thoughts and feelings about treatment options vary on an individual basis and are likely to depend on a number of factors, including attitudes toward and understanding of risk, which assumes a perception of probabilities, degrees, and time periods of process and outcome. Another factor is that family, culture, and language can collectively and individually influence healthcare decisions. There are, nevertheless, basic constructs of decision making that cut across social, economic, and political lines.

Principles of Decision Making

Decision making generally involves five phases: (1) definition of the problem, (2) analysis of the problem, (3) development of possible solutions, (4) selection of a perceived best solution, and (5) translation of the decision into action. Each phase contains its own set of steps. Yet research on decision making informs us that people generally have a deficient understanding of the policies and procedures they use in arriving at decisions, even given that they use insight—or dissected information—from previous experiences to guide judgment in new situations.

While decisions are influenced by biological and learned factors, behavior and emotion are also highly integrated elements of decision making. It stands, as the Greek philosopher Epictetus believed, that a given situation does not cause a reaction; rather, it is an individual's attitude about the situation that encourages a reaction to it. That is, because decision making is contingent on the generally unpredictable nature of daily life, a person's opinion or perception about a given situation—and not the situation itself—will cause that person to react in either a positive or negative manner. But when the person is able to become aware of and adjust the beliefs and perspectives that create the condition, the person can better manage the intrinsic and extrinsic factors that affect behavior.

This process functions on a cognitive level when an individual has a belief about a specific event. The belief, which may be a rational one that results in healthy emotions or an irrational one that leads to disturbed emotions, then contributes to an emotional consequence that respectively promotes or inhibits satisfaction. The irrational beliefs derive from a basic absolute—a "must" or a "should"—that represents a demanding and unrealistic perception of how things should be, while the individual directs blame at self and others by developing remarks that exaggerate the event. Such thinking often renders a person incapable of tolerating relatively high levels of frustration, a pattern that can be broken and redirected by challenging the irrational beliefs through rigorous questioning.

Of course, not every individual is unwilling to recognize adversity or finds a situation threatening rather than challenging. But every person does make decisions—and considers change—within an individually perceived set of physical, emotional, and environmental conditions that take place at continuous intervals. Depending on the person and the circumstances, individuals may be in a phase in which they are either unaware of the need to make a change, contemplating how to make a change, preparing to make a change, taking action on making a change, or adhering to a change that has been made already. Of further consequence is that a person can arrive at one phase only to engage in beliefs and behaviors that result in the individual reverting to a prior phase.

Beyond even these cursory tenets of individual, though universal, decision-making behavior lie more overt realities that complicate the means of informed decision making.

The Role and Flow of Information

Information, which can be acquired from numberless sources, has to be organized to help people think through and test basic assumptions about whatever it is that they are trying to accomplish. Of the various elements that comprise informed decision making, much of the process depends on the possession of skills to collect, analyze, and communicate information. This is true for physicians and all manner of health professionals, patients and their advocates, and any other individuals or groups who claim some sort of stake in healthcare.

The best reason for this reality is that during the span of the past 50 years or so, broad changes in social values have resulted in people becoming more inclined to assert their individual rights and autonomy while technological advancements have increasingly provided people with better access to

information. Many patients have consequently taken a more active role in making decisions about their healthcare. This has meant that physicians have had to orient themselves toward ceding some measure of the principal role in decision making. They also have had to prepare themselves to help a patient think through care options that lean less on practitioner knowledge and more on what the patient perceives to be the appropriate course of action.

At the same time, physicians have had to manage a patient population that in any instance may range from being well-informed to uninformed to misinformed about medical practices and standards of care and treatment. But there is now a growing body of evidence that conveys patients sometimes feel physicians have limitations in their awareness of effective courses of action. These seemingly opposite forces are the result of a more recent change in distribution channels that signal a shift from information being filtered and communicated predominantly by the physician to information being filtered and communicated predominantly by the World Wide Web.

But much as new information channels have changed their pattern behaviors, few patients elect to decide on a treatment that entirely disregards the physician's recommendation. In fact, a large majority of patients may actually prefer that the physician take responsibility for selecting a treatment option—so long as the physician comfortably and consistently informs them of courses of action. Even so, patients who are more actively involved in making decisions about their care will on occasion select a treatment alternative that departs from the physician's recommendation. In all but the most extreme or compelling cases, and since the patient is the end user of a decided-upon treatment, physicians must acknowledge—and theoretically accept—that the patient has autonomy to think and behave in associated ways.

Yet from practical, professional, and ethical points of view, it can be difficult to accept a patient's choice when it runs contrary to the physician's recommended approach. Even in instances in which the physician and patient have healthy or heated exchanges of perspective, one can be left to wonder whether all best efforts have been transmitted to influence a decision. There is, however, strong evidence that imparts any exchange of ideas about treatment alternatives that can be beneficial to effective decision making and that all sides more fully appreciate the decision, even if it remains the same as initially determined. But the quality of these interactions is heavily dependent on the primary cords of the doctor-patient relationship, cultural norms, and the behavior of the healthcare system, including any extent to which it is privatized or government run.

Ethical Issues and Conflicts of Interest

As healthcare systems throughout the world increasingly and intentionally wed delivery of care with cost of care, there must be consideration of the balance between autonomy and administrative issues, resources, work-related relationships, and amount of time allotted to visit with patients. One ethical issue that is likely to emerge is the potential for some treatment options and explanations to go unexplored due to either time or financial concerns, or both.

A similar ethical concern that has gained attention of late is the relationship between medical professionals and the pharmaceutical and medical device industries. Physicians possess the professional and legal authority to prescribe pharmaceutical therapies to patients, who are customarily in less of a position to appropriately evaluate which methods are the most cost- and condition-effective; physicians also have the facility to select which devices will be used for a procedure or treatment. But as companies throughout these industries continuously increase their marketing and advertising expenditures, there is a growing concern from many corners about whether industry and company representatives commonly influence physician prescribing behavior. What has been hotly debated, in particular, is the extent to which the objectivity of medical professionals, especially physicians, is compromised by industry- and company-funded perks.

Professional relationships with pharmaceutical and device-manufacturing companies raise the specter of potential conflicts of interest. But there is an argument that to completely disconnect the physician from either industry would be to limit the physician's access to information about useful treatment options, despite the fact that therapeutic research trials and journal articles are often funded by the pharmaceutical or device company that

stands to gain financial reward for the respective innovation. Physicians, by nature of their work, tend to concern themselves a great deal with accruing extensive knowledge about the benefits, contraindications, and costs of various pharmaceutical therapies and medical devices. To keep current and alert is a necessary practice if only to uphold the implicit directive of the Hippocratic Oath: to respect, and to do no harm unto, the patient.

Respect for the patient means keeping the patient actively involved in the decisions about care and treatment of his or her condition. This requires that the patient be well informed about care and treatment options, which is a task that has traditionally and fundamentally been the responsibility of the physician. Yet as changes have occurred in and around the doctor–patient relationship and throughout society, so too does the patient now share a mutual responsibility for the dimensions of healthcare decision making.

Importance

Above all, medicine is a practice. It essentially concerns the function and performance of human beings—and all human beings are fallible. But the impressive advancements of modern medical science and technology often render this an afterthought in light of presumably informed decisions and expected outcomes of healthcare.

Today as ever, because healthcare decisions emanate from individual values, feelings, and behaviors in the context of a rapidly changing society, economy, and polity, there is an increasing demand for approachable and more transparent exchanges of ideas and information between physicians and patients. This encourages practitioners and patients to learn from each other. Yet it requires that the right time and space conditions be established to enable cooperative decision making by individuals who posses diverse personality types. It also requires that people become clear about who will be involved in a healthcare decision and to what extent they intend to be involved during the course of care. And there is an additional and vital need for a better understanding of the processes through which people gather information and how they use that information to make healthcare decisions.

Lee H. Igel

See also Cognitive Psychology and Processes; Informed Consent

Further Readings

Alpert, J. S. (2008). Doctors and the drug industry: Further thoughts for dealing with potential conflicts of interest? *American Journal of Medicine, 121*(4), 253–255.

Braddock, C. H., III, Edwards, K. A., Hasenberg, N. M., Laidley, T. L., & Levinson, W. (1999). Informed decision making in outpatient practice: Time to get back to basics. *Journal of the American Medical Association, 282*(24), 2313–2320.

Conrad, P. (2007). *The medicalization of society: On the transformation of human conditions into treatable disorders*. Baltimore: Johns Hopkins University Press.

Ellis, A. (2000). Rational emotive behavior therapy. In R. Corsini & D. Wedding (Eds.), *Current psychotherapies* (6th ed., pp. 168–204). Itasca, IL: F. E. Peacock.

Epictetus. (1955). *Enchiridion* (G. Long, Trans.). Lancashire, UK: Promethean Press. (Original work published 1960)

Heritage, J., & Maynard, D. W. (Eds.). (2006). *Communication in medical care: Interaction between primary care physicians and patients*. Cambridge, UK: Cambridge University Press.

Kaplan, R. M., & Frosch, D. L. (2005). Decision making in medicine and health care. *Annual Review of Clinical Psychology, 1,* 525–556.

Kravitz, R. L., & Melnikow, J. (2001). Engaging patients in medical decision making. *British Medical Journal, 323,* 584–585.

Linzer, M., Konrad, T. R., Douglas, J., McMurray, J. E., Pathman, D. E., Williams, E. S., et al. (2000). Managed care, time pressure, and physician job satisfaction: Results from the physician worklife study. *Journal of General Internal Medicine, 15*(7), 441–450.

Montgomery, A. A., & Fahey, T. (2001). How do patients' treatment preferences compare with those of clinicians? *Quality in Health Care, 10,* i39–i43.

International Differences in Healthcare Systems

Healthcare systems can be distinguished on a number of dimensions and in a number of ways.

The need for care can be analyzed in terms of the burden of disease; here, developing countries differ considerably from industrialized nations. Healthcare incorporates a wide variety of types of service, whose delivery may be based on different combinations of providers, arranged in differing organizational structures. Financing these services may incorporate various combinations of public payment, social insurance, private insurance, and out-of-pocket payments. Reimbursement to providers may use various combinations of global budgets, fee-for-service, pay-for-performance, and capitation, which in turn imply differing incentive structures. System outcomes may be evaluated on a number of dimensions, including access, cost, quality, patient satisfaction, provider satisfaction, and health outcomes. No two systems are identical, and no system excels on all dimensions; trade-offs are inevitable.

Need for Care

The Burden of Disease

As the World Health Organization has noted, there are major differences in the causes of death. In a fact sheet, they estimate that for 2002, in low-income countries, nearly one third of deaths occur among children under 14 and less than one quarter of people reach the age of 70. Infectious diseases and the complications of pregnancy and childbirth are the leading causes of death. In high-income countries, more than two thirds live past age 70, and they tend to die of chronic diseases. Middle-income countries fall in between. However, chronic diseases (cardiovascular diseases in particular) are the leading causes of death across all categories of nations. The healthcare systems must accordingly determine what resources are required to meet the health needs of their populations.

An ongoing difficulty is that health expenditures are highly skewed. In general, the sickest 20% of the population accounts for about 80% of health expenditures, and this persists in all age-sex categories. In turn, this distribution of health expenditures affects the implication of various funding models, particularly since those identified as being at high risk of incurring high expenditures are not desirable customers for insurers.

The Role of Public Health

The foundation of all healthcare systems is public health. Ensuring clean water, clean air, and communicable disease control is a vital starting point for preventing disease and disability. So is appropriate engineering (road traffic accidents are leading causes of death in low- and middle-income countries), availability of safe and healthy foods, avoidance of tobacco, and other interventions coming under the rubric of "health promotion/disease prevention." In that sense, death by chronic disease is a success story; it means that people are living long enough to be affected by these conditions.

Financing, Delivery, and Allocation

Healthcare systems have a number of components. Although different writers may use slightly different nomenclatures and break down these functions in slightly different ways, they all note the importance of distinguishing between how services are paid for—often termed *financing*—and how they are organized, managed, and provided—often called *delivery*. Healthcare systems may also explicitly incorporate other key elements such as planning, monitoring, and evaluating, or leave these to the workings of market forces.

The missing link connecting financing and delivery, which has sometimes been termed *allocation,* refers to the incentive structures set up to manage how funds will flow from those who pay for care to those who deliver it. Saltman and von Otter have placed these allocation approaches on a continuum. At one end, patients follow money; funders allocate budgets to providers and people seeking that kind of care must go to those providers. At the other end of their continuum, money follows patients; providers are paid only to the extent that they attract clients. Unfortunately for those wishing clear reform prescriptions, there is no one best allocation model that can simultaneously ensure cost control, client responsiveness, and delivery of high-quality appropriate care; instead, one is often faced with policy trade-offs. Allocation is usually tied to reimbursement mechanisms.

Although certain combinations are more common than others, in theory, these dimensions of health systems can be viewed separately. One can flow public funds to private delivery, and one can support

public delivery through private funds (e.g., user fees for publicly operated services). Similarly, both public and private funders can embed various incentive structures in their reimbursement mechanisms.

Delivery

Components of Healthcare

Healthcare contains a number of components, including but not restricted to the following: diagnosis; disease and injury prevention, particularly of common conditions (including health assessments, screening, and immunization); treatment and management of diseases and injuries (both episodic and chronic); emergency services; health promotion and patient education, including encouraging patients to take active roles in their own health; rehabilitation (of both episodic and chronic conditions); long-term/chronic care, including personal and community support services and even housing; counseling and reassurance; referrals to/coordination of care with other professionals (including with hospitals and specialist care, and public health); reproductive care, including birth and delivery; mental-health care; palliative/end-of-life care; healthy child development, including well-baby care; and the provision of services to the population as a whole (public health programs), including the report and control of contagious and other diseases and ensuring occupational and environmental health and safety. This care may be provided in a number of locations, including physicians' offices, clinics, hospitals, nursing homes, homes, and workplaces. It may be provided by a variety of providers, including but not restricted to physicians (both specialists and generalists), nurses, dentists, pharmacists, traditional healers, allied health workers, rehabilitation professionals, personal support workers, and informal caregivers. Healthcare is often subdivided into primary care (the first point of contact), chronic care, public health, and secondary (hospital)/tertiary/quaternary (specialized/highly specialized hospital) care. Although different countries may arrange these services in different ways, similarity in human physiology means that there is, in general, less variability on what is required to manage a specific health problem. Although there may be differences in the prevalence of particular health problems across jurisdictions, broken arms will be treated similarly, and professionals trained in one jurisdiction can often—albeit often with difficulty—work in others.

The Public–Private Mix

There is considerably more variability in how this care is structured across systems. As analysts of comparative policy stress, no two healthcare systems are identical, although they may share common characteristics. Delivery models can be classified on the basis of ownership structure as public (owned and operated by the state), private not-for-profit (NFP), and private for-profit (FP). Other classifications subdivide these to incorporate additional characteristics such as autonomy and budgeting arrangements, including the extent of reliance on market incentives. For example, Deber has modified the Organisation for Economic Co-operation and Development (OECD) classification to subdivide the private FP category into small business/entrepreneurs (FP/s) and investor-owned corporations (FP/c); Preker and Harding have subdivided the public sector into those managed by the health system, and those that, although still public, have more managerial independence. For example, although hospitals in the United Kingdom are publicly owned and managed and part of the National Health Service (NHS), more recently the government has allowed them to become self-managing "trusts" with more independence from central control. In contrast, Canada relied on private NFP hospitals, many of which were originally owned and operated by religious organizations and other charitable organizations. (Confusingly, they are commonly referred to as "public hospitals," although they were neither publicly owned nor publicly managed and were usually governed by an independent board of directors.)

In general, such NFP organizations will not be bound by the same financial or administrative requirements that bind the public sector and can also draw on volunteers and receive charitable contributions. They may also go bankrupt if they cannot raise sufficient revenues, although in practice this rarely occurs. NFP organizations are motivated by multiple objectives, rather than just the financial bottom line, and are the most common ownership structure for hospitals in most industrialized countries. (An additional complexity is the extent to

which the governments, as major funders, try to control these NFP bodies. Indeed, some of these organizations can be considered "quasi-public" in that they are legally private but heavily influenced by government. This applies in some Canadian provinces, which regionalized hospital services and replaced the formerly independent hospital boards by regional health authorities.)

Most provider-run organizations would be categorized as FP/s; these include physicians' offices, many physiotherapy clinics, and indeed most of the private hospitals in countries such as Germany. They differ from FP/c organizations because they are not under the requirement of providing a return on investment to their shareholders. In some jurisdictions, small businesses may be incorporated for tax purposes (e.g., physicians may be allowed to incorporate), but their "profits" go to those who provide the clinical services rather than to independent shareholder/investors. Finally, FP/c organizations have shareholders who expect a return on their investment, and evoke concerns as to potential conflict between the goal of providing high-quality care and the goal of running a successful business. Accordingly, although FP/c hospitals have a significant presence in the United States, and are found in Australia, they are uncommon in most industrialized countries.

In most countries, governments have a role in regulating healthcare providers, although this power is often delegated to health professionals. The rationale is to make sure that the public is protected by ensuring that providers are qualified and that provider organizations meet certain standards of quality. Even in systems claiming to encourage market forces, people are not allowed to declare themselves to be physicians without having satisfied a licensing body. Neither do most jurisdictions allow anyone to set up a private hospital without ensuring that certain regulatory standards are met.

Financing

The World Health Organization has estimated that in 2002 the total global expenditure for health, per person per year, was approximately US$639. This ranged from US$2.90 in Burundi to US$6,103 in the United States. Although 18% of the world's population lives in the OECD countries (which themselves include some middle-income nations), they accounted for 80% of healthcare spending. To avoid comparing apples with oranges, the rest of this entry concentrates on industrialized nations.

The OECD identifies four main types of funding for health services: public payment through taxation/general revenues, public/quasi-public payment through social insurance, private insurance, and direct out-of-pocket payments. Premiums may be risk rated (based on the expected costs of services required), or based on other factors, including age-sex, income, and/or employment status. Systems vary in the mix of funding approaches, which may vary across type of service, and/or category of client (e.g., governments may pay for people with particular diseases, in particular age groups, and/or with particular incomes, while leaving others to different forms of coverage). As noted above, those at high risk of incurring high expenditures are unattractive clients for private insurers, and hence more likely to end up needing public subsidy.

Types of Healthcare Systems and Ownership

The OECD classifies health systems on the basis of their approach to financing and delivery. Beveridge-type countries, such as the United Kingdom, Sweden, Denmark, and Finland, couple public financing (through taxation) with public delivery. (Earlier versions of the categorization also used the term *Semashko systems* to refer to the systems in the former Soviet Union; these resembled Beveridge systems, having public financing and public delivery by salaried providers, but have largely been succeeded by other models.)

Bismarckian systems, also referred to as social health insurance, use quasi-public funding; most (or all) of the population are required to purchase health insurance from designated third-party payers (often referred to as sickness funds), with employers often paying a share of the costs. However, these payments are not risk related, and the benefits that must be provided are often subject to government regulations. Providers in these systems tend to be private, albeit often not-for-profit. Examples of Bismarckian systems are Germany, Austria, the Netherlands, France, and Belgium.

Private insurance models have a minimal role for public payment or public delivery. Financing is largely through private insurance purchased by individuals or their employers. In these models, premiums can be risk rated, and universal coverage is not guaranteed. Delivery is private, usually a mix of NFP and FP. However, it should be noted that these models are by no means exhaustive, and indeed are insufficient for most analysis; they do not fully capture new models of delivery (e.g., public contracting), and omit countries such as Canada and Australia, which employ tax-based financing to pay for private delivery. Neither do they deal with relationships within federal systems; for example, they do not separate tax-based financing that is centralized at the national level from financing decentralized to subnational units. Finally, they do not allow for variability across subsectors; for example, systems may have public payment for some services, while leaving others to private insurance or out-of-pocket payments.

Differences in Values

These models have varying assumptions about the relative roles of government, charities, and individuals and their families. The literature about health policy is accordingly linked to theories of the welfare state and of rights and responsibilities. There are also variations in views about the role of market forces in determining the allocation of healthcare. To the extent that care is given on the basis of need, individuals cannot be priced out of the market, violating some of the key assumptions underlying economic theory. All systems are mixed ones, with certain populations and certain services falling within the public realm, and others being private. However, systems vary in their decisions about who (and what) should be publicly paid for as well as in the extent to which providers will be regulated.

Reimbursement

Another source of variation is the way in which providers are paid for their services. Payers may reimburse individual providers and provider organizations (e.g., hospitals, health maintenance organizations). They may use varying combinations of global budgets, fee-for-service, pay-for-performance, and capitation. In turn, these imply differing incentive structures. For example, fee-for-service encourages the provision of more services, whereas global budgets encourage providers to minimize their expenditures in order to remain within their budgets.

Outcomes Evaluation

System outcomes may be evaluated on a number of dimensions, including access, cost, quality, patient satisfaction, provider satisfaction, and health outcomes. Again, no single system appears optimal.

Recently, it has been recognized that health systems have a global dimension. This is most evident in dealing with infectious diseases, which can easily spread across national borders. Other issues include migration of the health workforce.

A number of international bodies collect comparable data and publish useful comparisons.

Decision Making

The way in which healthcare systems are organized and delivered in turn affects medical decision making in a number of ways. Systems constrain the available choices. Regulations may affect what treatments can be offered; drugs, as an obvious example, may need approval from regulatory agencies before being made available for sale. They may affect which providers are allowed to practice, in terms of both which professions are recognized (e.g., traditional healers) and which individuals are licensed within particular jurisdictions. Financing approaches affect who can afford care.

In addition, systems help decide who the decision maker will be. Global budgets place decisions about who will get what treatment in the hands of the provider organization. Fee schedules place a considerable amount of decision making in the hands of those who negotiate the fee schedules (usually, physician organizations and payers). Different systems will empower different decision makers and in turn affect how various factors (including costs, demands, and clinical evidence) will affect medical decision making.

Raisa Deber

See also Consumer-Directed Health Plans

Further Readings

Allin, S., Bankauskaite,V., Dubois, H., Figueras, J., Golna, C., Grosse-Tebbe, S., et al. (2005). *Snapshots of health systems*. World Health Organization on behalf of the European Observatory on Health Systems and Policies, Copenhagen, Denmark.

Blank, R. H., & Burau, V. (2004). *Comparative health policy*. New York: Palgrave Macmillan.

Colombo, F., & Tapay, N. (2004). *Private health insurance in OECD countries: The benefits and costs for individuals and health systems* (OECD Health Working Papers, No. 15). Paris: OECD.

Commonwealth Fund: http://www.commonwealthfund.org

Davis, K., Schoen, C., Schoenbaum, S. C., Doty, M. M., Holmgren, A. L., Kriss, J. L., et al. (2007, May). *Mirror, mirror on the wall: An international update on the comparative performance of American health care*. New York: The Commonwealth Fund.

Deber, R. (2004). Delivering health care services: Public, not-for-profit, or private? In G. P. Marchildon, T. McIntosh, & P.-G. Forest (Eds.), *The fiscal sustainability of health care in Canada: Romanow papers* (Vol. 1, pp. 233–296). Toronto, Ontario, Canada: University of Toronto Press.

Docteur, E., & Oxley, H. (2003). *Health-care systems: Lessons from the reform experience* (OECD Health Working Papers, No. 9). Paris: OECD Publishing.

European Observatory on Health Systems and Policies: http://www.euro.who.int

Health Policy Monitor: http://www.hpm.org/index.jsp

Organisation for Economic Co-operation and Development: http://www.oecd.org/health

Organisation for Economic Co-operation and Development. (2004). *Towards high-performing health systems: Summary report*. Paris: OECD Health Project.

Preker, A. S., & Harding, A. (Eds.). (2003). *Innovations in health service delivery: The corporatization of public hospitals* (1st ed.). Washington, DC: The World Bank.

Saltman, R. B., Busse, R., & Figueras, J. (Eds.). (2004). *Social health insurance systems in Western Europe*. Buckingham, UK: Open University Press.

Saltman, R. B., & von Otter, C. (1992). Planned markets and public competition: Strategic reform in Northern European Health Systems. Philadelphia: Open University Press.

World Health Organization. (2007). *Fact sheet: Spending on health: A global overview*. Geneva: Author. Retrieved January 16, 2009, from http://www.who.int/mediacentre/factsheets/fs319.pdf

World Health Organization. (2007). *Fact sheet: The top ten causes of death*. Geneva: Author. Retrieved January 16, 2009, from http://www.who.int/mediacentre/factsheets/fs310.pdf

World Health Organization. (2007). *The world health report 2007: A safer future: Global public health security in the 21st century*. Geneva: Author.

Intraclass Correlation Coefficient

The intraclass correlation coefficient (ICC) measures the correlation of responses within class when responses are grouped into classes or groups, and there are a number of classes or groups. The ICC quantifies the variation between the clusters and can be defined as a proportion of total variation that is attributed to differences between the clusters (groups). Variation of a quantity is spread around its mean measured mainly by variance and standard deviation. Many situations in decision-making science make use of the ICC for drawing inference and assessing reliability. The ICC is used in a variety of situations for different purposes, including assessing homogeneity of outcomes or responses within a class or cluster in the context of a cluster survey, group randomized trial and multilevel studies, interrater agreement, and similarity of health/social responses within couples. Accounting for correlation of responses within group in a cluster or group randomized trial has important implications in terms of required sample size and statistical significance. Assessing agreement of raters on health states has been an important aspect of clinical decision making.

Thus, the ICC is used in studies involving correlation of responses within groups, assessment of interrater reliability, and similarity of responses in the dyads where dyad members are exchangeable. This entry provides an overview of the use of the ICC in studies involving groups or clusters with an example, followed by the use of the ICC in assessing rater agreements in reliability assessment and special cases of interrater reliability.

Studies Involving Clustering

Health outcomes of individuals in the same household or community tend to be more similar than those of individuals in other households and communities. This phenomenon may be due to similar levels of exposures, similar behaviors, or genetic predispositions. Due to this similarity, individuals in a group sharing some characteristics are unlikely to be independent with respect to their health outcomes, since responses of the individuals in a group show positive intraclass correlation. For example, transmission of an infectious agent or exposure to air pollution with its related health outcomes such as asthma will be more common among people within a particular community than in other communities due to different levels of exposures.

In developing countries, sampling frames are often not available for epidemiological surveys; the cluster sampling technique is recommended due to its logistic efficiency. Another advantage of cluster sampling is the comparative ease in enumerating groups of households or larger units such as census block, county, village, and so on, as clusters than as individuals. In studies at places where responses are naturally clustered, such as patients in general practices (patients within practitioners) or worksites (workers within worksites), a similar sampling scheme is applied. Similarly, in community trials, the unit of randomization is a group of people rather than an individual because intervention is applied to all people in a group that decreases the risk of contamination and increases the administrative efficiency. In all these situations involving some kind of clustering, there are two components of variation of responses: within-cluster/group variation and between-cluster/group variation. The ICC can be used to quantify and account for these two components of variations.

Between-clusters variation causes inflation of error variance. In the presence of positive intraclass correlation, application of standard statistical methods for calculating sample size (assuming no clustering) and statistical analysis later on will underestimate the error variance. This needs to be accounted for at the time of designing a study; otherwise, it reduces the power to give desired results. To have adequate power, the sample size has to be increased by using a design effect or variance inflation factor as described by Kish and Donner.

Kish described the design effect (D) as a function of average cluster size ($\bar{n}$) and ICC (ρ):

$$D = 1 + (\bar{n} - 1)\rho.$$

The ICC quantifies the variation between the clusters and can be defined as the proportion of total variation that is attributed to differences between the clusters or groups:

$$\rho = \alpha_b^2 / (\alpha_b^2 + \alpha_w^2),$$

where α_b^2 is the between-cluster component of variance and α_w^2 is the within-cluster component of variance.

For analysis of cluster studies, investigators should use methods that take into account clustering of responses. If researchers ignore clustering and analyze assuming independence of health outcomes in a group, they will incorrectly get large test statistics and small p values because they have ignored one component of variation between groups. The methods for analysis of cluster data that have been recommended to be more appropriate are those that simultaneously adjust for cluster- and individual-level covariates. There are a variety of methods, generally called multilevel, hierarchical, or random-effects methods. They are based on different types of statistical models, such as the generalized linear, mixed model, generalized estimating equations, and hierarchical Bayesian model.

Some examples of clustering include dental studies involving multiple teeth from an individual, studies where environmental exposures are measured at the area level for a group of individuals, community randomized trials for evaluation of health education intervention, studies on patients presenting at primary care facilities, and studies involving geographic clusters.

Example

Naveed Zafar Janjua conducted a study in 16 villages of a periurban setting in Karachi, Pakistan, to estimate the prevalence of hepatitis C virus (HCV) infection and to identify risk factors. The cluster sampling technique was used, in which clusters were villages. From within villages, households were selected using systematic sampling with random start. Previous studies indicate that reuse

of injection equipment by healthcare providers is one of the major risk factors. Furthermore, health outcomes (HCV infection in this case) are expected to be similar among populations presenting to the same healthcare provider. Reuse of injection equipment depends on the injection use behavior of the healthcare provider, which is more likely to be similar for patients in the same geographic area. Thus, clustering of HCV needs to be accounted for in sample size estimation and analysis.

To assess the relationship of injection use and HCV infection, information on injection use during 6 months was collected. Injections were categorized into greater than or equal to 5 and less than 5 during the last 6 months based on initial assessment of their relationship with HCV infection status. Logistic regression was performed by taking HCV infection as a dependent variable and injection use as a dichotomous independent variable. The same analysis was repeated, accounting for the correlated nature of responses in the cluster using the generalized estimating equation technique. Results for injection use and transfusions are presented in Table 1.

Results show that point estimates (odds ratio) in both instances, when accounting for clustering and when not accounting for clustering, were the same. However, the major difference was in the test statistics, *p* value, and confidence interval. Test statistics are smaller, *p* values are larger, and confidence intervals are wider when accounting for clustering. This is because when there is clustering, total variance is more than when there is no clustering. Thus, in situations that involve clustering, correlation of responses within a cluster should be accounted for during analysis.

Reliability Analysis

The ICC is also used to measure interrater reliability for two or more raters or judges when data are on an interval or ratio scale, such as a score. The ICC assesses rating reliability by comparing the variability of different ratings of the same subject with the total variation across all ratings and all subjects. This type of assessment was initially proposed by Shrout and Fleiss in 1979. They proposed three different types of ICCs depending on sampling design and underlying intent.

In decision analysis and medical and behavioral sciences research, the main interest is interrater

Table 1 Comparison of two techniques for estimation of odds ratio for association of injection use and HCV infection in Karachi, Pakistan

	Accounting for Clustering[a]				*Ignoring Clustering*[b]			
Variable	*Odds Ratio*	*Confidence Limits*	*Chi-Square*	*p*	*Odds Ratio*	*Confidence Limits*	*Chi-Square*	*p*
Injection received during past 6 months								
≤5	1.00							
>5	1.48	(1.09– 2.03)	6.14	.0132	1.48	(1.17– 1.88)	10.36	.0013
Number of transfusions received								
0	1.00				1.00			
1	1.78	(1.03– 3.08)	4.22	.0399	1.78	(1.08– 2.92)	5.21	.0225
≥2	3.50	(1.92– 6.36)	16.85	<.0001	3.55	(2.07– 6.08)	21.13	<.0001

Source: Unpublished data from Hepatitis C Infection Investigation in a periurban community in Karachi, Pakistan.

a. Generalized estimating equation. b. Logistic regression.

reliability or, in other words, interchangeability or replaceability of the rater. In simpler terms, researchers want to assess that all raters are in good agreement with each other so that they can interchange or replace any rater or use any rater from the pool without any bias. This is the case for random samples of objects or subjects from a large pool and a random sample of assessors or raters from the large pool. This means that raters used in the research are not the only raters and the researchers can replace one rater with another rater. In statistical terms, this is a completely randomized two-way analysis of variance (ANOVA) design. This type of ICC has been termed an ICC (2, 1) by Shrout and Fleiss.

Interpretation

An ICC ranges from 0 to 1. It reaches 1 when all raters give the same ratings—perfect interrater reliability. In such a case, all variation is due to characteristics of the subject or patient. Similar to other reliability coefficients, an ICC of .7 is considered adequate.

Limitations

An ICC is population specific. An estimate from one population may not be comparable with one from another because underlying variability, which is the key component in ICC estimation, is characteristic of population. It may vary from one population to another. Thus, the same instrument judged reliable for one population may be unreliable for another.

Example

In 2006, Quintana and colleagues conducted a study to develop decision-making criteria for cataract extraction through phacoemulsification in Spain. Investigators listed the indications through a literature review. A national panel of ophthalmologists (doers and nondoers of cataract extraction) recognized in the field was compiled. Investigators provided literature review and the list of indications to the panelists, and the panelists rated each indication for the appropriateness of performing phacoemulsification, considering the average patient and average physician in the year 2004. Appropriateness was defined as meaning that the "expected health benefit exceeds the expected negative consequences by a sufficiently wide margin to make cataract surgery worth performing."

An ICC was computed to assess the reliability of the 12 panelists' scores. Results revealed that the ICC among the 12 panelists was .69. Investigators concluded that their panel of experts showed acceptable agreement. The study included further analysis using Classification and Regression Trees (CART) for the creation of a decision tree.

Studies Involving Dyads

One type of ICC—the pairwise ICC—can be used to measure homogeneity or similarity of the responses among dyads, couples or pairs that cannot be classified into separate classes, such as twins, gay couples, and so on. For example, if a researcher is studying the weight of gay partners or twins, he or she may not know which one should go in which column. In such situations, members of the group are from the same class and are exchangeable. When members are from the same class, there is no natural way to classify them into separate columns. Pearson correlation computation requires assignment to two separate groups or classes, which is not appropriate. If the assignment of one or more pairs is reversed, the Pearson correlation will change. In such situations, an ICC can be computed as a measure of likeability or homogeneity *within* the pairs.

An ICC for twin or couple studies can be computed by using a one-way ANOVA design because there is only one dimension, couple/twin. This ICC with one-way ANOVA is available in SPSS and can also be computed using SAS. In SPSS, this statistic can be requested using "Reliability Analysis" and requesting the "One-way Random" option in statistics options.

Naveed Zafar Janjua

See also Diagnostic Process, Making a Diagnosis; Health Status Measurement, Reliability and Internal Consistency; Variance and Covariance

Further Readings

Donner, A., & Klar, N. (2000). *Design and analysis of cluster randomized trials in health research*. London: Arnold Press.

Griffin, D., & Gonzalez, R. (1995). Correlational analysis of dyad-level data in the exchangeable case. *Psychological Bulletin, 118*(3), 430–439.

McGraw, K. O., & Wong, S. P. (1996). Forming inferences about some intraclass correlations. *Psychological Methods, 1,* 30–46.

Shrout, P. E. (1998). Measurement reliability and agreement in psychiatry. *Statistical Methods in Medical Research, 7*(3), 301–317.

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin, 86,* 420–427.

Uebersax, J. (2006). *Intraclass correlation and related methods.* Retrieved April 6, 2008, from http://ourworld.compuserve.com/homepages/jsuebersax/icc.htm

Ukoumunne, O. C., Gulliford, M. C., Chinn, S., Sterne, J. A., & Burney, P. G. (1999). Methods for evaluating area-wide and organisation-based interventions in health and health care: A systematic review. *Health Technology Assessment, 3*(5), iii–92.

Intuition Versus Analysis

Intuition and analysis are distinct modes of making judgments and decisions, and the two have been contrasted at least as far back as ancient Greece. Intuition and analysis are broad concepts without clear definitions. Nevertheless, there is some consensus about their respective characteristics, advantages, disadvantages, bases, and mechanisms. Discussions that contrast intuition with analysis in medical decision making have traditionally been set in the context of expert diagnostic judgment. These discussions have been both descriptive (i.e., about how clinicians actually arrive at diagnoses) and prescriptive (i.e., how clinicians should arrive at diagnoses). More recent discussions of intuition and analysis have considered their roles in patient decision making (e.g., treatment decisions) and healthcare management and policy decisions. This entry consists of three parts. The first part consists of an overview of the characteristics of and bases for intuitions and analysis. The second part briefly summarizes the intuition versus analysis debate. The third part summarizes and contextualizes some of the suggestions about when intuition is likely to be more advantageous and when analysis is.

Characteristics and Bases

Intuition

Intuition, often referred to as a *gut feeling* or *hunch*, is characterized broadly as an immediate (i.e., fast) and effortless judgment or decision process. That said, a given problem can lead one to have conflicting intuitions (e.g., personal vs. professional), which suggests that although the intuitions themselves may come quickly, the final judgment or decision may not. Intuitions have two components: the content and a feeling about the content. The content can range from a vague sense that something is wrong, to a diagnosis, a perception of risk, or a prediction about a future affective state (e.g., how one will feel about being blind if the treatment fails). The feeling about the content is a sense of rightness or familiarity, or of fitting and making sense. The strength of this feeling, in turn, has been found to moderate the degree of confidence one has in that intuition (e.g., a "strong hunch" vs. a "sneaking suspicion"). Generally speaking, the greater one's confidence in the intuition (whether it be right or wrong), the more likely one is to believe and follow it, and the less one is open to evidence that contradicts the intuition.

Intuitions are also characterized as an opaque process, that is, in most cases one does not know precisely why one has a particular intuition. The result is that although the intuition itself can usually be communicated, the cues and steps that lead to the intuition cannot. This means that intuitions are closed to scrutiny (though subsequent reflection on the intuition has been shown to sometimes shed light on what might have led to it).

One basis of intuition is experience that has been internalized into tacit knowledge. Because the same event can be perceived, interpreted, and encoded differently by different people, some researchers have pointed out that although more (repeated) experience will lead to the development of intuitions, it will not necessarily lead to *expert* intuitions. The upshot is that the strength of an intuition does not necessarily reflect its truthfulness or accuracy. The basic mechanism of intuition is believed to involve pattern recognition, though several researchers have pointed out that some forms of intuition (e.g., insights) have a less constrained and more creative associative mechanism.

Analysis

Analysis, in contrast, is characterized broadly as a deliberate and conscious judgment or decision process. It is a multistep process, where the progression from one step to the next is systematic, linear, and principled insofar as it is based on the search for and application of appropriate rules, logic, or argument. The principles are transparent and the steps traceable, making the process inherently communicable, and thus open to scrutiny. There appears to be some disagreement about whether analysis is a fast or slow process, but a comprehensive perspective suggests that the speed of analysis depends on several factors, including, but not limited to, who is doing the analysis (e.g., novice, expert, or computer), and how complex the analysis is (e.g., the quick application of a clinical prediction rule vs. a lengthy cost-benefit analysis of implementing electronic health records in a hospital system).

The basis of an analysis—what defines and constrains it—are the principles that govern it. These can be broad (e.g., maxims of statistics) or narrow (e.g., properly plugging in the cue values in a clinical prediction rule), formal (e.g., linear weighted decision analytic models) or informal (e.g., listing the pros and cons of a treatment option).

Intuition Versus Analysis

Outside medicine, the debate surrounding intuition and analysis (more often referred to as "reason" or "reasoning") as distinct ways of knowing has a long philosophical history, some of which is characterized by one being pitted against the other. The debate within medicine, however, has largely centered on the role that intuition and analysis do and ought to play in expert clinical judgment, specifically diagnostic judgment. The cognitive revolution in psychology, with its paradigm of mind as information processor, led to the development of linear, multistep, hypothetico-deductive models of diagnostic reasoning. These multistep descriptive models of the diagnostic process typically have as their first step the collection of data or cues, which leads to a differential diagnosis, and then on to a final diagnosis. Eventually, research suggested that although novices (e.g., medical students) labored through an analytic process, expert diagnosticians usually did not. Rather, experts were more selective in the cues to which they attended, were more recursive in their process, and made quicker judgments.

The discovery that expert physicians and nurses were not analytic diagnosticians led to a reaction, still very much alive and especially prevalent in nursing, against teaching diagnosis as a linear, analytic process. The argument is, roughly, that because there is evidence that experts make diagnoses holistically and intuitively, intuition is a legitimate mode of diagnosis and its development needs to be incorporated into medical education curricula. The counterargument is that clinical practice ought to be based on science, not on idiosyncratic clinician intuitions that are subject to biases and distortions. The flames of this debate are further fueled by allegiances to either side of the related controversy surrounding the growing prominence of evidence-based medicine. Many supporters of intuition are also critics of evidence-based medicine. They argue that its prescriptions are based on mean results, from potentially irrelevant samples, which ignore a host of important contextual factors. Their remedy to what they call "cookbook medicine" is the contextually sensitive, holistic judgment (i.e., intuition) of the clinician. In contrast, some proponents of evidence-based medicine point out that, from its inception, evidence-based medicine has explicitly called for clinicians to exercise their judgment when applying guidelines. This is to ensure that clinicians can accommodate cases that merit the modification or rejection of guidelines (i.e., contextual sensitivity). They point out, however, that intuition has its problems. Chief among them is that there are important breakthroughs in medicine that would lead to better patient outcome if implemented, but they are resisted, despite supporting evidence, because they are counterintuitive and require conceptual change (e.g., the acceptance of the bacterial theory of ulcers).

These proponents of evidence-based medicine seem to be part of a third group of researchers, one that argues that both intuition and analysis are key to diagnostic reasoning and to clinical practice more broadly. They contend that to advocate one process over the other (i.e., intuition vs. analysis) is both descriptively inaccurate and prescriptively counterproductive. The solution they propose is to leverage the strengths of each process by knowing when to rely on analysis and when to rely on

clinical intuition. Some proponents in this third group have characterized intuition and analysis as lying on two ends of a continuum and have held that the solution lies in the middle ground. Evidence in the dual-process theory of cognition literature, however, suggests that intuition and analysis (or analytic thinking) involve separate cognitive systems, and so the solution would be characterized as a better coordination of the two.

Intuition and Analysis

The question of when we tend to use intuition and when we tend to use analysis is a complicated one as many factors play into the equation, including personal, contextual, and task factors. Personal factors including personality variables, age, and culture, as well as one's beliefs about intuition and analysis, can all play a role. Individual differences in the disposition to engage in intuitive or analytic thinking also play a role—some people tend to rely on their intuitions more than do others. Finally, the level of one's experience and expertise in an area, and one's knowledge and acceptance of analytic procedures make a difference as well. There are also several contextual factors that affect whether one relies on intuition or analysis, some of which are believed to interact with personal factors. Time, stress, emotion, and fatigue all play a role, as do the presence or absence of accountability and the need to communicate (e.g., in teamwork), as well as other factors such as whether one places a premium on speed or accuracy and the perceived novelty of the situation. The third set of factors has to do with the task itself, and includes things such as the number and complexity of cues as well as the mode of their presentation.

Making justified and informative prescriptions about how to combine or coordinate intuition and analysis is complicated because one must consider both when each tends to be used, and when each tends to yield better patient outcomes. This is further complicated by the fact that not everybody will measure the quality of outcomes in the same fashion. For example, satisfaction with the judgment or decision process is one typical target outcome. Although intuitive judgments tend to be more procedurally satisfying than are analytically derived ones, this is not always true for individuals more disposed to analytical thinking. Furthermore, intuitive judgments, even though they may feel better to make, can sometimes be worse than are analytically derived decisions in terms of other outcomes, such as health.

In the end, there is growing evidence and consensus that if better patient outcomes are the overall goal, then the idea of "intuition versus analysis" needs to be reframed in terms of "intuition and analysis."

Georges Potworowski

See also Biases in Human Prediction; Bounded Rationality and Emotions; Clinical Algorithms and Practice Guidelines; Decision Rules; Diagnostic Process, Making a Diagnosis; Dual-Process Theory; Errors in Clinical Reasoning; Judgment

Further Readings

Benner, P., & Tanner, C. (1987). Clinical judgment: How expert nurses use intuition. *American Journal of Nursing, 87,* 23–31.

English, I. (1993). Intuition as a function of the expert nurse: A critique of Benner's novice to expert model. *Journal of Advanced Nursing, 18,* 387–393.

Gilovich, T., Griffin, D., & Kahneman, D. (Eds.). (2002). *Heuristics and biases: The psychology of intuitive judgment*. Cambridge, UK: Cambridge University Press.

Greenhalgh, T. (2002). Intuition and evidence: Uneasy bedfellows? *British Journal of General Practice, 52,* 395–400.

Hamm, R. M. (1988). Clinical intuition and clinical analysis: Expertise and the cognitive continuum. In J. Dowey & A. Elstein (Eds.), *Professional judgement: A reader in clinical decision making* (pp. 78–109). Cambridge, UK: Cambridge University Press.

Myers, D. (2002). *Intuition: Its power and perils*. New Haven, CT: Yale University Press.

Ubel, P. A., & Loewenstein, G. (1997). The role of decision analysis in informed consent: Choosing between intuition and systematicity. *Social Science and Medicine, 44,* 647–656.

Irrational Persistence in Belief

In medicine, it is rational to base beliefs and practices on evidence. Belief regarding diagnosis or

treatment is established through the scientific method—at its best, randomized controlled trials assessing treatment effects, epidemiological measurements of risk, and objective assessments of the accuracy of tests and the impact of their use. The source of standards for evaluating medical practices is decision theory, as expressed in analyses that include best estimates of costs, event probabilities, and the utilities of health states and outcomes. When these sorts of evidence establish that a particular belief or practice is superior to others, one might expect the rational physician to immediately adopt the better practice and use it consistently. And yet it has been seen that physicians—each of them occasionally, perhaps, and some of them consistently—persist in beliefs that have been proven wrong or in suboptimal practice. This entry describes and assesses why physicians irrationally persist in beliefs and practices that are against the evidence.

It would be unreasonable to expect beliefs or practices to change immediately on publication of new evidence. No individual physician can monitor developments in the entire clinical scientific literature. Rather, a minority of physicians read about advances in areas they know, discuss the implications for practice, and communicate their conclusions through their own examples, verbal recommendations, and published summaries and guidelines. It takes time to translate research into practice, to work out its implications for particular clinical settings and disseminate these conclusions through opinion leaders to the rest of the community, and to spread these changes from multiple centers out to the periphery. Those whom a change has not yet reached are said to be out of date or ignorantly persisting. Those who are exposed to information that objective observers would say justifies a change, yet do not change, are irrationally persisting.

Psychological Processes

Several psychological processes can account for irrational persistence of incorrect beliefs and inadequate medical practices. They can be categorized as nonmotivational and motivational processes. First, many practices are executed automatically, so their original rationale is no longer accessible to the physician. Although the evidence may be stated in the physician's presence, or pass before the physician's eyes, it produces no change in behavior because its relevance is not recognized. Second, people may be motivated to persist in practices to avoid the work that change requires or to defend the sense that one is right. (Financial motivation is not addressed in this entry.) Some illustrative vignettes follow.

Automated Thinking

A physician has a detailed knowledge structure covering how to deal with type 2 diabetics. This illness script has variants for the patient's initial visit, early follow-up, established follow-up, and response to each type of crisis. It is so well learned that he has forgotten the reasons for the timing of each test, prescription, and bit of advice. Despite recent demonstrations that the tight control of these diabetics' blood pressure has as great a protective role in preventing adverse outcomes as the tight control of their blood sugar, the physician has maintained the focus on blood glucose control. Consequently, he was surprised when an audit revealed that while 60% of the type 2 diabetics in his practice have adequately controlled blood glucose, only 40% have adequately controlled blood pressure.

Defensive Motivation

A gastroenterologist takes pride in her ability to rapidly complete colonoscopic screenings for colorectal cancer, withdrawing the scope in 4 minutes on average despite the informal guidelines that suggest spending more time on the procedure. When a new study showed that those who take longer than 6 minutes are more likely to find advanced polyps, the physician tried to do the procedure more slowly but found it difficult. It felt wrong, as if she were dawdling or wasting time. She imagined the nurse commenting to colleagues that she was losing her touch. She did not notice that she found any more polyps when she did it slowly and speculated that she had superior perceptual abilities. After a week, she was performing the procedure as rapidly as ever.

Nonmotivational Bases

When a physician has seen new evidence that shows a belief or practice is incorrect or suboptimal, there

are two nonmotivational reasons the physician might continue to hold the belief. These differ in whether the physician recognizes the applicability of the new evidence. In the first case, the way in which the current belief depended on the old evidence is inaccessible to the physician; when the relevance of the new evidence is not recognized, the physician does not pay much attention to it. In the second case, even when the physician thinks about the new evidence and accepts that it is relevant to his or her practice, there are residual effects from having spent time believing the opposite.

Use of Knowledge Structures, Cultural Patterns, and Intellectual Artifacts

To elaborate the first example above, in which the physician did not adjust his management of diabetic patients, let us recognize that physicians' current rational behavior is not usually produced by the application of their rational facilities to the data of today's patient. To research the treatments and apply decision theory to the facts about each patient is beyond physicians' practical capability. It would take more time than they can afford just to look up the pertinent evidence, let alone to analyze that information to determine the best option. And many physicians may not know how to do that analysis.

Rather, the physician relies on the products of past reasoning. The physician in our vignette retained his own conclusions from earlier episodes in which he had thought about managing type 2 diabetics, experimenting with various treatment approaches, and referring to the primary scientific literature, meta-analytic reviews, or evidence-based guidelines. The basic structure of physicians' clinical knowledge is derived from textbooks and from their teachers' explanations, demonstrations, and corrections offered during supervision. Memorable learning experiences are provided by the critical reviews of cases in morbidity and mortality rounds. Textbooks, articles summarizing a clinical expert's approach to a particular clinical presentation, or Internet sites also influence the content of physicians' well-practiced scripts. The influence of others' knowledge continues when physicians seek advice from their colleagues about their current patients or compare their judgments against experts' conclusions in cases they read about.

Whether the original rational analysis supporting the now discredited practice was the physician's own or someone else's, the clinically relevant knowledge structure comes to mind through automatic pattern recognition, and the physician usually applies it with scant reflection. When the rationale is inaccessible in this way, one does not notice that it does not apply any more, and so the new evidence does not figure in determining how to manage the patient.

Seeking external knowledge about how to manage a patient is no guarantee one won't irrationally do something already proven to be inappropriate. The individuals one consults have likely trained in the same institutions and read the same literature. Non-evidence-based practices have inertia due to mutual social reinforcement. The power of this local medical culture is demonstrated by enduring differences observed in how the same condition is treated in different areas, differences that persist in the form of community standards even though all have access to the same evidence in the literature.

Finally, even the clinical literature can manifest this form of irrationality, providing physician readers with recommendations or statements of fact that have been previously disproven. Obviously, the published reference texts do not change when some of their contents have been proven wrong. Papers show up in a literature search, even if they have been superseded. Additionally, a physician could base a treatment on a recent publication, ignorant that its justification had been definitively contradicted years before. The clinical advice papers published soon after the new evidence comes out have been in the pipeline long enough that they may not cover the implications of the new facts.

Impact of Previously Held Beliefs Subsequent to Willing Belief Change

Having once believed something, one continues to hold that belief somewhat, even if one subsequently receives information that contradicts the belief and one fully believes the new information. When we learn something, we do work to elaborate the idea, to explain it, to link it with related ideas. The residual of that work is

still present in our knowledge base, even when we have subsequently learned that the original idea was not true. This mechanism is the purest expression of the phenomenon of irrational psychological persistence in belief. Because it is unmotivated, no amount of effort to increase the reward for physician rationality or the punishment for impurity of physician motivation can be guaranteed to eliminate this type of irrational persistence of discredited beliefs.

Motivational Processes

Two types of motivational process may be involved when a physician irrationally persists in believing ideas that have been discredited or in doing actions shown to be less than optimal. The first is simply the avoidance of the work of changing one's own habits, others' behavior, or an organization's functioning, when the physician knows something different should be done. The second type occurs when the physician is motivated, possibly unconsciously, to manipulate his or her own awareness to reduce the conflict felt when one's beliefs are contradicted.

The Avoidance of Unpleasant Work

Physicians' daily choices are made in the context of patients' expectations (e.g., patient expectation of an antibiotic for a cold or sinusitis), the standard operating procedures of the clinic, the third-party payer's policies, and the nursing staff's habits. The physician cannot make a change individually, without explaining it to colleagues and staff, requesting changes in institutional policies, and educating patients. It may seem easier simply to ignore the new evidence and continue practicing in the usual way. Physicians in a complex system may not actually know how to change the system, unless they have taken a special interest in the management processes. Research suggests that it takes simultaneous efforts on multiple fronts to change how physicians, staff, and patients think a disease should be treated. While one might view each physician as rational when he or she does not choose to lead such a change process, jointly all the physicians in the system can be considered irrational if they know they should treat patients in a different manner and yet they don't.

Defense of Cherished Beliefs

In the second vignette above, the physician, confronted with evidence that spending more time inspecting the colon makes colonoscopy a more accurate screen, could not see that the conclusion applied to her. It takes conscious effort to change one's own way of doing things. To be motivated to make this effort, it is necessary to believe the change is needed and feasible. At this juncture, physicians sometimes have distorted perceptions or illogical reasoning, with the unconscious motive of neutralizing the justification for change. Thus, the colonoscopist was convinced that she did not need to change because she did not notice that withdrawing her scope made her inspection any more accurate, where a statistical analysis (such as that in the published study) would require objective observation of many months of colonoscopies to detect a difference in the rate of polyp detection. Generally, physicians are unconsciously motivated to notice evidence that supports their current way of practice and not to notice evidence that contradicts it. When the evidence is ambiguous they interpret it as supporting their position.

Scientific Disagreement

This entry labeled as "irrational" persistence in a belief or practice when there is convincing evidence against it. Of course, the scientific process consists in just such disagreements. The new treatments that truly do save lives emerge from the garden of unproven and expensive treatments of the specialists. Fundamental advances start with the extreme minority opinions, as in the conception of the role of *Helicobacter pylori* in gastric ulcers or angiogenesis in cancer. One person's genius may be another person's irrational persistence in belief.

Final Thoughts

Disproven medical beliefs and practices persist irrationally for many reasons, from the individual physician's cognition to the inertia of medical systems and professional ideologies. In promoting rational medicine, it is useful to be aware of this variety, because the different causes require different corrective measures. Few of the mechanisms have been researched extensively in the medical domain. Without knowledge of which mechanism

of persistence is predominant in a situation, efforts to promote rational change may be misaimed.

Robert M. Hamm

See also Automatic Thinking; Bias in Scientific Studies; Bounded Rationality and Emotions; Clinical Algorithms and Practice Guidelines; Conflicts of Interest and Evidence-Based Clinical Medicine; Errors in Clinical Reasoning; Evidence-Based Medicine; Motivation

Further Readings

Abernathy, C. M., & Hamm, R. M. (1994). *Surgical scripts*. Philadelphia: Hanley & Belfus.

Anderson, C. A., Lepper, M. R., & Ross, L. (1980). Perseverance of social theories: The role of explanation in the persistence of discredited information. *Journal of Personality and Social Psychology, 39*(6), 1037–1049.

Baron, J. (1994). *Thinking and deciding* (2nd ed.). Cambridge, UK: Cambridge University Press.

Connelly, D. P., Rich, E. C., Curley, S. P., & Kelly, J. T. (1990). Knowledge resource preferences of family physicians. *Journal of Family Practice, 30*(3), 353–359.

Gonzales, R., Steiner, J. F., Lum, A., & Barrett, P. H., Jr. (1999). Decreasing antibiotic use in ambulatory practice: Impact of a multidimensional intervention on the treatment of uncomplicated acute bronchitis in adults. *Journal of the American Medical Association, 281*(16), 1512–1519.

Haynes, R. B., Sackett, D. L., Guyatt, G. H., & Tugwell, P. (2005). *Clinical epidemiology: How to do clinical practice research* (3rd ed.). Philadelphia: Lippincott Williams & Wilkins.

Mold, J. W., & Gregory, M. (2003). Best practices research. *Family Medicine, 35*, 131–134.

Shaughnessy, A. F., & Slawson, D. C. (2003). What happened to the valid POEMs? A survey of review articles on the treatment of type 2 diabetes. *British Medical Journal, 327*(7409), 266.

Tatsioni, A., Bonitsis, N. G., & Ioannidis, J. P. A. (2007). Persistence of contradicted claims in the literature. *Journal of the American Medical Association, 298,* 2517–2526.

Wennberg, J. E., Freeman, J. L., & Culp, W. J. (1987). Are hospital services rationed in New Haven or over-utilized in Boston? *Lancet, 1,* 1185–1189.

J

JUDGMENT

A judgment is an opinion as to what was, is, or will be some decision-significant state of the world, where a *decision* is a commitment to a course of action that is intended to serve the personal interests and values of particular people, for instance, a patient. The "decision-significant" part of the concept rests on two facts. The first is that the content of the judgment at least partly dictates the decision that is reached. The second is that the accuracy of the judgment imposes a ceiling on the extent to which the selected action really does serve the interests and values of the intended beneficiaries. The following are some judgment examples:

- *Was*: A pathologist concludes that a patient died of natural causes and therefore chooses to not ask the authorities to investigate a possible crime; if that conclusion is erroneous, a criminal would remain free to cause additional harm.
- *Is*: A pediatrician believes that a child's slow growth pattern is normal and thus declines to recommend hormone treatments; if that belief is incorrect, the window of opportunity for treating a hormone imbalance could be lost forever.
- *Will be*: A patient is convinced that a new herbal treatment would halt the progress of her cancer and hence decides to seek out that treatment; if that conviction is misguided, pursuing the new treatment could prove useless or, worse, preclude other, more effective treatment options.

Judgment Formats

Judgments appear in formats that are opposites in several dichotomies. The distinctions matter for several reasons: Judgments in different formats drive decisions in different ways; their accuracy must be appraised differently; and they rest on somewhat different psychological processes, with contrasting implications for judgment-training efforts.

Categorical Versus Quantitative Target

The character of the judgment target (the state of the world at issue) can be *categorical*, implying simple qualitative distinctions, as when a physician must make a differential diagnosis among several biologically disparate disease categories. Alternatively, the target might be inherently *quantitative*, corresponding to a point along some continuum, as when a physician tells a patient, "I would expect your recovery to take about 6 weeks."

Deterministic Versus Likelihood Assertions

In a *deterministic judgment*, the "judge," the person rendering that opinion, makes a flat-out, unqualified assertion about the target; for example, "You have early-stage breast cancer." In contrast, in a *likelihood judgment*, the judge qualifies the offered claim with an indication of associated chances; for example, "There are good odds that you have the disease, I'm afraid." Some people (e.g., patients) prefer that others (e.g., their doctors) provide them with deterministic rather than likelihood judgments, perhaps because such definitive

pronouncements seem more competent and reassuring in their clarity. Others, on the other hand, doubt the integrity and expertise of people who seemingly hide or fail to even recognize the uncertainty presumed to be present in most real-life medical situations. Also, recipients of judgments from other people often say that if there is doubt in their informants' minds (as there usually is), they want to know about it. This allows them to make trade-offs between uncertainty and value, in the spirit of technologies such as decision analysis. They further recognize that deterministic judgments force them to act as if those judgments were definitively true. For instance, an unqualified diagnosis of lung cancer implies that the patient must be treated as having lung cancer.

Verbal Versus Numerical Indications of Likelihood

The chances associated with likelihood judgments sometimes are articulated *verbally*, with ordinary words such as "remote," "likely," or "good." These words convey ordinal (sometimes called *qualitative*) differences in likelihood but not much more, at least not precisely. For example, one should be more surprised by the actual occurrence of a stroke given a previous indication of a "remote" chance rather than a "good" one. But how *much* more surprise is warranted is impossible to say.

Alternatively, chances can be expressed *numerically*. *Probability statements* are one form of numerical expression, as when a radiologist records an 80% chance that an image represents a tumor. Intermediate between the extremes of everyday words and probability statements are *likelihood scale ratings*. For instance, a protocol might require a diagnostician to express a degree of certainty of dementia by circling a number between 1 and 7, where the scale anchors are 1 = *definitely not dementia* and 7 = *definitely dementia*. Probability statements offer several advantages. They make it easy to precisely characterize and analyze judgment accuracy. And, unlike mere ratings, some probability values have standard, concrete interpretations. For example, when there are two alternatives, such as "dementia" and "not dementia," 50% should mean that one alternative is just as likely to be true as the other. Furthermore, in principle, at least, relationships among probability judgments should conform to the rules of probability theory. Opposing considerations such as these, however, some people prefer verbal expressions of likelihood because they allow for the acknowledgment of actual vagueness. They do not promise more precision than truly ambiguous circumstances justify.

Point Versus Interval Quantity Judgments

When the target is a quantity, the person could render a *point judgment*, a claim for a specific value; for example, "You should be recovered in 10 days, so there's no reason to reschedule your trip." Alternatively, the judgment might be expressed in *interval form*; for example, "Recovery should take between 1 and 2 weeks." Interval judgments seem both more comfortable and more realistic, since it is hard to imagine many point judgments being exactly on the mark. One can go even further, indicating explicitly *how sure* the judge is that the interval in question will capture the actual value of the target; for example, "I'm 90% sure that everything will be back to normal in 1 to 2 weeks." Such interval judgments coupled with probability statements are called *credible intervals*.

Other Uses of the Term *Judgment*

Confusingly, two other meanings for the term *judgment* besides the predominant one used here are fairly common. In legal contexts, such as malpractice lawsuits, the expression is frequently employed to describe a legal decision, as in "The judgment of the court is for the plaintiff." Within decision scholarship, the term is sometimes used to characterize a special kind of decision otherwise known as an *evaluation decision* or simply an *evaluation*. This is a person's indication of how much something is valued by that individual, a pronouncement that is not mere idle talk but that, instead, potentially can have significant consequences for that person. A patient's rating of treatment satisfaction is a good illustration, as is a supervisor's appraisal of a resident's performance.

Expressed Judgments Versus Underlying "True" Judgments

Systematic differences sometimes exist between what a person actually thinks and what that

person reports. A physician might believe that a patient's prognosis for the next 6 weeks is bleak. However, because she wants to avoid discouraging the patient, she keeps her opinion to herself or perhaps even deliberately misleads the patient with a more optimistic "white lie." To combat various incentives for purposely biasing reports of one's true judgments this way, researchers have developed procedures that provide offsetting incentives for candor. Studies suggest that providing diagnosticians with accuracy bonuses paid according to *proper scoring rules* applied to probabilistic diagnoses would be effective in achieving that goal.

Discrepancies between what a judge really believes and what that judge reports explicitly do not have to be the result of intentional deception. Research has revealed numerous instances of reliable *self-insight failures*, cases in which people appear to be honestly ignorant about their true thinking or feelings. That is one reason that, for many years, decision scholars have been reluctant to accept people's explicitly articulated judgments at face value. Instead, they have emphasized inferences of people's true opinions from their decisions in specially structured situations. One inference approach is sometimes referred to as a *Bayesian technique*, illustrated as follows.

Suppose that we seek a physician's true belief about the probability that a certain patient will be fully recovered from surgery in 2 months. Call that target event "Recovery." Imagine that the physician is offered a gamble denoted G_{Recovery} = [\$10, Recovery; \$0, Otherwise], which means that the physician receives \$10 if Recovery occurs but nothing if it does not. Also imagine another gamble represented by G_Q = [\$10, Blue; \$0, Red $|$ Q Blues, 1,000 – Q Reds], where the information after the symbol $|$ refers to a special kind of lottery. This gamble offers the same payoffs as G_{Recovery}. However, those payoffs are determined by a random drawing (to be performed 2 months hence) from an urn containing Q blue balls and 1,000 – Q red ones, where Q is some whole number between 0 and 1,000. The gamble pays \$10 if a blue ball is selected, otherwise nothing. The physician is shown many versions of G_Q varying according to Q. And each time, the physician is asked whether he or she prefers G_{Recovery} or G_Q, or instead is indifferent between them. Suppose that he or she is indifferent between G_{Recovery} and G_{650}. Then we must infer that the physician's "true" probability judgment that the patient will be recovered in 2 months is 65%. That is because, in the physician's eyes, everything about G_{Recovery} and G_{650} must be equivalent. This includes the chances of getting \$10, the only thing that could possibly have differed between G_{Recovery} and G_{650}. And every "reasonable person" would agree that the chance of drawing a blue ball from an urn with 650 blue ones and 350 red ones is 65%.

Note that the physician was never explicitly asked the difficult question: "What do you think is the probability of recovery?" Instead, he or she only had to make a series of decisions between pairs of transparently simple alternatives. In effect, the Bayesian procedure provides a window on the judgments that actually drive the physician's decisions, the true judgments that really matter from a decision-making perspective. They are not self-reports of internal opinions to which the physician might have poor cognitive access.

Judgment Sources

The judgments that inform medical decisions originate in several kinds of sources, including three major ones that entail significant distinguishing features: individuals, collectives, and devices.

Individuals

Some judgments come from individuals working alone. The means by which they arrive at those judgments are varied. In some instances, judges use *formalistic* procedures, which resemble (or are the same as) those that a statistician might employ. An especially simple example would be relying on existing records. Consider a primary care physician faced with a 50-year-old male patient with hypertension who has already suffered a stroke. Her judgment of his chances of suffering another stroke by age 60 informs her decisions about a management plan. The physician might easily adopt as her judgment the published rate of stroke recurrence for patients with the same characteristics. Other known individual judgment processes differ considerably. They include judgment according to similarities (e.g., between a given patient and a prototype for a given disease), the availability of particular instances (e.g., recent, memorable

patients), and action models or "stories" whereby alternative possibilities might come about (e.g., biological disease models that explain the emergence of the patient's signs and symptoms more or less adequately than do other models).

Collectives

Some judgments are derived from collectives rather than individuals. Sometimes collective judgments are the product of *interactive deliberations* among the people comprising that collective. For example, several physicians might use a face-to-face meeting to reach consensus on a patient diagnosis in an especially difficult case. In principle, such deliberations have the potential to yield far more accurate judgments than those reached by any one physician alone when each of the individuals participating in the discussion has expert knowledge about a different aspect of the problem. Studies have shown, however, that several factors mitigate against the full exploitation of such differential knowledge, including the tendency for deliberations to be dominated by what the participants know in common rather than their unique, specialized knowledge. Another way to arrive at collective judgments is *mechanically*. That is, the members of the collective might render their judgments independently and then have those opinions somehow aggregated into a single assessment for the whole group. For instance, a group of physicians' collective prognosis for a patient's recovery time could be computed as a simple average of their individual predictions. The accuracy of such averages is often surprisingly good.

Devices

Judgments are sometimes rendered by devices such as computer programs. The judgment rules built into these devices can rest on a variety of principles. The most popular are various kinds of linear statistical models. In "accuracy contests" between devices and human judges with access to identical facts—for example, the very same patient signs and symptoms—the devices have almost always won. A major reason is that, although human judgment can be highly unreliable, judgments from devices are usually perfectly consistent. However, an inherent advantage enjoyed by human judgments is that they do not have to rely on a predetermined, restricted array of facts. For instance, a physician typically can observe and request myriad facts about any given patient, even those not on a standardized checklist. Unfortunately, this potential asset is known to be a liability at times, too. Humans sometimes pay attention to information that is statistically worthless, or else they respond to diagnostic information improperly. Either way, their accuracy is actually worsened rather than improved.

Assessing and Analyzing Judgment Accuracy

When judgments are inaccurate, they lead doctors and patients to choose courses of action that leave patients worse off than otherwise. This implicates the importance of effectively assessing judgment accuracy, since studies have shown that even trained professionals can differ substantially in the quality of their judgments.

Judgments are said to be accurate to the degree that they exhibit good *external correspondence*—that is, there is a strong, reliable statistical relationship between the judgments and the actual states of the world that they are intended to anticipate. There are well-established measures of external correspondence for each of the judgment formats noted previously, for example, for deterministic and likelihood judgments for discrete categories and quantities. These measures can be used to inform the selection and compensation of judges as well as to determine the extent to which programs for improving judgment accuracy have succeeded or failed.

A surgeon who excels overall might be outstanding with respect to some aspects of surgical performance yet only average in terms of others. Properly recognized, the "average" aspects can be targeted for focused improvement efforts that eventually result in even better overall performance. Like surgical competence, judgment accuracy is not a unitary construct. That is why researchers have developed numerous techniques for decomposing measures of overall judgment accuracy into constituent elements or contributors. For example, the lens model and statistical methods from the forecasting literature are useful tools for analyzing point judgments for quantities, such as predictions of survival time. They permit conclusions about whether assessments are deficient because a diagnostician is unreliable, say, rather than because

she is consistently biased in one direction or another. Decompositions of commonly used accuracy scores for probability judgments (e.g., for "cancer" vs. "no cancer"), such as probability scores or Brier scores, provide similar windows on specific contributors to probability judgment deficiencies that might then be the focus of training efforts. ROC (receiver operating characteristic) analysis, grounded in signal detection theory, offers similar insights into the accuracy of likelihood scale ratings and even deterministic judgments.

J. Frank Yates and Lydia L. Chen

See also Brier Scores; Heuristics; Judgment Modes; Lens Model; Probability, Verbal Expressions of; Risk Perception; Social Judgment Theory

Further Readings

Cooksey, R. W. (1996). *Judgment analysis: Theory, methods, and applications.* San Diego, CA: Academic Press.

Hilden, J. (1991). The area under the ROC curve and its competitors. *Medical Decision Making, 11,* 95–101.

Larrick, R. P., & Soll, J. B. (2006). Intuitions about combining opinions: Misappreciation of the averaging principle. *Management Science, 52,* 111–127.

Larson, J. R., Jr., Christensen, C., Franz, T. M., & Abbot, A. S. (1998). Diagnosing groups: The pooling, management, and impact of shared and unshared case information in team-based medical decision making. *Journal of Personality and Social Psychology, 75*(1), 93–108.

Lee, J.-W., & Yates, J. F. (1992). How quantity judgment changes as the number of cues increases: An analytical framework and review. *Psychological Bulletin, 112,* 363–377.

Price, P. C., & Stone, E. R. (2004). Intuitive evaluation of likelihood judgment producers: Evidence for a confidence heuristic. *Journal of Behavioral Decision Making, 17*(1), 39–57.

Yates, J. F. (1990). *Judgment and decision making.* Englewood Cliffs, NJ: Prentice Hall.

Judgment Modes

A judgment is an opinion as to what was, is, or will be some decision-significant state of the world, for instance, a physician's conclusion that a patient has suffered a myocardial infarction. Judgment modes are qualitatively distinct means by which people arrive at their judgments. Contrast, say, a myocardial infarction diagnosis based solely on the physician's personal clinical experience with another derived from a validated formula applied to signs and symptoms on a checklist. This entry explains why judgment modes matter in medicine. It also describes major judgment modes that are especially useful to distinguish in medical practice. And, as appropriate, the entry further indicates specific practical implications of such distinctions.

Why Judgment Modes Matter

Judgments are important in medicine because their accuracy imposes a ceiling on the quality of the decisions they inform. That ceiling in turn sets bounds on the patient's well-being. A patient with severe chest muscle strain who is misdiagnosed as having had a myocardial infarction will be treated as a heart attack victim. This inappropriate treatment would be invasive and risky as well as needlessly expensive. Naturally, any physician or medical practice would like to minimize inaccurate judgments, be they diagnoses, prognoses, efficacy opinions, or any of the other myriad assessments that are required throughout every day in every clinic. Achieving that aim requires a deep understanding of precisely where those judgments originate. Such understanding makes it clearer how mistakes can occur and therefore what is sensible in efforts to prevent, correct, or compensate for them. If one actually misunderstands how particular medical judgments are achieved, then the resulting attempts to improve those assessments could easily backfire, making things worse. Assuming that judgments originate in procedures—that is, modes—that are fundamentally different from how they actually are generated is misunderstanding in the extreme.

A Judgment Mode Tree

Studies have shown that, as in most practical arenas, the judgments that support people's medical decisions can arise from sources as different from one another as apples and oranges, and hence the term *judgment modes* aptly describes those sources.

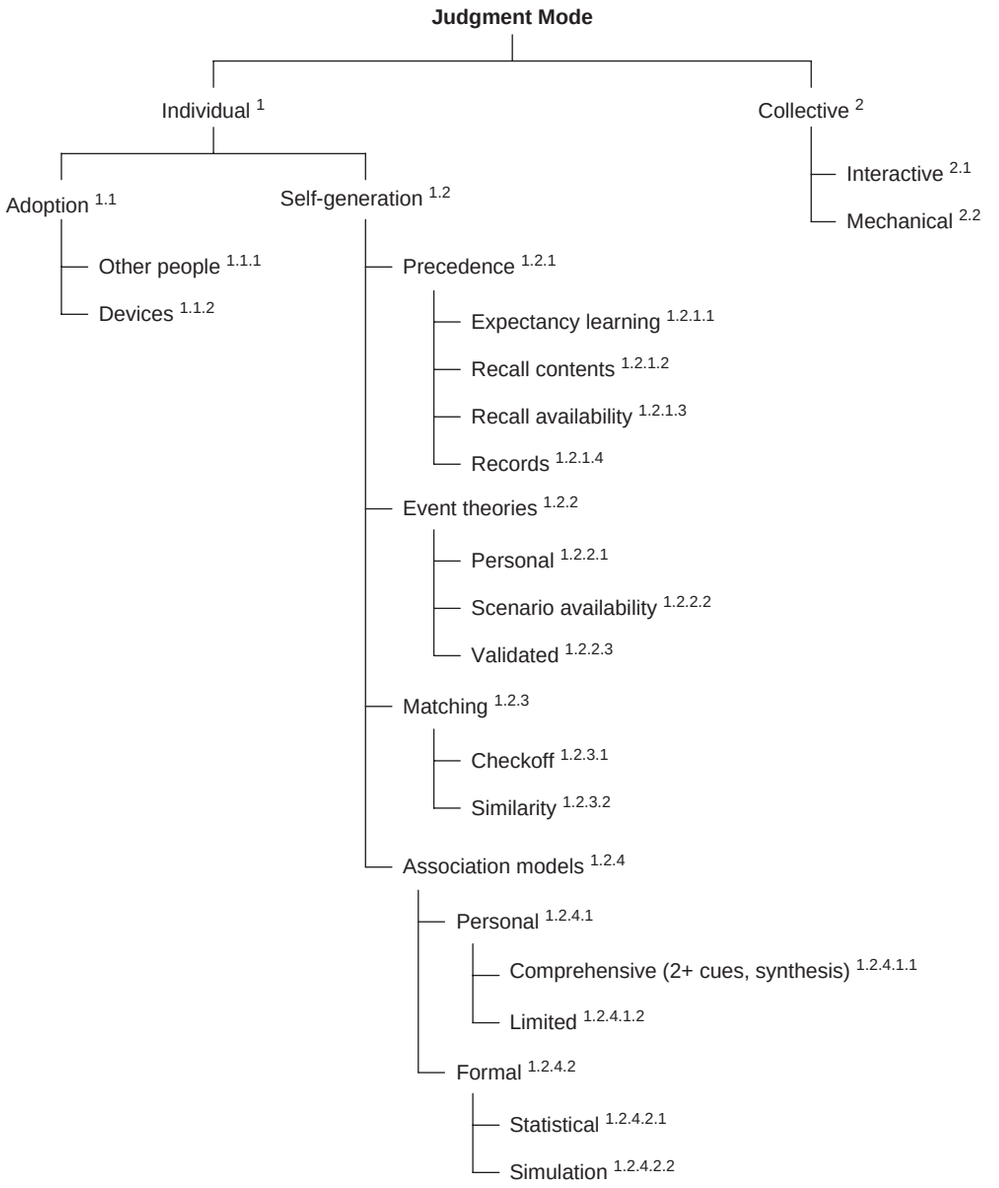


Figure 1 Judgment mode tree

Figure 1 shows a judgment mode tree. (The numbers on the nodes facilitate discussion.) This hierarchy is a taxonomy of major judgment modes, organized in a particular way. Specifically, there are reasons to expect the various modes to be invoked in roughly the order of a path from "northwest" to "southeast." For a given judgment problem, modes to the left and top of the tree are likely to be attempted before ones to the right and the bottom. It is important to bear in mind, however, that in a single decision episode, several different modes, applied one after another, easily might contribute to the judgment ultimately rendered.

Individual Versus Collective Modes

The first level of the tree distinguishes individual from collective modes. *Individual modes* (1) are those in which one person (e.g., the attending physician or, perhaps, a specialist whom the physician consults) provides the judgment in question. In contrast, in *collective modes* (2), the judgment is

supplied by several individuals working collaboratively in some manner (e.g., three physicians conferring to reach consensus on an especially challenging diagnosis). Individual modes normally have priority over collective modes if for no other reason than that people often work alone, sometimes by necessity. In addition, collective judgment is generally slower and more expensive, materially and emotionally; it entails higher process costs, such as the time and goodwill used up working through disagreements (e.g., between two physicians with opposing opinions about the true cause of a patient's complaint).

Individual Modes

The first major distinction among individual judgment modes concerns whether judgments are *adopted* (1.1) or are, instead, *self-generated* (1.2). Self-generated judgments are produced by the person who introduces them into deliberations for the decision problem at hand, whereas adopted judgments are acquired from some other source. Incidentally, it is common in decision scholarship to refer to the person who creates judgments for a given situation as the "judge." Consider, for instance, a physician deciding whether to apply standard heart attack treatment measures to a patient complaining of chest pain. If the physician makes the required diagnosis by herself, it is self-generated; she is the judge as well as the decider. But if she relies solely on the opinion of a colleague, the diagnosis is adopted.

Adoption Modes

Judgment via adoption is ubiquitous. It would be impossible for people to create personally all the judgments they need to make their decisions; they must depend on other sources. Sometimes those sources are humans, *other people* (1.1.1), as when a physician acquires the opinion of a pathologist who has examined a tissue sample. But at other times the sources are *devices* (1.1.2) of various kinds, such as computer programs that render probability assessments for potential diagnoses given presented signs and symptoms.

A decider who contemplates relying on adopted judgments ideally should resolve two practical concerns. The first is *accuracy*: How accurate would the adopted judgment be? For instance, how much more accurate than his own assessments should the physician expect the diagnoses of specialist Dr. Smith to be? The second concern is *cost*: How much would have to be paid—materially and otherwise—for the adopted judgment, and does the promised accuracy improvement outweigh the greater expense? For example, does the potentially improved accuracy of a new computerized diagnostic procedure more than offset the extra time and money it requires? There is reason to believe that both accuracy and cost concerns are often overlooked and that, when they are considered, they are not thought through adequately. Consider evidence that people's conclusions about others' expertise are strongly affected by factors (such as speech patterns) that easily can have little or nothing to do with objective accuracy indicators. Or take the fact that people often fail to ask questions about the functioning of judgment devices that have significant bearing on their appropriate use, such as questions about the information items the devices take into account and the items they ignore.

Self-Generation Modes

Under many conditions, although not all (e.g., where their inexperience is obvious), people tend to be overconfident about the quality of their own judgment. When that occurs, they should be expected to eschew adopted judgments for self-generated ones more than they should. And there are four main varieties of self-generation modes that might be pursued, each with its own special cases.

Precedence. The essence of *precedence modes* (1.2.1) is that the judge uses past occurrences in similar situations to inform the judgment needed presently. The most basic form this mode takes can be labeled *expectancy learning* (1.2.1.1). The core idea is the following: Consider some event of concern, say, a possible case of asthma. Furthermore, imagine that, in Clinical Setting 1, asthma occurs about 5% of the time, while in Setting 2, the rate is about 12%. The judge simply observes a large number of randomly presented cases (asthma vs. no asthma) in Setting 1 and does the same in Setting 2. There is no request to do anything like count, memorize, or even pay close attention to what is observed. Studies have shown that, in due

course, the judge will induce the fact that the asthma rates are different in the two contexts, as reflected in the judge's behavior.

Suppose the judge is told that, in a given setting, he will receive $100 if the next case that comes along is an asthma case. The judge is then allowed to choose the setting in which to exercise this opportunity. The judge will almost certainly pick Setting 2, where the asthma rate is 12%, rather than Setting 1, where that statistic is 5%; his expectancy for asthma is stronger in the former situation. Note that the judge was not asked to express a judgment explicitly. However, such a comparative likelihood judgment is implicit: "Asthma is more likely in Setting 2 than in Setting 1." If the judge were asked to state a probability judgment that the next patient in one of the settings will have asthma, there is no guarantee that that judgment would match the observed rate precisely. But the judgment would almost certainly be higher for Setting 2. The Setting 1 versus Setting 2 scenario is contrived, but the underlying principle is not. It is generalizable to more realistic circumstances, such as that in which "Setting 1" is replaced by "Patients like Ms. Jones."

The remaining precedence modes all entail some deliberate attempt to remember previous occurrences. In the *recall contents mode* (1.2.1.2), the judge tries to recall specific earlier instances when events similar to the one being considered presently actually happened. These are used to estimate, for example, the relative frequency of such past occurrences, which is then taken as the required probability judgment. Thus, suppose that, through such recall efforts, a physician estimates that, of the patients like Ms. Jones whom he has seen in the past, about 10% had asthma. He therefore concludes that there is a 10% chance that Ms. Jones has that condition, too.

The *recall availability* heuristic (1.2.1.3) is subtly but significantly different from the recall contents mode just described. In the latter, the judge uses the substance of what is remembered. Suppose that Ms. Jones's physician brings to mind 10 former patients similar to Ms. Jones and that one of them had asthma. That 1:10 ratio would yield the physician's probability judgment of asthma for Ms. Jones. In contrast, the inference in the recall availability approach is indirect. Ms. Jones's physician might try to bring to mind perhaps only a single case in which a patient similar to Ms. Jones had asthma. The physician next makes an assessment of how easy it was to recall that case, its "availability" for recall. The physician then invokes the key assumption underlying the recall availability heuristic: *The easier it is to recall a particular kind of event, the more often that event must have occurred in the past*. This assumption then justifies inferring that events similar to easy-to-recall past exemplars are highly likely to occur now also.

The *records mode* (1.2.1.4) is the "objective" or "scientific" variant of the precedence mode. The judge does not depend on fallible memory for a perhaps limited number of personally observed past cases. Instead, the judge calls upon reliable records of large, representative (if not exhaustive) samples of such cases. Thus, Ms. Jones's physician might consult an extensive database of valid records of patients who resemble Ms. Jones in relevant ways. If 13% of them had asthma, then Ms. Jones's doctor will take 13% as her probability of having asthma, too. This "evidence-based" approach is sometimes called an "actuarial" method, since it is basically the same as that used by insurance companies in arriving at the probability judgments that they use for setting premiums.

The key ideas underlying precedence modes are compelling. Nevertheless, those modes entail risks, too. One is *nonstationarity,* which essentially says that current tendencies are fundamentally different from those in the past. For example, shifts in environmental conditions might mean that the true asthma incidence changes substantially over time. Various psychological phenomena can compromise the adequacy of the nonobjective precedence modes also. For instance, recall is subject to a host of context influences, such as primacy and recency effects, whereby the earliest and the most recently observed cases, respectively, are especially likely to be remembered. The ease with which past instances can be brought to mind is also influenced by numerous other factors that have little to do with how often those instances actually occurred, such as their vividness (e.g., the breast cancer death of a prominent person).

Event Theories. The defining feature of *event theory modes* (1.2.2) is that, in some fashion or another, the judge draws on a theory of how the event in question literally comes about in nature. In

some cases, those theories are *personal* ones (1.2.2.1), sometimes described as "naive" in that typically they are buttressed not by rigorous scholarship but instead by plausible lay intuitions. Two examples of widely accepted theories in early medical history that almost certainly began as personal theories are illustrative. One is the "doctrine of signatures," according to which, supposedly, one could predict the efficacy of medicines derived from a plant by its resemblance to the organ of concern, for instance, the liverwort for treating disorders of the liver. Another is the miasma or "bad air" theory for explaining (and therefore predicting) diseases such as cholera.

The principle underlying judgment according to *scenario availability* (1.2.2.2) is similar to that deployed in the recall availability heuristic: *An event is judged likely to occur to the extent that a scenario giving rise to it is easy to imagine*. Thus, a physician's prognosis for a diabetes patient might be driven by how easily he can envision the patient adhering to a recommended treatment regimen. The operative principle is plausible but also perhaps overly self-generous: "If I personally cannot easily imagine how something can occur, that means it probably *can't* occur." One reason for having less than complete faith in this principle is that people have a hard time anticipating and even understanding the full range of complicated interactions that occur among the forces at play in many real-life scenarios, including ones in medicine.

Validated event theories (1.2.2.3) are the engines that drive modern scientific medicine. Such a theory provides an account for how a particular condition, such as an infectious disease (e.g., acute bronchitis), arises and how it progresses over time, affecting particular organs in specified ways and in a specified order. That sequence directly guides prognostic judgment. And backward reasoning from particular signs and symptoms helps narrow down differential diagnoses.

Judges turn to event theories relatively early for a given case because humans have a natural need to understand the world, not merely to predict events accurately. One hazard of event theories, though, is that they likely give short shrift to uncertainty. After all, if one firmly believes that an event is the end product of a specific A → B → C →… sequence of occurrences, there is little or no room for uncertainty.

Matching. The key feature of judgment via *matching modes* (1.2.3) is that, during the course of the judgment process, the judge matches one or more features of the event in question with some "reference," such as a prototype. Some algorithm-like diagnostic procedures provide good illustrations of the *checkoff* (1.2.3.1) variant of matching. Each possibility in a differential diagnosis is implicitly defined by a prototypical case consisting of several signs and symptoms, including, perhaps, test results. In a fashion similar to the parlor game "20 Questions," the diagnostician successively eliminates possible diagnoses whose prototypes require features missing from the case at hand until, eventually, only one diagnosis is left. In *similarity* (1.2.3.2) versions of matching, multiple characteristics of the given case are compared with corresponding features of the prototype more holistically, yielding an assessment of the degree of overall similarity between the case and the prototype. The judgment rests directly on that similarity assessment. For example, even if a given patient's profile does not exactly fit the classic pneumonia victim prototype, if the similarity is strong enough, pneumonia is the diagnosis rendered. Judgment according to the representativeness heuristic is illustrative.

Judgment via checkoffs in principle should be less demanding than judgment per similarity. However, an advantage of the latter is that it more readily acknowledges uncertainty. Unfortunately, that mode can also yield judgments that violate key formal principles such as Bayes's theorem.

Association Models. Judgment according to *association models* (1.2.4) seeks to exploit presumed statistical associations between the events in question and easily observed facts. For instance, "risk factors" such as hypertension are often used to sharpen stroke predictions, since strokes are thought to be (and are) especially common for people with high blood pressure.

In *personal* (1.2.4.1) forms of association model modes, the judge relies on personal intuitions and reasoning to derive a judgment. In special cases that are *comprehensive* (1.2.4.1.1), the judge attempts to exploit the predictiveness of at least two different facts in reaching judgments (e.g., hypertension and family history for stroke judgments). The anchoring and adjustment heuristic is a commonly discussed,

simple illustration, whereby the judge uses some general facts (e.g., a local incidence rate) to provide an initial ballpark judgment and then more case-specific facts (e.g., test results) to move that initial assessment in appropriate directions. Intuitive approximations of the logic of linear regression are other, perhaps more sophisticated, examples. In *limited* (1.2.4.1.2) forms of personal association modeling, the judge relies on just a single fact, perhaps deliberately flouting the principle that bringing to bear greater amounts of information can never reduce the statistical predictability of any event. Fast-and-frugal heuristics such as the "take the best rule" are illustrative. In that procedure, a physician would make a diagnosis solely on the basis of the one sign or symptom thought to be most predictive, ignoring all the rest. Limited association model procedures can sometimes outperform comprehensive ones. For instance, in attempting to synthesize the predictive value of multiple considerations, the judge might fail to recognize that, if some of those considerations are strongly correlated with one another, this redundancy should greatly affect how those facts are used. The burdens of juggling several facts at the same time can also reduce the judge's reliability, thereby undercutting judgment accuracy even further.

Formal (1.2.4.2) versions of the association models mode entail attempts to do the best job of "objectively" taking advantage of associations between facts and events of interest, almost always with the aid of computers. *Statistical* (1.2.4.2.1) variants include well-known linear regression–type procedures, such as discriminant function analysis, and also Bayesian updating routines. *Simulation* (1.2.4.2.2) variants take several forms. Some involve computer programs that are intended to mimic the routines that a recognized expert human judge (e.g., diagnostician) uses in actual practice. Artificial neural networks, another popular variety, are programs intended to imitate the means by which human neural networks are thought to perform mental tasks such as classifying stimuli into various categories.

It is noteworthy that, although most formal modes rely on objective association measures derived from formal records, personal modes do not. Instead, they depend on judges' own opinions about how various facts are correlated with the events at issue. And research has shown that the processes by which people arrive at their relationship beliefs are vulnerable to several forces that can introduce biases. For example, people's intuitions about how data should be used to draw relationship conclusions are often significantly different from standard statistical rules such as likelihood ratios. This clearly imposes bounds on how accurate judgments resting on personal association models can be.

Another Perspective: Deliberative Versus Nondeliberative Modes. All the various self-generation modes have been described as if they were executed consciously and purposefully. This is a reasonable assumption for "public" modes such as those resting on records and statistical models. But it is not always reasonable for modes carried out in the heads of individual judges. In fact, there is considerable evidence for another important mode distinction that underlies those displayed explicitly in the mode tree—the distinction between what may be called *deliberative* and *nondeliberative* modes.

Deliberative modes, sometimes said to involve System 2 thinking, are characterized by features such as control, effort, and awareness. In contrast, when a person applies nondeliberative modes, sometimes associated with terms such as *System 1*, *automaticity*, and *intuition,* the judgment process is often initiated on its own and cannot be stopped; once it begins, it requires virtually no mental capacity or effort, and it can function outside the person's awareness. In fact, the judge is likely to be unable to accurately describe how judgments are reached. Some modes listed in the mode tree, such as expectancy learning, might be nondeliberative virtually always. Others, however, are likely to transition from deliberative to nondeliberative as a result of experience. Thus, after a physician has made hundreds of diagnoses on the basis of a personal association model for a condition such as pneumonia, she is likely to lose awareness of how she arrives at those diagnoses.

The deliberative/nondeliberative distinction highlights the practical significance of modes. There have long been attempts to improve people's judgments by merely educating them about the existence of various biases. Such efforts have seldom succeeded. There is reason to believe that this is partly because the awareness approach presumes

that people are making their judgments on the basis of deliberative processes when they actually rest on ones that are nondeliberative, and hence beyond people's personal control. Markedly different strategies are therefore required to achieve improvements.

Collective Modes

There are two kinds of collective modes (2), distinguished by how the differing opinions of multiple judges are synthesized into a final, collaborative judgment. (Note that the judges do not have to be humans; some or all of them might be devices such as computer programs.) *Interactive* (2.1) synthesis basically entails discussion in some form, as when several clinicians hold a meeting and reach consensus about the diagnosis for a difficult case. In contrast, in *mechanical* (2.2) synthesis, the participating judges work independently and report their judgments individually. These assessments are then combined with a formula of some kind. For example, the recovery-time predictions of three different physicians might be simply averaged to yield a composite prediction that is reported to the patient.

There are many reasons to expect collective modes to yield more accurate judgments than individual modes ("Two heads are better than one,"…). But there is also reason to suspect that that potential goes unrealized in many interactive situations. For instance, people often choose to discuss things that they know in common rather than to introduce into conversation information or expertise that they hold uniquely. On the other hand, even simple averaging of multiple individual judgments typically yields composite judgments that are markedly (and surprisingly) more accurate than those of any one person.

J. Frank Yates and Andrea Angott

See also Differential Diagnosis; Heuristics; Intuition Versus Analysis; Judgment; Teaching Diagnostic Clinical Reasoning

Further Readings

Collins, R. D. (2003). *Algorithmic diagnosis of symptoms and signs: A cost-effective approach* (2nd ed.). Philadelphia: Lippincott Williams & Wilkins.

Evans, J. St. B. T. (2008). Dual-processing accounts of reasoning, judgment, and social cognition. *Annual Review of Psychology, 59,* 255–278.

Gilovich, T., Griffin, D., & Kahneman, D. (Eds.). (2002). *Heuristics and biases: The psychology of intuitive judgment.* New York: Cambridge University Press.

Stern, S. D., Altkorn, D., & Cifu, A. (2005). *Symptom to diagnosis: An evidence-based guide*. New York: McGraw-Hill.

Todd, P. M. (2000). Précis of *Simple heuristics that make us smart*. *Behavioral and Brain Sciences*, *23*, 727–780.

Yates, J. F. (1990). *Judgment and decision making.* Englewood Cliffs, NJ: Prentice Hall.

K

Kaplan-Meier Analysis

See Survival Analysis

L

Law and Court Decision Making

The law presumes that medical decision making will take place in the clinical setting without judicial participation. Medical decision making is ordinarily a private matter between the physician and the patient and, in some cases, the patient's family. There is no general legal requirement that medical decision making be judicially supervised.

One reason is practical: If there were such a requirement, both the courts and medical practice would come to a grinding halt. Another reason is jurisprudential: The legal system in the United States is adversarial, which means that first there must be a controversy before the participation of courts can occur, and one party to that controversy must seek the involvement of the judicial system. In practice, the vast range of medical decision-making matters do not involve disputes—or at least not irresolvable ones. Only in the most unusual cases are the offices of the judicial branch of government sought.

The most common situations in which recourse to the courts is sought in medical decision making are those involving the questions of whether a patient lacks decision-making capacity and, if so, who has the authority to make decisions for that person; whether a surrogate for a patient who lacks decision-making capacity is making an appropriate decision; and when the decisions that parents make for their children are not in the child's best interests. Other less common situations are instances in which patients or surrogates seek to do something very much out of the ordinary, such as the use of one child to benefit another (e.g., the transplantation of a kidney from one sibling to another) or the sterilization of persons lacking decision-making capacity.

Competent Patients

"Every human being of adult years and sound mind has a right to determine what shall be done with his own body." In 1914, Judge Benjamin Cardozo of the New York Court of Appeals—and later a justice of the United States Supreme Court—wrote these words in a judicial opinion in a lawsuit brought by a woman against a hospital where she claimed to have been subjected to surgical treatment without her consent (*Schloendorff v. Society of New York Hospital,* 105 N.E.2d [N.Y. 1914]). They have been repeated countless times since in virtually every medical decision-making case to have been decided by any court. Their origins are much older than their relatively recent legal vintage would suggest. The law of battery, which requires consent to any form of bodily "touching"—including surgery—dates back hundreds of years in English law, from which our own law is derived.

The requirement of consent to medical treatment underwent a slow transformation in the first half of the 20th century, culminating, in the 1970s, in a more sophisticated and complex requirement of informed consent to medical treatment. Today, informed consent is, in effect, the law's model of

medical decision making, prescribing how decisions about medical treatment must be made in order to be legally acceptable. Furthermore, the mainstream ethical analysis of medical decision making adheres to the same, or a very similar, model. The core principle of this model is autonomy—that individuals have the right to be free from unwanted interferences with their bodily integrity.

Autonomy is implemented in medical decision making through the requirement of consent to treatment; that is, a physician may not administer therapeutic or diagnostic procedures to a patient without the patient's consent, and the newer requirement of informed consent recognizes that mere permission does not constitute consent. Ethically and legally valid consent involves giving permission on the basis of an understanding of information about what the matter to which permission is given entails. In other words, for permission to treat to rise to the level of consent, the person giving it must understand the consequences of the medical treatment and suitable alternatives to that treatment (including no treatment) and the consequences of those alternative treatments.

Incompetent Patients

The foregoing analysis is applicable to autonomous individuals—to people who have the capacity to make decisions about their medical treatment. The following discussion applies to those who do not have such decision-making capacity.

Meaning of Incompetence

The notion of incompetence is closely related to the concept of consent. In fact, incompetence, as it is understood in the contemporary law of medical decision making, is defined as the lack of capacity to give informed consent. Giving informed consent entails the ability to communicate—to receive information and to render a decision. But it also entails the ability to understand this information and to use it in a rational manner. This definition of incompetence—or lack of decision-making capacity—is far easier to state than to apply. Difficult questions arise in application concerning the meaning of "understanding" of the information and "rational" use of it.

Determination of Incompetence

Ordinarily, incompetence is assessed in the clinical setting by the patient's physician. Sometimes the attending physician will be uncertain about a patient's decision-making capacity and seek a consultation from a psychiatrist, psychologist, or neurologist, and in instances where there is an intractable dispute about a person's decision-making capacity, a determination by a judge (referred to as an adjudication of incompetence) may be obtained.

Because, in fact, most adults possess decision-making capacity and because the law presumes that adults possess decision-making capacity, an assessment to determine incapacity ordinarily should occur (and probably does occur) only when the patient's behavior or condition is such as to raise strong suspicions of incapacity. At one extreme is the unconscious patient who clearly lacks decision-making capacity; more difficult cases involve patients who are demented, intoxicated, obtunded, mentally ill, or mentally retarded, some of whom may be intellectually compromised but not so much that they lack decision-making capacity. Children below the age of 18 are presumed to lack decision-making capacity, but more mature minors may in fact have the capacity to make some or all decisions about their medical treatment, and the law recognizes their authority to do so at least in limited circumstances—generally those involving minors who are married, are or have been pregnant, or are being treated for mental illness, drug use, or sexually transmitted diseases.

Deciding for Others

Decision making for patients who lack the capacity to make their own decisions is, like decision making for competent persons, based on autonomy. Even though an individual may have lost the capacity to make decisions, decisions can still be made for that person in a way that attempts to further the same values that the person would consider if making the decision personally.

Advance Directives

The most direct way for this to happen is through an advance directive, which is a written

instrument by which one, while still competent, gives instructions about one's medical treatment that may need to be administered after one has lost decision-making capacity. Such instruments are usually referred to as *living wills*, though sometimes they are referred to as *medical directives, health care directives,* or *instruction directives*.

A less direct, but more flexible, means of achieving this end is for an individual to appoint an agent to make medical decisions for him or her at some future time after having lost decision-making capacity. The instrument by which this is accomplished is referred to as a *healthcare power of attorney*, and the individual so appointed is referred to as a *healthcare agent* or *healthcare proxy*.

When a person has executed an advance directive and later loses decision-making capacity, the advance directive should be the starting point for decision making. If the advance directive gives instructions, it should be followed by the patient's physician to the extent that the instructions are relevant to the particular decisions that actually need to be made. Often, however, this is not the case. Individuals drafting instructions may not have extraordinary foresight about the decisions that will need to be made, or their instructions may be so general that they require a great deal of interpretation to be applied to those particular decisions. Even then their instructions may not be relevant.

In contrast, an advance directive that appoints an agent to make decisions provides a great deal more flexibility but considerably less guidance, unless the patient combines it with some directions about the kind of treatment he or she does or does not wish to have. In many instances of decision making for patients who lack decision-making capacity, however, there is no advance directive of either type, and decisions must be made in a different fashion.

Procedures for Surrogate Decision Making

The consequence of being determined to lack decision-making capacity is that one is disenfranchised from making medical decisions for oneself, and someone else must of necessity make such decisions. The general term for such an individual is *surrogate*, but there are a number of different types of surrogates, depending on how they assume this position.

In some instances, an individual, in anticipation of a possible future loss of decision-making capacity, designates someone to be his or her surrogate should the loss of decision-making capacity in fact occur. Such a patient-designated surrogate is referred to as a *proxy* or *agent*.

The means by which this appointment is made may be a formal, legal one involving the execution of a document known as a *health care power of attorney*, or it may be a more informal, usually oral, one. Indeed, it may even be more informal than that, as when a patient who is accompanied by another person to a medical appointment may, by including that other individual in discussions with a physician, tacitly authorize the other to be his or her proxy then, in the future, or both.

When an individual fails to designate a proxy, a surrogate may be named pursuant to law. Many states have legislation that sets forth a list of individuals who are deemed to be the surrogate for an individual who lacks decision-making capacity. These individuals—referred to as *statutory surrogates*—are usually close family members—such as spouse, parents, adult offspring, and siblings—and occasionally more distant relatives by blood or marriage. Further down the list may be friends, clergy, or healthcare professionals.

If legislation of this kind does not exist, the customary practice in the health professions is for close and involved family members to act as surrogates for patients. In some states, judicial opinions have expressly approved of this practice. Where there has been no such formal judicial approval, a sound legal argument can still be made that it is legally acceptable for close family members to act as surrogates. Such a surrogate is a *common-law surrogate*.

In cases in which a patient has no close family member to serve as a surrogate, a surrogate may be appointed by a court. This type of surrogate is usually referred to as a *guardian* (though the terminology is different in a few states). When circumstances dictate the need for rapid action in appointing a guardian, this can sometimes be done quickly, though for a limited period of time. Ordinarily, the appointment of a guardian requires a judicial hearing with testimony from a psychiatrist or psychologist about the patient's decision-making capacity.

Under limited circumstances, recourse to the courts for surrogate decision making is appropriate

even if a surrogate already exists. One such instance arises when it is unclear who should serve as surrogate. This may occur because a patient has executed more than one healthcare power of attorney appointing more than one person to serve as his or her agent. It may occur because there is more than one family member in a statutory class of persons authorized to act as a surrogate. Or it may occur in a case of common-law surrogacy because a patient has more than one ready, willing, and able family member to act as surrogate. When there is irresolvable conflict among these individuals, or when it is just not clear who has the authority to serve as surrogate, recourse to the courts to sort out the matter is warranted. Also, if a particular individual clearly has the legal authority to act, but that individual lacks the capacity to make a decision, a judicial proceeding might be instituted to disqualify the individual. The same is true if there is a serious conflict of interest between the surrogate and the patient, such that the surrogate is clearly motivated to act in his own best interests rather than the patient's.

Emergencies

In a life-or-death emergency, there is no need for a surrogate decision maker. A physician is legally authorized to administer life-saving medical treatment without the patient's or a surrogate's consent, unless there has been a refusal of the treatment in question by the patient or a surrogate prior to the occurrence of the emergency.

Standards for Surrogate Decision Making

When a surrogate makes a decision for an incompetent patient, the surrogate is required to follow a set of legal standards in so doing. These are referred to as *surrogate decision-making standards*. The standards are prescribed either by judicial decisions or by legislation. The predominant standard is referred to as the *substituted judgment standard*. Pursuant to this standard, the role of the surrogate is to determine insofar as is reasonably possible what the now incompetent patient would decide if the patient were able to make a decision. Put another way, the surrogate is to determine the patient's *probable wishes* about medical treatment.

In so doing, the surrogate is permitted to rely on any discussions that the surrogate or others had with the patient before the patient lost decision-making capacity; on any oral statements the patient made; and on any written statements the patient made that do not satisfy the legal requisites of an advance directive. In taking this information into account, the surrogate should consider it in the context in which it was articulated, for instance, whether it was made in a casual way or under more serious circumstances.

In addition to direct statements by the patient, other evidence is relevant to a decision made pursuant to the substituted judgment standard. The surrogate may also take into account such matters as the patient's age and life expectancy with or without the contemplated treatment; the probable side effects of treatment, including any suffering, pain, or disorientation that the treatment may cause; the potential benefits of treatment; the patient's religious beliefs or value system; and the quality of the patient's life with or without treatment.

Decision making under the substituted judgment standard is an inferential process. If there is direct and relevant evidence of the patient's wishes about treatment—such as that contained in a living will—that should guide the decision making and a surrogate is not even necessary, at least as to those components of decision making to which the direct and relevant evidence applies. However, in many instances, that is not the case. Either a living will does not exist, or, if it does, it requires interpretation to be applied, or it does not address in a direct way the particular question at issue.

A very small number of states require more exacting evidence to guide the surrogate in making a decision than is permitted under the substituted judgment standard. This standard is often referred to as the *clear and convincing evidence* standard, but a better term is the *actual intent* standard because it requires that decisions made by the surrogate be based on the patient's actual wishes about treatment rather than on the patient's probable wishes, as is required under the substituted judgment standard. Thus, the surrogate may not infer the patient's wishes from, for example, the patient's religious beliefs or value system, but must rely on clear and convincing statements made by the patient. In the strictest version of this standard, the surrogate is merely a conduit for the patient's

wishes and plays no independent role in attempting to discern what those wishes are.

In some instances, there will be no evidence of the patient's wishes, inferential or direct. In these situations, the surrogate is legally obligated to make medical decisions according to the *best interests* standard. When forgoing the treatment in question would bring about the patient's death, some states require that treatment be administered on the ground that life is always preferable to death. Most states recognize, however, that sometimes the burdens to the patient of continuing life far outweigh any benefits to the patient and therefore permit the surrogate to discontinue treatment on the grounds that its administration would not be in the patient's best interest.

Alan Meisel

See also Advance Directives and End-of-Life Decision Making; Bioethics; Informed Consent; Informed Decision Making; Patient Rights; Risk-Benefit Trade-Off; Surrogate Decision Making; Terminating Treatment, Physician Perspective

Further Readings

Berg, J. W., Appelbaum, P. S., Parker, L. S., & Lidz, C. W. (2001). *Informed consent: Legal theory and clinical practice*. New York: Oxford University Press.

Meisel, A., & Cerminara, K. (2008). *The right to die: The law of end-of-life decisionmaking*. New York: Wolters Kluwer Law and Business.

League Tables for Incremental Cost-Effectiveness Ratios

A cost-effectiveness league table is a listing of health interventions ranked by their incremental cost-effectiveness ratios (ICERs) presented in terms of cost per life years, or cost per quality-adjusted life years (QALYs) gained. A typical table (see Table 1) starts with the most favorable (lowest ICER) intervention and ends with the least favorable one (highest ICER). League tables are used to place findings of a cost-effectiveness analysis in a broader context and help determine whether a specific intervention presents "good value for the money." They may be used for informing resource allocation decisions: Under a fixed budget constraint, healthcare resources are allocated starting with interventions with the lowest ICER and moving to higher ICERs until the entire budget is consumed.

Cost-Effectiveness Analysis and Rationing

Cost-effectiveness analysis is a method of economic evaluation in which costs and outcomes of a program and at least one alternative are compared. The difference in cost (incremental cost) is divided by the difference in outcomes (incremental effect) to derive the incremental cost-effectiveness ratio. Although any natural unit of outcome can be used to determine a program's effect, the common metrics used are life years or QALYs gained because they allow for comparisons across diverse treatments and diseases.

An analysis using cost per QALY as an outcome measure is sometimes referenced as a cost-utility analysis. This is the most widely used method for informing resource allocation decisions in healthcare. As opposed to benefit-cost analysis, cost-effectiveness studies do not present health outcomes in monetary terms, which would permit a straightforward comparison of costs and benefits to determine whether an intervention is worthwhile. As a result, the relative value for the money of an intervention can only be interpreted by a reference to an external standard. This standard can be a benchmark or threshold value (e.g., $50,000 per QALY gained) below which an intervention can be considered to be "good value for the money" or a comparison of the relative cost-effectiveness of various interventions of which some may be already covered by health plans.

History of League Table Presentation

The presentation of league tables (so called after the tables used for British soccer league standings) and comparisons between healthcare interventions in terms of their relative cost-effectiveness became fashionable in the 1980s. Since a common metric is used (life years or QALYs), league tables can be useful for the comparison and ranking of diverse interventions to improve health, from public health or environmental programs to medical technology. One of the first league tables was presented by

Table I League tables for incremental cost-effectiveness ratios

Intervention and Comparator	*Cost-Effectiveness Ratio (2002 Dollars)*
Cochlear implant as compared with no implant in profoundly deaf children (average hearing loss > 90 dB for both ears)	Cost-saving
Treatment with tinzaparin sodium as compared with unfractionated heparin in patients with deep vein thrombosis	Cost-saving
Long-term androgen-deprivation with radiation therapy (RT) as compared with short-term androgen-deprivation with RT in men with histologically confirmed adenocarcinoma of the prostate	$1,100/QALY
Interferon therapy as compared with no treatment in patients with chronic hepatitis B virus (HBV) infection, elevated aminotranferase levels, and no cirrhosis	$6,000/QALY
Warfarin therapy as compared with aspirin therapy in patients with chronic atrial defibrillation, varying risk of stroke, and no contraindications to anticoagulation therapy at age 70	$16,000/QALY
Linezolid therapy as compared with vancomycin therapy in patients with ventilator-associated pneumonia	$30,000/QALY
Clopidogrel therapy as compared with aspirin therapy in patients with a prior stroke, requiring secondary prophylaxis of vascular events at age 63	$31,000/QALY
Treatment with drotrecogin alfa (activated protein C) with 96-hour intravenous infusion at 24 μg/kg/hr plus usual care as compared with placebo plus usual care in patients with severe sepsis in the intensive care unit	$51,000/QALY
Cardiac resynchronization therapy as compared with medical therapy in patients with reduced ventricular function and prolonged QRS	$110,000/QALY
Adolescent meningococcal vaccination (single dose) as compared with no vaccination in a hypothetical U.S. population cohort of children at age 11	$130,000/QALY
Left ventricular assist device, as compared with optimal medical management, in patients with heart failure who are not candidates for a heart transplant	$800,000/QALY
National minipool blood supply testing for West Nile virus over entire year as compared with national minipool blood supply testing for half of the year in transfusion recipients with and without underlying immunocompromise in patients aged 60 or older	Dominated
Surgery in 70-year-old men with a new diagnosis of prostate cancer as compared with watchful waiting	Dominated

Source: Tufts Medical Center Cost-Effectiveness Analysis Registry (https://research.tufts-nemc.org/cear/default.aspx).

Notes: The cost-effectiveness ratio is the incremental costs divided by the incremental benefits (QALYs gained). The cost-effectiveness estimates listed are point estimates from the original articles. A more detailed description may be found at the Cost-Effectiveness Registry (https://research.tufts-nemc.org/cear/default.aspx).

John Graham and James Vaupel and included estimates of cost per life saved and cost per life year saved. Allan Williams presented a league table for the United Kingdom in his seminal work on the economics of coronary artery bypass grafting (CABG). Other examples of comprehensive league

tables can be found in the work of Tammy Tengs and colleagues, who presented in 1996 a list of more than 500 "life-saving" interventions and their relative cost-effectiveness and, more recently, in the data presented in the Tufts Medical Center Cost-Effectiveness Analysis Registry, gathered through an extensive review of cost-effectiveness analyses published since 1976.

Limitations of League Tables

League tables are useful for decision makers only if all competing interventions are listed and ranked in the table. This task, however, involves an enormous analytical effort and is unlikely to be feasible. Another concern has been raised about the comparability of the methodology used to determine the ICER in various studies.

Information presented in a league table may be helpful for decision makers only if they are confident that the methodology used in the source studies presented in the table is relatively homogeneous. The variability in the methods used for conducting cost-effectiveness analyses has been well documented. Cost-effectiveness analyses may differ by the choice of the comparison interventions, the study perspective, the range of costs and consequences considered, the time horizon of the analysis, methods for estimating health preferences, the discount rate applied, and the clinical assumptions used. In addition, league tables usually present only a point estimate of the cost-effectiveness ratio and do not include measures of uncertainty and variability around that point estimate.

Standardization of Studies Presented in League Tables

In an effort to standardize the practice of cost-effectiveness analysis, the U.S. Panel on Cost-Effectiveness in Health and Medicine recommended the reporting of a reference case ICER that can be compared with those of other interventions for the same or a different health condition. The reference case should adopt a societal perspective and community or patient preferences utility weights, and it should feature use of net costs, appropriate incremental comparisons, and discounting of costs and QALYs at the same rate. A league table that presents studies that follow the Panel recommendations may enhance comparisons among studies.

Other scholars have proposed various ways to improve cost-effectiveness league tables and make them more relevant to decision makers. These suggestions include, but are not limited to, presentation of an incremental analysis of the value of a new intervention over the best available practice for the disease considered, providing details on the comparator treatment used to calculate the ICER, and presentation of the confidence limits for the ICERs. It is critical, however, that league tables to be used by the relevant decision makers be kept simple and include only information relevant for the decision maker.

Dan Greenberg and Peter J. Neumann

See also Cost-Effectiveness Analysis; Cost-Utility Analysis; Quality-Adjusted Life Years (QALYs); Rationing; Reference Case

Further Readings

Chapman, R. H., Stone, P. W., Sandberg, E. A., Bell, C., & Neumann, P. J. (2000). A comprehensive league table of cost-utility ratios and sub-table of panel-worthy studies. *Medical Decision Making*, *20*, 451–467.

Drummond, M. F., Schipher, M. J., Torrance, G. W., O'Brien, B. J., & Stoddart, G. L. (2005). *Methods for the economic evaluation of health care programmes*. Oxford, UK: Oxford University Press.

Drummond, M., Torrance, G., & Mason, J. (1993). Cost-effectiveness league tables: More harm than good? *Social Science and Medicine*, *37*, 33–40.

Gold, M. R., Siegel, J. E., Russell, L. B., & Weinstein, M. C. (Eds.). (1996). *Cost-effectiveness in health and medicine*. New York: Oxford University Press.

Graham, J., & Vaupel, J. (1981). Value of a life: What difference does it make? *Risk Analysis*, *1*, 89–95.

Mason, J. M. (1994). Cost per QALY league tables: Their role in pharmacoeconomic analysis. *PharmacoEconomics, 5*, 472–481.

Mauskopf, J., Rutten, F., & Schonfeld, W. (2003). Cost effectiveness league tables: Valuable guidance for decision makers? *PharmacoEconomics, 21*, 991–1000.

Neumann, P. J. (2005). *Using cost-effectiveness analysis to improve health care: Opportunity and barriers*. New York: Oxford University Press.

Tengs, T. O., Adams, M. E., Pliskin, J. S., Safran, D. G., Siegel, J. E., Weinstein, M. C., et al. (1996). Five-hundred life-saving interventions and their cost-effectiveness. *Risk Analysis*, *15*, 369–390.

Tufts-New England Medical Center. (2008). *Cost-effectiveness analysis registry*. Retrieved March 19, 2008, from https://research.tufts-nemc.org/cear/default.aspx

Williams, A. (1985). Economics of coronary artery bypass grafting. *British Medical Journal*, *292*, 326–329.

Learning and Memory in Medical Training

The knowledge and skills an experienced physician has acquired over the years are very impressive. It is often difficult to believe and even more difficult to understand how a student with very limited knowledge and skills develops into a competent physician who is able to tackle most of her or his professional problems. To get a better understanding of how this is possible, one has to take a closer look at medical education.

Biomedical and Encapsulated Knowledge

During their first years in medical school, students acquire a large body of knowledge concerning the basic sciences. This type of knowledge, which will further be referred to as biomedical knowledge, entails subjects such as physiology, anatomy, microbiology, and pathology. Biomedical knowledge will develop into rich and elaborate causal networks that explain the causes and consequences of disease in terms of general underlying processes. Most medical curricula are divided into preclinical years, which mainly focus on biomedical knowledge, and a clinical program consisting of clinical courses and practicals. During the clinical years, students are often for the first time confronted with real patients, and they can witness the impact of a disease on a patient's life. This clinical experience also provides students with the opportunity to establish links with their elaborate biomedical knowledge.

Encapsulated Knowledge

This integration of biomedical and clinical knowledge is most clearly established by Schmidt and Boshuizen's theory of knowledge encapsulation. It is assumed that through extensive and repeated application of biomedical knowledge and through confrontation with clinical problems, the students' elaborate biomedical networks of knowledge will eventually be subsumed under higher-level concepts with the same explanatory power. In other words, experienced physicians who are asked to explain the signs and symptoms of a patient will most likely use much more encapsulated concepts in their explanations than a less-experienced medical student.

For instance, if a student is required to explain the shortness of breath of a patient with a heart condition, her or his response may look like this: "If the ability of the heart to pump the blood forward from the left side is diminished, the body does not receive enough oxygen. The pressure in the veins of the lung increases and may result in fluid accumulation in the lung, leading to shortness of breath." An experienced physician, on the other hand, will refrain from this detailed, causal explanation and may respond by saying that shortness of breath in this case results from left-sided heart failure. This does not imply, however, that the physician does not know the detailed explanation provided by the student—studies have shown that physicians can easily produce the student's explanation if the task requires it—but it does imply that the concept of "left-sided heart failure" incorporates or encapsulates the detailed, causal description.

Similar results were found in clinical case studies using free recall. In these studies, using the so-called clinical case paradigm, participants of different levels of expertise were required to study a clinical case description, provide a diagnosis, and write down everything they could remember from the case (in a free order). Each case reported some contextual information, the complaint, findings from history-taking and physical examination, relevant laboratory data, and some additional findings (e.g., X-rays, ECGs). What many of these studies have shown is that advanced medical students not only remembered more from the case description than less-advanced students; they also remembered more from the case than experienced physicians. This phenomenon has been dubbed *the intermediate effect*. That is, the participants from an intermediate level of expertise remembered most details from the case. In line with the knowledge encapsulation theory, physicians will refrain from a detailed recall of all the facts and findings in a clinical case. They will instead use higher-level concepts that incorporate or summarize much of the information provided in the case. For example,

consider the following fragment from a clinical case description:

> A 45-year-old woman who has always had a fast heartbeat has been admitted to the hospital. Physical examination reveals an enlarged thyroid gland that is elastic on palpation. Her weight is 50 kilos and she is 1.70 m tall.

Instead of reproducing the provided information, which will be done by most (advanced) students, an experienced physician will immediately recognize this pattern of signs and symptoms as belonging to a condition called hyperthyroidism. Generating many of these inferences during case processing and mainly reporting them during recall leads to a short but highly relevant account of a clinical case.

Research Findings

The idea that biomedical knowledge becomes encapsulated into clinical knowledge has led to several predictions, which have been confirmed by experimental studies. These studies have shown that encapsulated knowledge is more readily accessed by physicians than biomedical knowledge; pathophysiological explanations (i.e., explanations of the signs and symptoms in a clinical case) by experts contain less biomedical and more encapsulating concepts than those by students; and recall protocols of physicians dealing with cases within their area of expertise contain more encapsulations than those of physicians outside their expertise area. These findings seem to suggest that the role of biomedical knowledge is rather limited as compared with encapsulated knowledge. Studies by Norman and colleagues at McMaster University, however, have shown that biomedical knowledge plays a crucial role in providing coherence between otherwise unrelated signs and symptoms. They have shown that students who were asked to study a list of features associated with a number of diseases performed less well on a recall task after a delay of 1 week than students who had to learn a causal relationship between these features. That is, causal relationships (i.e., biomedical knowledge) clarify coherence among symptoms in a way that simple association does not. Interestingly, these researchers also showed that students spontaneously developed encapsulated concepts, as evidenced by better performance on a recognition test presenting new concepts encapsulating the learned causal mechanisms. These findings also demonstrate that the process of knowledge encapsulation can start relatively early and will continue throughout the physician's professional career.

Scripts

The development of encapsulated knowledge also triggers a second major shift in the student's knowledge organization. As students during their later years of training encounter more and more patients, their encapsulated knowledge is restructured into a type of narrative that has been called a script. The development of scripts is not something that is unique to medicine; people develop scripts for many activities that are done on a regular basis and consist of more than one step. A classic example is the restaurant script. Most people are very familiar with the sequence of events that take place when they enter a restaurant. This sequence of events (waiting to be seated, getting the menu, ordering the food, etc.) is very similar for most restaurants, and our scripts help us to anticipate these events and to act accordingly. Similarly, a physician who has encountered a certain disease many times will develop a script, or better, an illness script, that enables him or her to deal with the problem more efficiently.

In medicine, illness scripts are cognitive structures that contain relatively little knowledge about pathophysiological causes and symptoms and complaints (as a result of encapsulation), but a wealth of clinically relevant information about the so-called enabling conditions of disease (i.e., contextual information about the conditions that make the acquisition of a disease more likely, such as heredity factors). Advanced levels of expertise are characterized by elaborate knowledge about enabling conditions because it enables the physician to rule out many diseases and to focus on those diseases that are most likely. For example, if a woman enters the consulting room complaining about fever-like symptoms in the middle of a flu epidemic, the doctor will obviously think of flu. However, if the woman also tells the doctor that she has recently visited a malaria-infected region, then this "enabling condition" may lead to an alternative diagnostic hypothesis.

Elaborate knowledge about enabling conditions helps physicians improve their diagnostic performance. This point was elegantly demonstrated in the 1980s by Schmidt and colleagues at Maastricht University in the Netherlands. They gave experienced and inexperienced general practitioners cases with enabling conditions and cases without enabling conditions. Removing the enabling conditions had hardly any effect on the diagnostic performance of the inexperienced practitioners because they could not use this information anyway. The experienced practitioners, on the other hand, although their performance was better than that of the inexperienced practitioners, showed a dramatic drop in accuracy: Their diagnostic performance was roughly half as good as when they had the enabling conditions. This and other studies on illness scripts indicate that elaborate knowledge about enabling conditions increases over the years and is essential for physicians to come up with the correct diagnostic alternative.

It is important to note that the acquisition of enabling conditions is to a large extent based on clinical experience, and formal education seems to play a minor role in this process. Furthermore, the successful application of illness scripts depends heavily on the physician's memory. That is, while solving a problem, physicians have to search for the most appropriate script in their long-term memory by matching it with the information about the patient. If this verification process is successful, a particular script becomes instantiated. However, an instantiated script does not necessarily become decontextualized after being used, but remains available in long-term memory and may be used if a similar problem occurs in the future. In other words, illness scripts exist at various levels of generality, ranging from representations of disease categories, to prototypes, to representations of previously seen patients.

Medical Expertise Development

In an overview article on medical expertise development, Schmidt and Rikers summarize the transitory states a medical student has to go through in order to become an experienced physician:

1. Development of elaborate knowledge explaining the causes and consequences of disease using primarily biomedical concepts.
2. Development of encapsulated knowledge. The detailed biomedical concepts become integrated with higher-order (clinical) concepts with the same explanatory powers.
3. Development of illness scripts. The encounter with many different manifestations of a disease in patients will lead to a rich and elaborate illness script. In particular, the illness script stores the enabling conditions or contextual factors of disease, which play a crucial role in diagnostic performance.
4. Development of interpreted instances of illness scripts as exemplars of the particular disease. These instances are available in long-term memory and may be used to diagnose a similar problem in the future.

Implications

Based on this overview on learning and memory, a number of implications for medical education can be identified. First, the development of encapsulated knowledge should be facilitated by modern curricula through integrated teaching (i.e., the integration of biomedical and clinical science). Second, to promote knowledge encapsulation and the formation of illness scripts, students should work with patient problems early on in the curriculum. Finally, during clerkships and other postings, students should reflect on patient problems they have encountered, preferably with the help of an experienced coach and in small groups of peers. In this way, they will develop adequate knowledge structures that are essential to their becoming proficient physicians.

Remy Rikers, Sílvia Mamede, Elisabeth van Rijen, and Henk G. Schmidt

See also Automatic Thinking; Cognitive Psychology and Processes; Heuristics; Medical Errors and Errors in Healthcare Delivery

Further Readings

Elstein, A. S., Shulmann, L. S., & Sprafka, S. A. (1978). *Medical problem-solving: An analysis of clinical reasoning.* Cambridge, MA: Harvard University Press.

Hobus, P. P. M., Schmidt, H. G., Boshuizen, H. P. A., & Patel, V. L. (1987). Contextual factors in the activation of first diagnostic hypotheses: Expert-novice differences. *Medical Education, 21,* 471–476.

Norman, G. R. (2005). Research in clinical reasoning: Past history and current trends. *Medical Education, 39,* 418–427.

Patel, V. L., Evans, D. A., & Groen, G. J. (1989). Biomedical knowledge and clinical reasoning. In D. A. Evans & V. L. Patel (Eds.), *Cognitive science in medicine: Biomedical modeling* (pp. 49–108). Cambridge: MIT Press.

Schmidt, H. G., & Boshuizen, H. P. A. (1993). On the origin of intermediate effects in clinical case recall. *Memory and Cognition, 21,* 338–351.

Schmidt, H. G., & Rikers, R. M. (2007). How expertise develops in medicine: Knowledge encapsulation and illness script formation. *Medical Education*, *41*, 1133–1139.

Lens Model

The lens model is a general framework originated by Egon Brunswik (1903–1955) that describes the situation that people face when using multiple items of imperfect information, or *cues*, to make judgments about an uncertain environment. The cues mediate between the judgment and a *distal variable*, which is the observation or event that is the object of judgment. There are numerous representations of the lens model in the literature, but all include seven elements: (1) a distal variable that is the focus of judgment; (2) a set of cues, or *proximal variables*; (3) a judgment about the distal variable, based on the cues; (4) imperfect relations, called *cue validities*, between the cues and the distal variable; (5) imperfect relations, called *cue utilizations*, between the cues and the judgment; (6) interrelations among the cues; and (7) the relation between the cue and the judgment, called *accuracy* or *achievement*.

Description

Brunswik chose the analogy of the lens to describe how organisms perceive a distal object through the "lens" of the surface, or proximal, data that are available to the person making a judgment. Figure 1 is a typical representation of the lens model. The left side of the lens is called the environmental side, and the right side is called the subject side. Each of the seven common elements of the lens model is described in turn.

Distal Variable

On the left side of the lens model is the distal variable (Y_e), also called the *criterion*. It represents an event or observation that is the focus of judgment, such as the presence of a disease, the severity of an illness, or the correct dosage of a drug. In medicine, this is often called the gold standard. It is "distal" relative to the person making a judgment because it is not directly available to him or her.

Cues

The cues (X) are the surface data that are available for making inferences about the distal variable. The cues include, for example, the symptoms, features of images, or test results that are available to the person at the time the judgment is made. The cues are imperfect. They may be subject to measurement error, and they may lack some of the information needed to make an ideal judgment. This limitation is common in judgment situations that occur in medicine.

Judgment

On the right side of the lens model is the judgment (Y_s), which is usually considered to be continuous. If the distal variable is continuous, then the judgment is measured in the same units as the distal variable. If the distal variable is binary or categorical, then the judgment is usually thought of as a probability. In either case, the judgment represents a person's attempt, implicitly or explicitly, to assess or predict the value of the distal variable, either in the present or at some future time. The judgment is assumed to be empirically verifiable, at least in principle. Judgment is necessary because the distal variable is not directly observable, either because it is obscured or hidden in some way or because it occurs in the future. Judgments of preference or value, such as how pleasing a painting is, or the desirability of a dessert item, are not addressed by the lens model.

Medical judgments that have been studied using the lens model include severity of depression, severity of rheumatoid arthritis, diagnosis of pulmonary embolism, treatment of upper respiratory tract infection, diagnosis of otitis media, physician practice patterns in hypertension, diagnosis of pneumonia, and mammography screening.

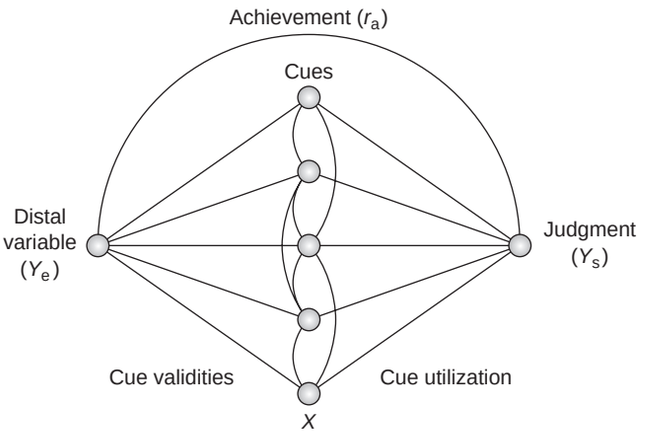


Figure 1 The lens model

Cue Validities

Lines from the cues to the distal variable represent the cue validities. These will differ among cues because some will be more strongly related to the distal variable than others. Furthermore, some cues may be linearly related to the distal variable, while others have a nonlinear relationship. Typically, the relationship between the cues and the distal variable is assumed to be additive, but this is not always the case. No single cue is perfectly valid, nor is the set of cues sufficient to make a perfect determination of the distal variable. The uncertain relation between the cues and the distal variable is called *environmental uncertainty*, and that uncertainty creates the need for judgment. Environmental uncertainty sets an upper limit on judgmental accuracy.

Cue Utilizations

Lines from the cues to the judgment represent cue utilization by the judge. Cues will differ in their relative importance to the judge, and judgments are probabilistically related to the cues due to the inconsistency or unreliability of judgment. Furthermore, cue utilizations do not always match cue validities, resulting in judgments that are less accurate than they could be.

Cue Intercorrelations

The cues are not independent, but are generally correlated with one another. This is represented in the lens model by the lines connecting the cues. It results in *causal ambiguity*, that is, the effects of an individual cue cannot be determined because of covariance with other cues. But these correlations also create the opportunity for vicarious functioning, which Brunswik believed to be a central focus of psychology. Vicarious functioning means that when one or more cues are missing, the judge may be able to substitute other cues that are correlated with the missing ones.

Accuracy

The arc between the judgment and the distal variable represents accuracy or achievement. The goal of the person making judgments is to achieve the greatest level of empirical accuracy possible.

Concepts of Brunswik's Probabilistic Functionalism

Brunswik emphasized that behavior is goal directed. Understanding a judgment requires understanding what the judge is trying to accomplish. In perception and judgment, accuracy is the goal. The lens model describes why judgments are not perfectly accurate.

Central to Brunswik's theory are the ideas of symmetry and parallel concepts. The lens model is symmetric; for each concept on one side of the model there is a parallel concept on the other. Cue validities on the environmental side parallel the cue utilizations on the subject side. If cue utilizations

do not match cue validities, the judgment will be less accurate than it could be. The relations between the cues and the criterion on the environmental side of the lens are probabilistic; that is, there is irreducible uncertainty in the environment that the judge must cope with. The parallel concept on the subject side of the lens model is inconsistency or unreliability of judgment. In other words, the cues are not perfectly related to the judgment. Unreliability can be another source of error in judgment.

Brunswik also distinguished between surface and depth in the lens model. Surface variables are cues that are available to the judge. The judge must rely on surface variables to make inferences about depth variables—the observations or gold standard that are not available to the judge because they are inaccessible or will not occur until some future time. Relations between surface and depth variables are imperfect both on the left side of the lens (environmental uncertainty) and on the right side of the lens (judgmental consistency).

Finally, the idea of causal ambiguity is important in Brunswik's theory. The lines between the cues represent relations (correlations) among them. Correlations among cues make it difficult to attribute causality to individual cues. Causal ambiguity induces *quasi-rational* judgment, that is, judgment that involves elements of both intuitive and analytic processes.

Brunswik believed that the consequences of interrelations among cues (causal ambiguity and vicarious functioning) should be the central focus of psychology. This led to his emphasis on representative design of experiments, rather than experiments that manipulate only one variable or manipulate several variables in an orthogonal design. It is not possible to study causal ambiguity or vicarious functioning without using several interrelated cues.

The Lens Model Equation

The lens model suggests three factors that limit judgmental accuracy. One is the uncertainty in the environment that results because the cues are not perfect predictors of the distal event. Another is the imperfect relation between the cues and the judgment; that is, judgments are not perfectly reliable. The third is that the judge may not use the cues in optimal fashion—the cue utilizations may not match the cue validities.

After Brunswik's death, a mathematical expression of the lens model, called the lens model equation, was developed by Kenneth Hammond and his colleagues, and refined by Ledyard R Tucker. That equation is based on two multiple regression models. One model regresses the distal variable on the cues and another regresses the judgment on the same cues. The lens model equation shows that, with appropriate assumptions, achievement is approximated by the product of environmental uncertainty, judgmental consistency, and the match between cue validities and cue utilizations.

The lens model has been used as a framework for research on judgment within the tradition of researchers who are primarily interested in describing the empirical accuracy of judgment.

Thomas R. Stewart

See also Social Judgment Theory

Further Readings

Brunswik, E. (1952). *The conceptual framework of psychology*. Chicago: University of Chicago Press.

Brunswik, E. (1956). *Perception and the representative design of psychological experiments*. Berkeley: University of California Press.

Hammond, K. R., & Stewart, T. R. (Eds.). (2001). *The essential Brunswik: Beginnings, explications, applications*. New York: Oxford University Press.

Hursch, C. J., Hammond, K. R., & Hursch, J. L. (1964). Some methodological considerations in multiple-cue probability studies. *Psychological Review, 71*(1), 42–60.

Tucker, L. R. (1964). A suggested alternative formulation in the developments by Hursch, Hammond, and Hursch, and by Hammond, Hursch, and Todd. *Psychological Review, 71*(6), 528–530.

Life Expectancy

Life expectancy is considered to be the length of survival of a person or a patient. On one hand, this

length of survival can be calculated as life expectancy at birth, which represents the number of years an infant is supposed to live throughout his or her life. This information is of importance for political, administrative, and insurance purposes but has less meaning in daily practice for a physician. On the other hand, life expectancy can also be calculated as life expectancy at any time point in the life of a person; this then represents the number of years a person is supposed to live throughout the rest of his or her life. A physician is especially interested in this latter life expectancy, as for medical decision making it can be important to know, at a certain time point, how much longer a patient will have to live.

In the following, life expectancy in general, defined as life expectancy at birth and its changes over the decades, will be briefly discussed, as this information is relevant for present and future health systems as well as for prevalence and incidence of certain diseases. This will be followed by a more specific discussion of life expectancy at a given time point in life and about tools that can be used to make predictions of life expectancy accurate and reliable and therefore useful for medical decision making.

Life Expectancy in General

Life expectancies are assessed by considering the age-specific death rates of a population. These age-specific death rates are calculated by dividing the number of deaths in a given age-group by the total population of that age-group. They are expressed as an average annual rate in a 100,000-person population. These values represent the base to develop life tables which then can be used to estimate the probability of surviving from one age to the next. Based on that, the life expectancy at birth represents the average number of years infants are supposed to live if they were to experience the death rates present in the year of birth throughout their life. Moreover, life tables can be used to calculate the number of remaining life years at a certain age, if one considers the age-specific death rates for each year a person will age in the future.

Improvements in sanitation, nutrition, and medical knowledge have resulted in a constant increase of life expectancy during the past decades throughout the world. The greatest improvements have been observed in the most developed parts, such as North America, Europe, and Japan. For example, life expectancy at birth in the United States in 1900 was 47 years and reached 78 years in 2005. Life expectancy in India in the middle of the 20th century was around 39 years and reached 64 years in 2005. It is of note that the calculations of life tables do not consider any expected changes of life expectancy in the future. This means that the calculations of life expectancy are rather hypothetical, as they effectively assume that current death rates are "frozen" and will not change in the future.

A major exception to the general improvement of life expectancy has been noted in countries with a high incidence of AIDS, as AIDS has become the leading cause of death in these regions. This is the case especially in sub-Saharan Africa, where a significant decrease in life expectancy has been observed. The overall life expectancy in sub-Saharan Africa has dropped sharply over the past 10 years, for females from 51.1 years to 46.3 years, and for males from 47.3 years to 44.8 years.

Apart from discrepancies in life expectancy between countries, there are also variations between certain groups within a country. For example, in most countries there is a significant difference in life expectancy between men and women, with women outliving men by several years. In 2004, life expectancy in the United States for females was 80.4 years and for males 75.2 years. From 1900 to the late 1970s, the sex gap in life expectancy widened from 2.0 years to 7.8 years. Since its peak in the 1970s, the sex gap has been narrowing and was 5.2 years in 2004, with men's life expectancy improving at a faster rate than that of women.

There are also significant differences in life expectancy between different racial and ethnic groups. For example, in 2004, the life expectancy in the United States for the African American population was 73.1 years. Life expectancy for the Caucasian population in the same year was 78.3 years, which results in a difference in life expectancy of 5.2 years. However, this difference has lessened in recent years. The widest Caucasian–African American life-expectancy gap was observed in 1989 and was 7.1 years. Among the four major race-sex groups in 2004, Caucasian females continued to have the highest life expectancy at birth (80.8 years), followed by African American females (76.3 years), Caucasian males (75.7 years), and African American males (69.5 years).

Poverty also has a substantial effect on life expectancy. In some countries, life expectancy is substantially longer in the wealthier areas compared with the poorer areas. Despite improvements in basic healthcare and improved access to medical care for the general population, the gap seems to be increasing as life expectancy continues to increase faster in wealthy communities relative to less prosperous communities.

Based on these facts, life expectancy figures are a useful statistical tool to summarize the current health status and development status of a population, which makes them useful for political issues. The above numbers are also of relevance for insurance and administrative purposes, since the aging population will be associated with an increase of incidence and prevalence of diseases observed especially in older patients. Such diseases could be malignant diseases or neurological diseases (e.g., Alzheimer's disease). As a result, these diseases will have growing importance for physicians in the future and might change the "landscape" of medical specialties.

Life Expectancy and Medical Decision Making

A physician is especially interested in the life expectancy that can be calculated at any time point in the life of a patient and that represents the number of years this patient is supposed to live throughout the rest of his or her life. This information is useful for several purposes.

First, when dealing with a patient in an advanced stage of an incurable disease, an estimation of life expectancy is necessary to judge if aggressive treatment is still indicated or if only palliative measures are to be taken. Similar considerations are applicable when dealing with an asymptomatic patient diagnosed with a slow progressing but potentially harmful disease, where a limited life expectancy would not allow for the disease to become symptomatic and therefore aggressive treatment is not warranted. Many such treatments have significant side effects, which may only be considered acceptable if a patient is likely to live long enough to experience any subsequent benefit.

Second, at diagnosis of a possibly incurable disease such as cancer, a patient would like to know if such disease might affect his or her life expectancy and, if yes, to what extent. This would allow the patient and his or her family to cope with impending death and make suitable plans for the remaining life span.

Third, predictions of life expectancy are relevant for the rational use of limited healthcare resources, since treating patients with expensive regimens without any benefit for the patient wastes valuable resources.

Most of the above considerations are of paramount importance in the treatment of prostate cancer. Prostate cancer is a very frequent disease, but the natural history of prostate cancer is rather favorable. Untreated prostate cancer patients often show clinical progression-free survivals of 10 to 15 years. This survival strongly depends on the grade of prostate cancer, where low-grade cancers show favorable long-term outcomes and high-grade cancers lead to progression and death. It can be postulated that treatment of patients diagnosed with a slow progressing disease such as prostate cancer and with insufficient life expectancy to experience disease-specific morbidity or mortality represents overtreatment. This overtreatment may unnecessarily add to costs, complications, early and late onset morbidities, and treatment-related mortality. Thus, from societal as well as individual perspectives, individuals with suboptimal life expectancy should not be considered for certain aggressive therapy options but should be offered more conservative and less harmful options or surveillance only. This consideration is particularly true for prostate cancer patients; however, it is also of relevance for the treatment of other diseases such as chronic organ failure (calling for transplantation), orthopedic problems, and so on.

The above considerations clearly indicate the importance of life expectancy–based treatment decision making and patient counseling. However, accurate prediction of life expectancy in patients represents a challenge. There are several possible bases for a physician's predictions about the life expectancy of a patient: the physician's clinical experience, life tables, comorbidity indices, and multivariable predictive tools.

Clinical Experience

To predict the life expectancy of an individual patient, a physician can rely on his or her personal knowledge and clinical experience. Several

publications addressing the ability of physicians to predict life expectancy of patients suggest that physicians have only poor ability to do so. For example, it has been shown that physicians overestimated the life expectancy of patients with metastatic cancers by nearly 100%. Another study showed that physicians were poor in predicting life expectancy in prostate cancer patients, where the survival beyond 10 years was under- or overestimated by 75% and 50%, respectively. Moreover, the overall accuracy of life-expectancy predictions was only around .68, where an accuracy of 1.0 would have been a perfect prediction and an accuracy of .5 would have been a prediction as good as a toss of a coin, which means pure chance. When estimating the life expectancy of an individual patient, a physician will most likely compare this patient's case with similar cases the physician has dealt with in the past. Unfortunately, such estimates are often biased by many factors, for example, the fact that positive events are more easily remembered than negative events (recall bias), that physicians might remember a unique patient rather than their general experience, or that they unconsciously want the predicted outcome to come true (control bias). To base medical decision making on such estimates is very likely associated with inappropriate decision making and may result in overtreatment as well as in undertreatment of individual patients.

Life Tables

Life tables represent the oldest and possibly most widely accessible tool for prediction of life expectancy. They may be used to calculate the average life expectancies of patients at selected ages. For example, based on the death rates observed in 2004, a person aged 50 years could expect to live an average of 30.9 more years for a total of 80.9 years. A person aged 65 years could expect to live an average of 18.7 more years for a total of 83.7 years. Recent publications evaluating the accuracy of life tables to predict life expectancy in patients treated for prostate cancer showed that life tables had only limited ability to predict the true length of survival. The survival of prostate cancer patients was overestimated by life tables at 5, 10, and 15 years after treatment by up to 50%, 36%, and 5%, respectively. The accuracy of the predictions was only between .60 and .65, which represents poor accuracy. There are several explanations for why these predictions are rather poor. As already mentioned above, the current death rates for life table calculations are frozen at the time of calculation, and no adjustment is done for life expectancy improvements in the future. This may result in departures from real survival. Moreover, life tables' predictions may be undermined due to grouping of all kinds of patients according to age strata. Patients can be in good or in poor health status, which might be associated with survival above or below average, respectively. Life tables do not adjust for these individual characteristics, which limits their applicability in daily practice. Finally, life tables are devised to provide an average number of remaining life years in a general population, and they are not devised to provide these predictions in one specific group of patients with one specific disease. Therefore, life tables represent an easily accessible tool for life-expectancy predictions, but their value for medical decision making in an individual patient is limited.

Comorbidity Indices

Another tool to improve predictions of life expectancy in patients is the use of comorbidity indices. Comorbidities are defined as the coexistence of two or more chronic conditions in a patient, and they are considered an important predictor of mortality. Comorbidity indices are devised to obtain a standardized categorization of comorbidities, which is then used to predict the risk of mortality in a population or in an individual patient. The Charlson Comorbidity Index is probably the most widely used comorbidity index. Further indices are the Chronic Disease Score, the Index of Coexisting Disease (ICED), the Cumulative Illness Rating Scale, the Kaplan-Feinstein Index, and the American Society of Anesthesiologists (ASA) physical status classification, to mention only a few. The most important feature of these comorbidity indices is their ability to predict overall survival independently of age. In other words, a certain comorbidity load may have a similar impact on survival in a young patient as it has in an older patient.

A limitation of comorbidity indices is the lack of proof that they are generalizable to all kinds of different populations. It is not certain that results achieved in a population used to develop a

comorbidity index and characterized by a certain disease can be extrapolated to another population characterized by another disease. Caution is recommended when these indices are used in patient populations other than that of the development population unless studies are available that confirm their reliability. In prostate cancer patients, most of the above-mentioned indices were evaluated and proved to be more or less similar in predicting the overall survival of patients. Their accuracy ranged from .61 to .68, which is equivalent to poor or moderate accuracy. This somewhat limited accuracy can be explained by the fact that comorbidities account significantly for overall mortality, but the variable age cannot be omitted without losing accuracy. Therefore, comorbidity indices provide a standardized and objective assessment of the risk of mortality in a patient, but their use in isolation is insufficient to predict life expectancy in an individual patient.

Multivariable Predictive Models

If the isolated use of a predictor is insufficient for medical decision making, the combined use of several predictors often improves the accuracy of life-expectancy predictions. Therefore, a further tool to improve predictions of life expectancy in an individual patient is a multivariable predictive model combining several predictors such as age, comorbidities, risk factors, health-conscious behavior, and so on. Here again, many models are available in the medical literature, some predicting survival in healthy patients, others in patients with certain early-stage diseases or advanced-stage diseases.

Generally, these multivariable predictive models are developed in a large group of patients with sufficient follow-up and a sufficient number of cases relevant to the outcome of interest. They are the result of a thorough and objective statistical analysis and therefore the risk of biased predictions is lower. The modes of practical application of these tools are manifold and vary between look-up tables, nomograms, mathematical formulas, online versions, and so on. The multivariable approach of these models allows considering several variables in a standardized fashion and improves the prediction of life expectancy. Again in prostate cancer patients, the accuracy of such models can vary from .69 to .84. This represents moderate to good accuracy in predicting the survival of patients. Compared with the other options for predicting life expectancy, these multivariable models seem to be the most appropriate tool to be used in medical decision making. However, as with the comorbidity indices, most of them were devised in a population with a certain disease characteristic, and generalizability is not assured unless studies provide this evidence.

Jochen Walz

See also Biases in Human Prediction; Decision Making in Advanced Disease; Disability-Adjusted Life Years (DALYs); Mortality

Further Readings

Albertsen, P. C., Hanley, J. A., & Fine, J. (2005). 20-year outcomes following conservative management of clinically localized prostate cancer. *Journal of the American Medical Association, 293,* 2095–2101.

Charlson, M. E., Pompei, P., Ales, K. L., & MacKenzie, C. R. (1987). A new method of classifying prognostic comorbidity in longitudinal studies: Development and validation. *Journal of Chronic Diseases, 40,* 373–383.

Chow, E., Davis, L., Panzarella, T., Hayter, C., Szumacher, E., Loblaw, A., et al. (2005). Accuracy of survival prediction by palliative radiation oncologists. *International Journal of Radiation Oncology, Biology, Physics*, *61,* 870–873.

Cowen, M. E., Halasyamani, L. K., & Kattan, M. W. (2006). Predicting life expectancy in men with clinically localized prostate cancer. *The Journal of Urology, 175,* 99–103.

Greenfield, S., Apolone, G., McNeil, B. J., & Cleary, P. D. (1993). The importance of co-existent disease in the occurrence of postoperative complications and one-year recovery in patients undergoing total hip replacement: Comorbidity and outcomes after hip replacement. *Medical Care, 31,* 141–154.

Henderson, R., Jones, M., & Stare, J. (2001). Accuracy of point predictions in survival analysis. *Statistics in Medicine, 20,* 3083–3096.

Johansson, J. E., Andren, O., Andersson, S. O., Dickman, P. W., Holmberg, L., Magnuson, A., et al. (2004). Natural history of early, localized prostate cancer. *Journal of the American Medical Association, 291,* 2713–2719.

National Center for Health Statistics. (2008). *Health, United States, 2007*. Hyattsville, MD: Author. Retrieved January 8, 2009, from http://www.cdc.gov/nchs/fastats/lifexpec.htm

Walz, J., Gallina, A., Perrotte, P., Jeldres, C., Trinh, Q. D., Hutterer, G. C., et al. (2007). Clinicians are poor raters of life-expectancy before radical prostatectomy or definitive radiotherapy for localized prostate cancer. *BJU International, 100,* 1254–1258.

Walz, J., Gallina, A., Saad, F., Montorsi, F., Perrotte, P., Shariat, S. F., et al. (2007). A nomogram predicting 10-year life expectancy in candidates for radical prostatectomy or radiotherapy for prostate cancer. *Journal of Clinical Oncology, 25,* 3576–3581.

Likelihood Ratio

Understanding the performance of diagnostic tests is fundamental to clinical decision making. The primary measures for assessing diagnostic test performance are sensitivity, specificity, positive and negative predictive value, and likelihood ratios. The likelihood ratio (*LR*) incorporates sensitivity and specificity into a single parameter and allows one to determine how much a positive or negative test result changes the likelihood that a patient has the disease of interest. Specifically,

$$LR(+) = \text{sensitivity}/(1 - \text{specificity})$$
(or true positive rate/false positive rate);

$$LR(-) = (1 - \text{sensitivity})/\text{specificity}$$
(or false negative rate/true negative rate).

The likelihood ratio positive indicates how much more likely a positive test result will be seen in someone with disease, relative to someone without the disease of interest. As a rule of thumb, a likelihood ratio positive above 10 provides strong evidence to rule in a diagnosis in most circumstances. Similarly, the likelihood ratio negative indicates how much less likely a negative test result will be for someone with disease relative to someone without the disease of interest. Generally speaking, a likelihood ratio negative below 0.1 provides strong evidence to rule out a diagnosis. Using the following adaptation of Bayes's theorem, one can calculate the posttest probability of disease:

$$\text{Posttest odds} = \text{Pretest odds} \times LR$$
(pretest odds = $p/(1 - p)$, where p = prior probability of disease);

$$\text{Posttest probability} = \text{Posttest odds}/(\text{Posttest odds} + 1).$$

Examples

Example 1

Consider the impact of a testing for recurrent primary colorectal cancer with stool guaiac cards among asymptomatic subjects following a previous curative resection. The prior probability (p) of colorectal cancer in this population is approximately .04 (and pretest odds = $p/(1 - p)$ = .04/.96 = .042). A set of six stool guaiac cards has a sensitivity of .25 and a specificity of .97 in detecting asymptomatic cancer (in subjects not on nonsteroidal anti-inflammatory drugs).

For a positive test:

$$LR(+) = \text{sensitivity}/(1 - \text{specificity}) = .25/.03 = 8.3;$$

$$\text{Posttest odds} = \text{Pretest odds} \times LR = .042 \times 8.3 = .35;$$

$$\text{Posttest probability} = .35/(1 + .35) = .26.$$

For a negative test:

$$LR(-) = (1 - \text{sensitivity})/\text{specificity} = .75/.97 = .77;$$

$$\text{Posttest odds} = \text{Pretest odds} \times LR = .042 \times .77 = .032;$$

$$\text{Posttest probability} = .032/(1 + .032) = .03.$$

Thus, a positive guaiac series substantially increased the probability of recurrent cancer from 4% to 26%, whereas a negative guaiac series only modestly reduced the probability of recurrent cancer from 4% to 3%. The high false negative rate of the test (1 – sensitivity) accounts for its failure to substantially lower the probability of disease.

In some cases, it is more useful to calculate likelihood ratios for multiple categories (rather than using a single cutpoint).

Example 2

A 9-month-old infant presents to the emergency department with fever and increased irritability. The white blood cell (WBC) count is markedly elevated (21,000). What is the possibility that this infant has bacterial meningitis or bacteremia? Table 1 shows the distribution of WBC results among 2,240 febrile infants (63 infants with confirmed bacterial meningitis or bacteremia and 2,177 infants without either condition).

For each interval, the probabilities for results within that interval were used to calculate interval likelihood ratios (*iLR*s). Inspection of the *iLR*s reveals that infants with *both* high and low WBC counts have an increased likelihood of having bacterial meningitis or bacteremia; this information would have been lost had a single cutoff been selected. Based on the prior probability of bacterial meningitis or bacteremia in febrile infants at the institution where the study was conducted (3%, which is equal to a prior odds of .031), one can estimate the posttest probability of disease for the patient as follows:

$$\text{Posttest odds} = \text{Prior odds} \times iLR\ (\geq 20{,}000) = .031 \times 3.4 = .105;$$

$$\text{Posttest probability} = .105/(.105 + 1) = .095.$$

Implications

Thus, likelihood ratios allow one to predict the risk of disease, given a particular test result, across varying prior probabilities of disease. Indeed, likelihood ratios tend to be more stable than sensitivity and specificity to changes in the prevalence of disease. Moreover, likelihood ratios allow one to revise disease probability for a sequence of diagnostic tests, that is, where the posttest odds for one test become the pretest odds for a second, *independent* diagnostic test. There are a number of methodological issues that may affect the accuracy of likelihood ratios for probability revision, however, including verification bias, ascertainment bias, incorporation bias, and spectrum bias. For example, spectrum bias may increase the sensitivity of a diagnostic test (with generally little impact on

Table 1 White blood cell count as a predictor of bacterial meningitis or bacteremia among febrile infants

WBC Count (per mm^3)	*Meningitis or Bacteremia*		*Likelihood Ratio*
	Yes	*No*	
<5,000	5	96	
	8%	4%	2.0
5,000–9,999	18	854	
	29%	39%	0.7
10,000–14,999	8	790	
	12%	36%	0.3
15,000–19,999	17	286	
	27%	13%	2.1
≥20,000	15	151	
	24%	7%	3.4
Total	63	2,177	
	100%	100%	

Note: Percentages add to 99% in the "No" column due to rounding.

specificity) if the study sample is skewed toward patients with more serious illness; this bias would be expected to increase $LR(+)$ for a given diagnostic test. Much work remains to be done to compile likelihood ratios across diagnostic technologies in different populations (with differing spectra of illness) and to improve access to likelihood ratios and pretest probabilities of disease for clinical decision making.

David A. Katz

See also Diagnostic Tests; Positivity Criterion and Cutoff Values

Further Readings

Deeks, J. J., & Altman, D. G. (2004). Diagnostic tests 4: Likelihood ratios. *British Medical Journal*, *329*, 168–169.

Jaeschke, R., Guyatt, G., & Lijmer, J. (2002). Diagnostic tests. In G. Guyatt & D. Rennie (Eds.), *Users' guides to the medical literature: Essentials for evidence-based clinical practice* (pp. 187–217). Chicago: AMA Press.

Katz, D. A. (1999). A primer on measures of treatment effectiveness and diagnostic test performance. *Wisconsin Medical Journal, 98*(2), 37–43.

Newman, T. B., Browner, W. S., & Cummings, S. R. (2007). Designing studies of medical tests. In *Designing clinical research* (3rd ed.). Philadelphia: Lippincott Williams & Wilkins.

Sackett, D. L., Haynes, R. B., Guyatt, G., & Tugwell, P. (1991). *Clinical epidemiology: A basic science for clinical medicine* (2nd ed.). Boston: Little, Brown.

Whiting, P., Rutjes, A. W. S., Reitsma, J. B., Glas, A. S., Bossuyt, P. M., & Kleijnen, J. (2004). Sources of variation and bias in studies of diagnostic accuracy: A systematic review. *Annals of Internal Medicine, 140*(3), 189–202.

Logic Regression

Logic regression is an adaptive regression and classification tool to address problems arising when data of mostly binary covariates are analyzed and the interactions between these predictors are of main interest to predict future outcomes or to identify variables that are associated with a particular outcome. Logic regression should not be confused with logistic regression: For logistic regression, the response is binary; for logic regression, the covariates are binary, but the response and the regression model can have any form. Binary covariates arise in many medical settings, such as the diagnosis of disease using phenotypic features, the identification of factors that contribute to emergency room crises, and the identification of genotypes that are associated with a particular disease. Often, the interaction between those binary predictors is of particular interest. Given a set of binary covariates, logic regression creates new predictors by considering Boolean ("logic") combinations of the binary covariates and has the capability to embed those into a regression framework. As an example, this allows for statements such as "the odds of suffering an adverse response in the emergency room are three times higher for subjects above 65 years of age who have high blood pressure or breathing problems." The logic regression framework includes many forms of classification and regression (such as linear and logistic regression, the Cox proportional hazards model, and more). In general, any type of model can be considered, as long as an objective (scoring) function can be defined. The model search is carried out using simulated annealing, a stochastic search algorithm commonly used in high-dimensional data problems. Model selection is performed via cross-validation or permutation tests, which implicitly address multiple comparisons problems. A Markov chain Monte Carlo–based extension of logic regression to create ensembles of plausible covariate combinations and measures of variance importance has also been implemented. The logic regression software is freely available as a contributed package to the statistical environment, R, and can be downloaded from the Comprehensive R Archive Network.

Description

In many medical and public health–related settings, a number of binary variables are collected, and the aim is the prediction of a particular response or the selection of covariates associated with the response. The former includes, for example, the task to predict which incoming patient should be admitted to critical care from a set of medical markers and records and the determination of what conditions

are responsible for emergency room crises. Genomic studies, in particular, single nucleotide polymorphism (SNP) association studies, are an instance of the latter. For example, researchers studied the relation between 88 SNPs and their association with restenosis development among 779 subjects, with the main question of interest being the search for a combination of SNPs that best explains the variation in the phenotype. In these settings, the interaction of several of those binary variables often predicts the outcome or explains the relationship of relevant covariates to the outcome better than the individual covariates alone. Thus, from a statistical perspective, this represents a very challenging task, since in a typical setting the number of possible interactions between the predictors can be immense.

Logic regression is an adaptive regression and classification tool to address exactly these types of problems. Given a set of binary covariates, logic regression creates new predictors for the response by considering Boolean combinations of the binary covariates. One of its most appealing features is that the Boolean terms can be searched for using a regression framework, and valid statistical inferences can be made. The resulting model incorporating these interactions typically is very easy to interpret. The logic regression approach and its usefulness can best be explained in the context of SNP association studies. Statistical approaches to evaluate higher-order interactions between SNPs, or between SNPs and environmental variables, are critical for analyzing complex diseases, as higher susceptibility is likely to be related to the interaction of multiple SNPs and environmental factors. SNPs are typically recorded as a single binary variable (wild-type vs. variant) or two dummy variables using dominant and recessive coding. For example, a very simple Boolean term for a model in the restenosis example is *"nonwild-type SNP CBS and wild-type SNP TP53,"* indicating that the combination of at least one variant allele in SNP CBS and no variants in SNP TP53 is associated with higher susceptibility to disease. The effect sizes seen in complex diseases are typically very small, and therefore the power to detect those small effect sizes can crucially depend on whether methods to simultaneously investigate SNPs and environmental variables are employed and on how the stochastic model is specified. Prediction of the outcome status in this context is often not of primary importance, and not bound for much success, as the small effect sizes almost invariably result in poor sensitivity and specificity for any prediction method. That said, any method (including prediction approaches) that generates some measure of variable importance can be used successfully in this setting, when there is an emphasis on variable selection.

The framework of logic regression includes many forms of classification and regression (such as linear and logistic regression, the Cox proportional hazards model). In general, any type of model can be considered, as long as an objective function can be defined (such as likelihoods or partial likelihoods, deviances, or residual sums of squares). The model search is carried out using a simulated annealing algorithm, and model selection is performed via cross-validation and permutation tests, which implicitly address the multiple comparisons problems. The model search and model selection procedures result in a single model that specifies the relationship of the predictor variables with the response. However, there might be many plausible models. For example, if in the model described above *SNP TP54* was in strong linkage disequilibrium (highly correlated) with *SNP TP53*, then the Boolean term *"nonwild-type SNP CBS and wild-type SNP TP54"* could also be plausible. This issue is addressed in the Markov chain Monte Carlo–based extension of logic regression, creating ensembles of plausible covariate combinations and deriving measures that assess the importance of single covariates as well as higher-order terms.

Model Search and Model Selection

The search for good scoring models is carried out using a simulated annealing algorithm. This probabilistic search strategy is based on a move set that allows for alterations of the Boolean terms in a model (such as adding a variable, deleting a variable, or changing an *and* into an *or*). Two examples for the restenosis data are *"nonwild-type SNP CBS or wild-type SNP TP53"* and *"nonwild-type SNP CBS and (wild-type SNP TP53 or variant SNP CBS)."* The model with the new Boolean term replaces the current model in the search if it has a better score according to the objective function. Otherwise, it replaces the current model with a

probability less than 1. This probability depends on the difference between the scores of the competing models and an annealing parameter that governs the stage of the search procedure—the further the annealing has progressed, the less likely it is that the new model will be accepted if the score worsens. The obvious advantage of such a stochastic annealing approach compared with greedy search strategies is the fact that local minima can be avoided (at the price of increased computing time).

Model selection has to be employed to avoid overfitting, and for this purpose a definition of model size is required. For a fixed number of Boolean terms in a logic model, the model size is defined as the combined total number of variables in all Boolean terms of the model (or equivalently, the total number of leaves in the tree representation of the Boolean term). Candidate models of various sizes are generated by prohibiting moves in the model search algorithm that result in models beyond the allowed complexity. Among the candidate models, either the model with the best predictive performance is selected (using cross-validation, if prediction is of main interest) or, alternatively, the largest model that does not overfit the data (as assessed by permutation tests, if association is of main interest). The predictive performance can, of course, also be assessed using a training and test set approach when sufficient data are available.

Comparison With Other Approaches

There is a wealth of approaches for regression and classification problems in the machine learning, computer science, and statistics literature that can be used when most or all of the covariates are binary, and some of these approaches also involve modeling interactions between those predictors. Boolean functions of these binary covariates in particular have played a major role in the machine learning literature. Most of these methods, however, are intended for classification and prediction only and are not embedded in a regression framework. In particular, in contrast to logic regression, many of these approaches are only applicable for binary outcomes. Noteworthy exceptions are, for example, MARS and CART (and derivations thereof, such as random forests and boosting), which work for continuous outcomes as well. Other approaches have been augmented to handle continuous outcomes by transforming the problem back into a classification setting (such as the extension SWAP1R of the original SWAP1 algorithm, which learns regression rules in disjunctive normal form). Tree-based models that have linear equations instead of numeric values in the terminal nodes are known as treed models in the statistical literature; however, some similar approaches (either rule-based, such as R^2, or tree-based, such as M5) have also been proposed in the computer science literature.

An important aspect that differentiates logic regression from most other approaches is that higher-order interactions between binary variables can be detected and valid statistical inference about their *association* to a response variable can be made, using a regression framework. As previously described, this property is particularly useful in SNP association studies (the susceptibility for complex diseases is likely to be related to the interaction of multiple SNPs and environmental factors), where the sensitivity and specificity of prediction methods are very low but statistically significant associations (e.g., departures from randomness) can often be detected. That said, many tools from the statistical learning literature, developed to deal with high-dimensional search spaces, have been adapted or extended and applied to multimarker SNP data (this includes, e.g., neural networks and random forests). And while these approaches do not generate interpretable genetic models and are usually only used for prediction purposes, they do allow for the generation of measures of variable importance, which can be very useful for variable selection in these association studies. Other methods to elucidate SNP-SNP and SNP-environment interactions directly have also been proposed, such as the multifactor dimensionality reduction (MDR) technique. The latter differs from logic regression in several ways. It aims at finding interactions that decrease the heterogeneity in the response among the subclasses it defines; however, it is not embedded in a regression setting. Moreover, another very important distinction that it shares with almost every other approach (for prediction and classification algorithms in general and methods for SNP association studies in particular) is the fact that it performs a greedy search. This has the advantage of fast computing times; however, it comes with the high risk of being trapped in local extrema.

Software

The logic regression software is freely available as an R package from the Comprehensive R Archive Network, which also includes functions to carry out Monte Carlo logic regression and contains the logic regression manual. Further support (such as examples and help files) is available from the logic regression Web site. Note that the software requires complete records, and thus, other approaches such as imputations have to be employed first if data are missing and/or a complete case analysis is not desirable.

Ingo Ruczinski and Charles Kooperberg

See also Decision Trees, Advanced Techniques in Constructing; Logistic Regression; Prediction Rules and Modeling; Recursive Partitioning

Further Readings

Comprehensive R Archive Network: http://cran.r-project.org

Dai, J., Ruczinski, I., LeBlanc, M., & Kooperberg, C. (2006). Imputation methods to improve inference in SNP association studies. *Genetic Epidemiology, 30*(8), 690–702.

Kooperberg, C., & Ruczinski, I. (2005). Identifying interacting SNPs using Monte Carlo logic regression. *Genetic Epidemiology, 28*(2), 157–170.

Logic Regression: http://bear.fhcrc.org/~ingor/logic

Ruczinski, I. (2000). *Logic regression and statistical issues related to the protein folding problem.* Unpublished PhD thesis, University of Washington, Seattle.

Ruczinski, I., Kooperberg, C., & LeBlanc, M. (2003). Logic regression. *Journal of Computational and Graphical Statistics, 12*(3), 475–511.

Ruczinski, I., Kooperberg, C., & LeBlanc, M. (2004). Exploring interactions in high dimensional genomic data: An overview of logic regression, with applications. *Journal of Multivariate Analysis, 90*, 178–195.

Logistic Regression

Often the aim of medical research projects is to study the relationship between an outcome variable on the one hand, say Y, and one or more other variables on the other hand, say $X_1, X_2, \ldots, X_p$. For instance, Y = diastolic blood pressure, and X_1 = age, X_2 = gender, X_3 = body weight, X_4 = salt intake, and so on. In this context, the outcome variable Y is also called the dependent variable or the response variable. The Xs are also called independent variables, explanatory variables, predictor variables, or simply covariates. Based on a sample of subjects from the population of interest in which the dependent and independent variables are observed, the statistical relationship between the variables can be studied using a statistical regression model. The subjects might comprise a completely random sample from the population, or they might be selected on the Xs. For instance, the sample might be stratified on age and gender, but given the Xs the subjects are supposed to be randomly sampled. In general, a statistical regression model describes in mathematical terms the distribution of Y given the Xs. There are numerous regression models available. Every regression model is meant for a specific type of outcome variable. For instance, if Y has a normal distribution, the appropriate model is the linear regression model, which is the most well-known regression model. It assumes that the mean of Y is a linear combination of the Xs and unknown parameters, $\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p$. This expression is called the linear predictor and is an ingredient of almost all regression models. Very often in medical research the outcome variable is dichotomous, say $Y = 1$ if some event of interest has occurred, and $Y = 0$ if the event did not occur. For instance, in epidemiology, the outcome event mostly is the occurrence of a certain disease. In clinical research, many clinical trials have a dichotomous endpoint, for instance, the patient is cured or not, the patient survives or not, an adverse drug reaction occurs or not, and so on. The logistic regression model is the most well-known regression model for dichotomous outcome. Although there are several other regression models available for dichotomous outcome, in medical research it is by far the most-used model.

The Model

Let $\pi = P(Y = 1)$ denote the probability that the event occurs for a subject. This probability might

depend on the values of the predictor variable of the subject; thus, $\pi(x) = P(Y = 1|x)$, where x stands for the values of all the predictor variables of a subject. The most straightforward regression model, analogous to the linear regression model, would be $\pi = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$. Using standard likelihood methods, the parameters could be estimated, confidence intervals constructed, and null hypotheses tested. Though the βs have a clear interpretation as risk differences and the model might describe the data well in many instances, it is hardly used in practice. The disadvantage of this model is that the linear predictor on the right-hand side of the equation is not necessarily restricted to a number between 0 and 1 but might be any number between $-\infty$ and $+\infty$. Therefore, it might happen that the model predicts probabilities smaller than 0 or greater than 1. A way to get around this is to transform π through a monotone function g that maps the interval (0, 1) into $(-\infty, +\infty)$, leading to models of the form $g(\pi) = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p$. The transformation g is called the link function. It links probabilities to values of the linear predictor. Alternatively, the model is written as $\pi = f(\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p)$, with f the inverse of g, called the inverse link function. Numerous choices for g or f are possible, each leading to another regression model for dichotomous outcome. Well-known choices for the link function are the $g(\pi) = \ln(-\ln(\pi))$, leading to a model known as the complementary log-log model, and the inverse cumulative standard normal distribution function leading to what is called the probit model. The latter model is very popular in economics and social sciences. However, in medical research by far the most popular choice is the logit (or log-odds) link function, $g(\pi) = \ln(\pi/(1 - \pi))$. This leads to the *logistic regression model*:

$$\ln\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p.$$

Alternatively the model is written as

$$\pi = \frac{\exp(\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p)}{1 + \exp(\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p)}.$$

Since the inverse link function $f(z) = \exp(z)/(1 + \exp(z))$ is the cumulative distribution function of a logistic distribution, the model is called the logistic regression model.

Interpretation of the Parameters

Dichotomous Predictor Variable

Let X_1 be a dichotomous predictor variable, for instance, $X_1 = 1$ for women and $X_1 = 0$ for men. According to the model, the predicted odds are

$$\frac{\pi}{1-\pi} = \exp(\beta_0 + \beta_1 + \beta_2 x_2 + \ldots + \beta_p x_p) \text{ for a woman,}$$

$$\frac{\pi}{1-\pi} = \exp(\beta_0 + \beta_2 x_2 + \ldots + \beta_p x_p) \text{ for a man.}$$

Thus, if a woman and a man have the same values on the other predictors $X_2, \ldots, X_p$, the odds ratio of a woman relative to a man is

$$OR_{X_1 = 1 \text{ VERSUS } X_1 = 0} = \exp(\beta_1).$$

Apparently, for a dichotomous predictor variable X taking values 0 and 1, the corresponding β is interpreted as the log-odds ratio comparing category $X = 1$ with category $X = 0$, keeping the values of the other predictor variables fixed. Since the other predictor variables are kept equal in the comparison of $X = 1$ to $X = 0$, the (log) odds ratio is said to be "adjusted" or "corrected" for the other predictor variables.

Polytomous Predictor Variable

Suppose X is a categorical predictor variable with three or more categories. For instance, X is the race of a subject labeled as $X = 1$ for white, $X = 2$ for black, and $X = 3$ otherwise. There are many possibilities to represent a categorical variable in a regression model. The most popular way is through what are called dummy variables. For example, two dummy variables are defined as

$X_1 = 1$ if race is black, and $X_1 = 0$ otherwise;

$X_2 = 1$ if race is other, and $X_2 = 0$ otherwise

and put into the model. There may be other predictor variables $X_3, \ldots, X_p$ in the model as well. The two dummy variables here correspond to the categories black and other, but this choice is arbitrary. The predicted odds for the different races are

$$\frac{\pi}{1-\pi} = \exp(\beta_0 + \beta_3 x_3 + \ldots + \beta_p x_p) \text{ for white race,}$$

$$\frac{\pi}{1-\pi} = \exp\left(\beta_0 + \beta_1 + \beta_3 x_3 + \ldots + \beta_p x_p\right) \text{ for black race,}$$

$$\frac{\pi}{1-\pi} = \exp\left(\beta_0 + \beta_2 + \beta_3 x_3 + \ldots + \beta_p x_p\right) \text{ for other race.}$$

Thus, comparing black with white subjects who have the same values on all other predictor variables gives an odds ratio $\exp(\beta_1)$ or a log-odds ratio β_1. Comparing a subject with other race to white subjects keeping the other predictor variables fixed leads to an odds ratio $\exp(\beta_2)$ or a log-odds ratio β_2. Note that the category for which no dummy variable is made automatically serves as the reference category. The choice of the reference category is arbitrary. In general, if a categorical predictor variable has k categories, first one of them is chosen to be the reference. For the other categories a dummy variable is made. The β of a dummy variable is then interpreted as the log-odds ratio comparing the corresponding category with the reference category, adjusted for the other independent variables in the model. The above way of defining the dummy variables is only one way of representing a categorical predictor in a regression model, albeit the most popular one. Other definitions lead to other interpretations of the corresponding βs. Many logistic regression computer programs automatically set up dummy variables for categorical predictors. It is then essential that the user know exactly how these are defined, since otherwise the corresponding βs cannot be interpreted.

Continuous Predictor Variable

Consider a continuous predictor variable, say X_1 = age. Then, following the same reasoning as for a dichotomous predictor, it can be seen that the corresponding β can be interpreted as the odds ratio of a subject with arbitrary age as compared with a subject who is one year younger, adjusted for all other predictors in the model. In general, for a continuous predictor X, $\exp(\beta)$ is interpreted as the factor by which the odds are multiplied if X increases with one unit, or β is interpreted as the increase in log odds of the event per unit increase of the continuous predictor variable.

Fitting the Model

The parameters and corresponding standard errors are estimated following the standard maximum likelihood method. Only in very simple cases is it feasible to do the calculations without using a computer. Confidence intervals can be calculated and hypothesis tests can be carried out by one of the three methods that are available within the likelihood theory: Wald's method, the likelihood ratio method, or the score method. Statistical software packages always provide the results from the method of Wald. Some packages give in addition the results of the score and/or likelihood ratio test as well.

Example

In a study on complications of bone marrow transplantation, interest was in predicting factors for the occurrence of acute graft-versus-host disease (AGVHD). A study group of 166 bone marrow transplantation patients was available in which the occurrence of AGVHD was observed together with a large number of potential prognostic variables. One of the models that were fitted predicted AGVHD on the basis of the following three predictor variables:

AGEDON (age of the donor [years])

MATCH (= 1 if sexes of donor and recipient match, = 0 if they do not match)

DIAG (underlying diagnosis: A = severe aplastic anemia, B = acute nonlymphoblastic leukemia, C = acute lymphoblastic leukemia)

The diagnosis is a categorical predictor that was represented in the model by two dummy variables indicating diagnosis A (DIAG_A) and B (DIAG_B), while C was serving as reference category. The results are given in the following table.

Predictor Variable	*b*	*SE*	*p Value*	*exp(b)*	*95% CI for exp(b)*	
					Lower	*Upper*
DIAG_A	2.174	.662	.001	8.794	2.401	32.194
DIAG_B	0.816	.556	.142	2.261	0.761	6.723
AGEDON	0.036	.015	.016	1.037	1.007	1.068
MATCH	–0.723	.361	.045	0.485	0.239	0.984
CONSTANT	–2.153	.582	.000	0.116		

The second column, with heading *b*, gives the estimates of the regression coefficients β. The CONSTANT is the estimate of β_0. For instance, the regression coefficient $b = -0.723$ corresponding to MATCH is the log-odds ratio of sex being matched relative to sex not being matched. The corresponding odds ratio is exp(–0.723) = 0.485 and is given in the fifth column. It means that if sex of donor and patient are matched, the odds of AGVHD are about half of the odds of a patient with unmatched sex, adjusted for the age of the donor and for the underlying diagnosis. Or, in other words, if the sexes are not matched, the odds of AGVHD are $1/0.485 \approx 2$ times higher than if they are matched. The approximate 95% confidence interval according to the method of Wald is given in the last two columns. The true odds ratio of sexes being matched versus unmatched is with 95% confidence between 0.239 and 0.984. The (Wald) *p* value for the null hypothesis H_0: $\beta_{MATCH} = 0$ is given in Column 4. Thus, the odds ratio of MATCH is just statistically significantly smaller than 1 at the .05 significance level. The dummy variable DIAG_A compares diagnosis A with C. The odds ratio 8.794 means that the odds of AGVHD of patients with diagnosis A are almost 9 times higher than for patients with diagnosis C. The coefficient of AGEDON means that the log-odds of AGVHD increases by 0.036 for each year that the donor is older, when the values of the other predictors are kept fixed. In other words, the odds of AGVHD increase by a factor exp(0.036) = 1.007 (Column 5) per year increase of age of the donor. The results of the model also can be used to predict the probability of a future patient getting AGVHD, depending on his or her values on the predictor variables. For instance, for a patient with diagnosis A who is going to have a 37-year-old donor of the same sex, the predicted probability of AGVHD is

$$\frac{\exp(-2.153 + 2.174 \times 1 + 0.816 \times 0 + 0.036 \times 37 - 0.723 \times 1)}{1 + \exp(-2.153 + 2.174 \times 1 + 0.816 \times 0 + 0.036 \times 37 - 0.723 \times 1)} = .652.$$

Remarks

Case-Control Studies

A remarkable property of the logistic model is that it can be used to model the probability of being a case, while the βs have the same interpretation as for cohort study data (except for β_0, which has a different interpretation). Often, in particular in case-control study settings, events or cases are rare. Then it is allowed to interpret odds ratios as risk ratios or relative risks.

Exact Logistic Regression

The usual way of fitting logistic models is by likelihood methods, which are approximate methods. This puts restrictions on the number of β parameters that are allowed. An often-used rule of thumb is that the number of βs should not be larger than the square root of the number of observed events or number of events divided by 10. Otherwise, exact logistic regression should be used. Exact logistic regression is computationally extensive and software for it is scarce.

Conditional Logistic Regression

This is a modification of ordinary logistic regression, useful for stratified data with many strata, such as in matched case-control studies.

Theo Stijnen

See also Cox Proportional Hazards Regression; Logic Regression; Ordinary Least Squares Regression

Further Readings

Hosmer, D. W., & Lemeshow, S. (2000). *Applied logistic regression* (2nd ed.). New York: Wiley.

Log-Rank Test

The log-rank test is a statistical test used to compare two or more groups of subjects with respect to the corresponding durations of time to the occurrence of a specific event. The log-rank test is frequently used in medical research to compare the survival times between patients with the same condition or disease who can be grouped by treatment, age, sex, or some other factors such as specific disease characteristics. Similar to other statistical tests, the log-rank test result is reported as a p value between 0 and 1. A smaller p value indicates stronger evidence for the differences between the tested groups. A commonly used convention is to interpret a p value of .05 or less as evidence of statistically significant differences.

Background

Statistical tests do comparisons between different populations by determining how likely the observed differences in the data collected from those populations are due to chance only, under the assumption that those populations are the same with respect to the quantity under comparison (the null hypothesis). A p value of .05 means that, under the null hypothesis, the chance of observing such a difference in the collected data is 5%. The power of a statistical test is the chance of it correctly rejecting the null hypothesis under a specified difference between the populations under testing. For any well-designed statistical test, a larger sample size results in greater power.

Some statistical tests assume that the distribution of the quantity of interest has a certain mathematical format. For survival times, this can be Weibull, log-normal, gamma, or others. Such tests are called parametric tests. Other statistical tests do not make such distribution assumptions. These tests are called nonparametric tests. A common way to conduct a nonparametric test is to combine and then rank the data from the populations under study. The test statistic then depends only on the ranks of the data, not on the exact value of the data. Such tests are called rank tests. The log-rank test is a rank test. It does not require specification of the underlying survival time distribution. Nonparametric tests are thought to be more robust because they are not subject to bias arising from misspecification of the parametric distributions. Parametric tests, on the other hand, can be more powerful.

A survival function specifies the probability that the survival time is greater than a given number. The most common way of estimating a survival function is the Kaplan-Meier method. And the most common way to compare different estimated survival functions is the log-rank test.

History

The log-rank test was first derived by Nethan Mantel in 1966. The name *log-rank* was first used by Richard Peto and Julian Peto in 1972. They showed the optimality of the test under certain conditions. The "log" in the test name comes from the fact that, at a given time point, the factor used in the test, "the number of failures divided by the number of subjects at risk," is an estimate for the change of the logarithm of the survival function at that time point. Stringent justification of the properties of the log-rank test needs modern mathematical theory.

Failure and Censoring

The branch of statistical methods dealing with time-to-event data is called *survival analysis*. The event of interest can be death, disease onset, disease recurrence, or some other well-defined end point. It is usually called "failure." However, the "failure" can be a positive event in some applications, such as the end of unemployment in economics studies.

A unique feature of time-to-event data is that they are subject to censoring. When the event of interest does not happen to a subject during the study period, the time-to-event for that subject is censored. This does not mean that the subject's participation in the study does not provide any useful information. It still provides the information that the time-to-event of this subject is longer than the censoring time. The presence of censored time-to-event data precludes the use of other commonly used statistical tests, such as the *t* test. Simply ignoring censored data can result in biased test results. For example, when the follow-up times are the same for all subjects, the data from subjects with a longer time-to-event are more likely to be censored. The log-rank test is used to solve this problem.

Underlying Assumptions

An important implicit assumption for the log-rank test is independent censoring. This requires that, at any given time point, subjects cannot be censored because they appear to have a higher or lower failure risk.

While a small *p* value indicates differences between the groups being tested, a large *p* value does not necessarily mean that no differences exist. In some scenarios, the groups being tested have differences in their failure risks, but the log-rank test cannot detect those differences. For example, suppose that, as compared with Group B, subjects in Group A have a higher failure risk early on but a lower risk at later stages. In such a case, the log-rank test may fail to give evidence for the differences between these two groups. When the failure risks over time for one group are proportional to those for the other, the log-rank test is the most powerful test to detect such a difference. The log-rank test places a uniform weight on the between-group differences in the early and late stages and sums them up. The Wilcoxon test places higher weights on early differences. Thomas Fleming and David Harrington and others proposed alternative tests that place more weight on the differences occurring in the middle or late stages.

When to Use a Log-Rank Test

Different statistical tests should be used for different types of data. When the data are continuous and roughly normally distributed, the *t* test can be used. For count data in contingency tables, when the counts are not too small, the chi-square test can be used. If some of the counts in a contingency table are too small (< 5), then the Fisher's exact test can be used. When the data are about time durations to the occurrence of an event, the log-rank test can be used. To simultaneously estimate the effects of multiple factors on the time-to-event, the Cox proportional hazards model can be used. The testing results from a Cox model with only group indicators as independent variables (or predictors) are the same as those resulting from a log-rank test.

Procedures

To prepare data for a log-rank test, first, a meaningful and well-defined time origin (or baseline time) is chosen. For example, in medical research, the time origin is usually the date of diagnosis or the date treatment begins. Then two variables are used to record the data. One is a time variable that measures for each subject in the study the duration from the time origin to the failure event or the end of follow-up, whichever happens first. The other is an indicator variable, which shows whether the time variable for each subject represents failure or censoring.

The procedures of the log-rank test are as follows. First, the data set is sorted according to the time variables in ascending order. All the subsequent time orders will refer to this sorted order (as opposed to chronological order by calendar time). At any time point, all the subjects who still remain in the study (have not failed or been censored yet) are called at risk (of failure). Consider a simple case of comparison between two groups. Suppose the two groups have the same distribution of time to failure (the null hypothesis). Then at any time point, all the subjects who are still at risk have the same chance to fail. Thus, conditional on the observed total number of failures from both groups at a given time point, the expected number of failures in any group should be proportional to the number of at-risk subjects in that group, while the sum of the expected numbers of failures in the two groups should be the observed total number of failures at that time point. Choosing one of the two groups, the difference between its observed and expected numbers of failures at a time point is

calculated. The variance of this difference is estimated. The log-rank statistic is the quotient of the square of the sum of the above differences at all failure time points (from the two groups combined), divided by the sum of the estimated variance at all failure time points. Mathematically, choosing the other group in the above calculation will give the same result. Under the null hypothesis, the differences between the observed and expected number of failures should be small. A large sum of such differences indicates deviation from the null hypothesis. Consequently, a large log-rank statistic is evidence for differences between the two groups. Then the question is how large is "large." A distribution table answers this question by matching each value of a statistic to a p value that describes its ranking position among all possible values of the statistic. Note that here the tail probability of a distribution is used. This means that a large statistic corresponds to a small p value. When the sample size is large, the log-rank statistic has a chi-square distribution. This distribution is independent of the sample size, as long as it is sufficiently large. A conservative rule is that, if the number of observed failure times in each group is greater than or equal to 30, then the sample size is sufficiently large. If this sample size condition is met, then the corresponding p value can be obtained from the chi-square distribution table.

When there are more than two groups under testing, the procedure is similar, except that a variance-covariance matrix is needed to account for the correlation between groups. The log-rank statistic has a chi-square distribution with its degrees of freedom equaling the number of groups minus 1. The chi-square distributions of different degrees of freedom are also tabulated. When making comparisons between multiple groups, a large p value indicates no differences between any two of the groups, and a small p value indicates that at least two groups under testing are different. Note that a small p value does not mean that all the groups are different from each other.

Statisticians use the term *random variable* to describe an unknown quantity that has a chance to be each of many values. For example, the survival time of each of the patients in a study is a random variable since it is unknown at the beginning of the study. The underlying theory for the log-rank test is the central limit theorem. The basic version of this theorem states that the mean of many independent and identically distributed random variables has a normal distribution, regardless of the distribution of those random variables. In the log-rank statistic, those terms of the differences between observed and expected numbers of failures are also random variables. The log-rank statistic is in the form of a weighted average over these random variables, which are from the same distribution family, but not independent and identically distributed. The central limit theorem can be applied to the log-rank statistic after some manipulations via mathematical techniques and theories. The square of a random variable with a standard normal distribution has a chi-square distribution (with 1 degree of freedom). The sum of the squares of k independent random variables with a standard normal distribution is a chi-square distribution with k degrees of freedom, where k is a positive integer. That is how the distributions of log-rank tests are derived.

Software Programs

Most statistical software programs can be used to conduct log-rank tests. Examples include SAS, S-PLUS, R, BMDP, SPSS, and Stata.

Xuelin Huang

See also Cox Proportional Hazards Regression; Sample Size and Power; Statistical Testing: Overview; Survival Analysis

Further Readings

Fleming, T. R., & Harrington, D. P. (1991). *Counting processes and survival analysis*. Hoboken, NJ: Wiley.

Kalbfleisch, J. D., & Prentice, R. L. (2002). *The statistical analysis of failure time data* (2nd ed.). Hoboken, NJ: Wiley.

Klein, J. P., & Moeschberger, M. L. (1997). *Survival analysis: Techniques for censored and truncated data*. New York: Springer.

Mantel, N. (1966). Evaluation of survival data and two new rank order statistics arising in its consideration. *Cancer Chemotherapy Reports*, *50*(3), 163–170.

Peto, R., & Peto, J. (1972). Asymptotically efficient rank invariant test procedures. *Journal of the Royal Statistical Society, A*, *135*(2), 185–207.

Loss Aversion

See Risk Aversion

Lottery

In today's general parlance, the term *lottery* is used to refer to a type of gambling where there is a small (even infinitesimally small) chance for a gain and a large (even exponentially large) chance for an individual losing. In economic parlance, a lottery is a situation that involves an uncertain payoff. Consideration of lotteries by economists has generated many questions related to decision-making behavior in humans. This entry discusses the relationship of lotteries and auctions to human decision making, particularly medical decision making.

Decision-Making Behavior

The purchase of lottery tickets has been used in decision theory to illustrate the example of individuals with risk-seeking tendencies or attitudes in economic models of decision making. This is because a risk-seeking individual may be willing to purchase a lottery ticket even though the cost of that ticket is much more than the expected value of winning the lottery on the basis of that ticket.

In 1948, Milton Friedman and L. J. Savage asked a key question about lotteries: Why do people buy both lottery tickets and insurance against losses? That would seem to make them both risk seeking (lottery) and risk-averse (insurance against losses) at the same time. The proffered answer for Friedman and Savage was that part of the individual's utility function is concave and part is convex. Over one part of the function's range, some humans wish to play it safe, but over another part of the range of the function, these same humans are willing to take gambles. A simple model of decision making can be constructed using simple lotteries and giving the decision maker the flexibility to choose among two actions: play or not.

Yet, the term *lottery* is also used in a much more general sense in the history of economic thought in expected utility. In this more general sense in expected utility theory, risky alternatives are modeled in terms of "prospects." Here, the term *prospect* has been used interchangeably with *lottery*. And here, one could have the phrase *the attempt to model risky alternatives as "prospects" or "lotteries."*

John Nash, the Nobel Prize–winning game theorist, developed the notion of the Nash equilibrium for strategic noncooperative games in a setting that is often described in the economic decision-making literature as "choices over lotteries."

Auctions

An auction is a sale of an item based on bids. Auctions may be low-bid auctions (where an individual—the auctioneer—asks for a first bid, which may be the predetermined minimum price acceptable to the individual owner of the item put up for sale). Bidding in a low-bid auction starts low, and as the auctioneer raises the size of the bid and the auctioneer's suggested higher price is matched by bidders, the process continues until a highest bid is achieved and a winner identified. A high-bid auction starts off with the auctioneer starting the auction with a high asking price, which is then lowered until some bidder is willing to accept the auctioneer's price (or a predetermined minimum price is reached).

Auctions are not lotteries. Auctions and lotteries differ in their consequences and the relationships between the individual or group that places the item as the prize for the auction or the lottery and the bidders for that prize. In many circumstances, the individual or group putting up the prize for an auction will tend to receive more money for their prize in an auction than in a lottery. In addition, auctions can be "less fair" than lotteries because individuals or groups with the deepest pockets (the most wealth) can take over the bidding in any auction until other restrictions of the auctions are put into place that may increase the fairness of the auction. In addition, in any auction, it must be recognized that—unless restrictions are put into place—the more-moneyed bidders can also engage in behaviors with other bidders at the auction to shape the outcome of the auction (collusion).

Open Auctions

Most people are familiar with open auctions where all bidders are present in the audience and when a bid is placed an individual can look over and identify the bidder who has just made the previous bid before the placement of his or her

next bid. This type of auction allows collusion of sorts by allowing individuals to meet before the auction and plan how to bid on certain items for their own advantage over other bidders present in the same open room. An individual may bid for himself or herself, or the individual may represent another individual (or group) who is not in the room at the time of the bidding but may be in communication with the representative telephonically or electronically. Although attempts to minimize or restrict collusion may be in place, collusion is not ruled out in an open auction. One way to minimize collusion once an auction has started is to do away with the open auction entirely and to hold a closed auction.

Closed Auctions

A closed auction is an auction where no bidder is able to see another bidder's bid on a prize at any point in time from the start of the auction to the end of the auction. Given the fact that no bidder sees another bidder's bid, the bidding must take place "sequentially," where all bids are made and turned in so that no bidder sees what another bidder has bid. What is disclosed to all bidders through an announcement before the start of the next round of bidding is the highest bid that has been made up to that point in time. This is called *closed auction sequential bidding*, and the sequence continues through round after round of bidding, opening of bids, announcement of the highest bid after each round of bidding, and then continuing on through another round of bidding until the highest bid is reached. To carry out closed auction sequential bidding, all bidders need to be enclosed in their own bidding rooms with no opportunity for collaboration, discussion, or collusion once the auction starts until the auction ends. The only information each bidder receives is the highest bid at the end of each round of bidding before going into the next round of bidding. While auctions today are seen in governmental selling of licenses, for example, cell phone licenses, auctions are also found in health and medical care arenas.

Patient Care Auctions and Patient Debt Auctions

Hospital medicine auctions are seen today in two areas: patient care and patient debt. In a patient care auction, patient care provision is auctioned off to the highest bidder, for example, in auctions for the pricing of maternity care in the contracting market between insurance companies as buyers and hospitals as sellers. In a patient debt auction, a hospital turns to online auctions to sell their growing patient debt, hoping to receive bids by debt buyers and collection agencies.

Both auctions can affect market performance (e.g., pricing of maternity care in a locality, region, or country), market structure (e.g., by influencing more hospitals to become involved in the bidding processes), and market behavior (e.g., bidding behaviors, user preferences, and future impacts of successful bidders on those hospitals offering up the items—patient care or patient debt—for auction sale). For example, with the use of patient care auctions, if optimally satisfactory delivery and maternity care is not provided by the highest bidder, the insurance companies may have a backlash from disgruntled mothers, families, and employers. With the use of patient debt auctions, the hospitals in a specific geographic area—after the sale of their patient debt—may be affected by the debt-collection strategies of the debt buyers and collection agencies who are the most successful bidders for hospital-patient debt and who are now doing the collecting.

Medical Decision Making

Shifting from economics to medicine and medical decision making, one might suggest that a "lottery," defined as a gambling event with a small to infinitesimally small chance of winning a large prize, would have to be considered a "nonstandard decision" in medical decision making. Indeed, in terms of bounds of reasonableness in medical decision making, Jerome P. Kassirer and Stephen G. Pauker discuss the *toss-up*, where there is about a 50:50 chance of a gain or a loss in medical decision making. Kassirer and Pauker describe the toss-up decision as one where the physician can give over all decision making to the patient. Kassirer and Pauker do allow for patients who strongly consider what their own preferences are in terms of such toss-up medical decisions that they face in their own lives.

Consider the following example. Idiopathic pulmonary fibrosis (IPF) of the lung is a disorder of unknown cause associated with a high mortality rate. Up to the present, two questions remain unclear. First, it is unclear medically *what* treatment

modality should be applied in the attempt to ameliorate the disease process associated with this condition. Second, it is unclear *when* in the course of the disease the best chance of reversal (if any) exists and hence when treatment should be attempted. This is a disease entity where there has been little breakthrough in research and development, and many key research questions remain to be understood related to all aspects of IPF.

Now consider a case of end-stage pulmonary fibrosis of any type. The reality is that if a treatment was researched and developed that offered even a less-than-50% chance of a benefit accruing to individuals, this treatment would be heralded as a tremendous gain in the management of this end-stage condition, which otherwise will uniformly end in death in the short term or medium term rather than the long term. For example, a newly developed drug therapy for end-stage pulmonary fibrosis that gives a 40% chance of better survival over the next 1 to 5 years would be heralded as a significant advancement in treatment of this condition.

The above examples provide an opportunity to better understand the basis of toss-up decision making and the basis of medical decision making in general in the case of a severe disease with quality-of-life and survival consequences. First, there must be a severe disease, where a severe disease is characterized in terms of its consequences (related to the individual's quality of life and/or survival) in the short, medium, and long term. Second, the individual whose life is affected by the severe disease must have the capacity to make decisions on his or her own behalf and must be willing to accept the consequences of risky decisions in medicine. Third, the treatment itself must be available or accessible. Fourth, there must be a competent physician (and medical care team) in place skilled with the use of the treatment, able and willing to treat the patient, and able to help manage the consequences of the treatment in the patient and support the patient whether or not the treatment is successful.

Risky decisions in medicine include risky decision events which can have as consequences expected results that are as poor as those found in high-risk lotteries—with small chances of securing a medical gain (benefit) and grave chances of occurrence of adverse outcomes (risks)—that are entered into by humans. Here, such risky decision events should only be entered into by competent adult patients who are supported by highly qualified medical care teams in medical care settings where clear communication among all parties is embarked on and achieved. But the decision events should be entered into by these individuals only after careful prospective pre-event deliberations and careful retrospective postevent evaluations.

Dennis J. Mazur

See also Expected Utility Theory; Prospect Theory

Further Readings

Ashton, J. (1893). *A history of English lotteries*. London: Field & Tuer/Leadenhall Press.

Bakir, N. O. (2004, August). *Evaluation of information bundles in engineering decisions*. Unpublished PhD dissertation, Texas A&M University.

Bradley, R. E. (2001). *Euler and the Genoese lottery*. Retrieved June 5, 2008, from http://oldhome.adelphi.edu/~bradley/Euler/neworleans.pdf

Bu, Tian-Ming, Deng, Xiaotie, & Qi, Qi. (2008). Forward-looking Nash equilibrium for keyword auction. *Information Processing Letters*, *105,* 41–46.

Friedman, M., & Savage, L. J. (1948). The utility analysis of choices involving risk. *Journal of Political Economy, 56,* 279–304.

Howard, R. A. (1967). Value of information lotteries. *IEEE Transactions on Systems Science and Cybernetics*, *3,* 54–60.

Kassirer, J. P., & Pauker, S. G. (1981). The toss-up. *New England Journal of Medicine, 305,* 1467–1469.

Klarreich, E., Arrow, K., Aumann, R., McMillan, J., Milgrom, P., Myerson, R., et al. (n.d.). *The bidding game*. Written for Beyond Discovery®: The Path from Research to Human Benefit, a project of the National Academy of Sciences. Retrieved June 5, 2008, from http://www.beyonddiscovery.org/content/view.txt.asp?a=3681

Lotteries. Their origin and history. Noted lotteries of England, France, and America. Calculation of chances and choice of numbers. The ethics of lotteries and the economy of their use. (1875, November 7). *The New York Times*, p. 2.

Raven, J. (1991). The abolition of the English state lotteries. *The Historical Journal, 34,* 371–389.

Richards, R. D. (1953). The lottery in the history of English government finance. *Economic History, 3,* 57–76.

Seville, A. (1999). The Italian roots of the lottery. *History Today, 49,* 17–23.

M

Managing Variability and Uncertainty

Clinical decision making involves the use of diverse strategies to generate and test potential solutions for problems that are presented by patients. It involves using, acquiring, and interpreting the indicators and then generating and evaluating hypotheses. Most health decisions occur in contexts of scientific uncertainty. Thus, the notion of uncertainty should be at the heart of exchanges between professionals and their patients. Accordingly, the failure to integrate the concept of uncertainty in routine medical practice remains a major obstacle to informed decisions by patients.

The first section of this entry reviews the definition of the concepts of variability and uncertainty and makes the distinction with other similar concepts. The second section briefly summarizes how physicians understand and react to uncertainty. It also proposes potential strategies to alleviate the burden of managing uncertainty in routine clinical decision making. The last section highlights the gaps in knowledge and areas for further research.

Definitions

If *ignorance* is defined as an absence of knowledge of the available issues or options as well as of their probability, *uncertainty* is defined as knowledge of the issues or options available but with an absence of the knowledge of their probability. Risk is defined as knowledge of the available issues or options and their probabilities. In some literature, the concept of variability is distinguished from that of uncertainty and refers to the heterogeneity of subjects included in analyses. Although the distinction between variability and uncertainty has clear implications from a decision analysis perspective, for clinicians and their patients, variability is only one of the many sources of uncertainty.

Types of Uncertainty

Diagnosis and management of health problems are full of uncertainty. Sometimes, the probabilistic nature of the diagnosis made by physicians makes it difficult to choose the "best" course of action. Scientific evidence that imparts conflicting results regarding treatment options (i.e., balance between risks and benefits) or the absence or insufficiency of scientific evidence makes this choice even more difficult. Moreover, the probabilistic aspect of the evidence that is drawn from populations implies uncertain outcomes for the individual. Consequently, patients and physicians need help in addressing their decisional needs and in resolving uncertainty when making decisions.

Physicians' Reaction to Uncertainty

Both patients and their physicians have difficulty grasping the concept of uncertainty, specifically when dealing with numbers and probabilities. Physicians express concerns with communicating risk to patients and may not have the necessary

skills to do it. In brief, the medical problem and the characteristics of the patient create the uncertainty inherent in the clinical encounter. The characteristics of physicians influence their reaction to uncertainty. In turn, the decision-making process occurring during the clinical encounter between a patient and a physician is under the influence of the uncertainty inherent in the clinical encounter and of the physician's reactions to uncertainty. Patients and physicians interact to produce a set of decisions that in some cases will be translated into physicians' behavior. The decision outcome and, on some occasions, the physician's behavior may be modified by external sources such as source of payment, setting of the practice, and so on. The reaction of physicians to uncertainty is composed of four main constructs: anxiety due to uncertainty, concern about bad outcomes, reluctance to disclose uncertainty to patients, and reluctance to disclose mistakes to other physicians. The reaction of physicians toward uncertainty was shown to be significantly associated with disclosure of uncertainty by physicians to patients during clinical encounters, resource use and costs, and the intention of physicians to engage in shared decision making. Therefore, strategies to alleviate the burden of managing uncertainty in clinical decision making are of utmost importance in ensuring quality of care, patient safety, and control of costs for the healthcare system.

Strategies

For both health providers and patients to accept and manage uncertainty, there is a need for, first, recognizing and accessing the level of uncertainty that is present in the decision-making process. It is in this context that there is considerable interest today in the idea of shared decision making.

Shared Decision Making

Shared decision making is defined as a joint process shared by the physician and the patient. It rests on the best evidence as to the risks and benefits of all available options, including doing nothing. It includes the following components: establishing a context in which patients' views about treatment options are valued and seen as necessary; transferring technical information; making sure patients understand this information; helping patients base their preference on the best evidence; eliciting patients' preferences; sharing treatment recommendations; and making explicit the component of uncertainty in the clinical decision-making process. Consequently, fostering shared decision making in clinical settings has the potential to help both health providers and patients recognize the uncertainty that is present in the decision-making process, a first step for managing uncertainty in routine clinical decisions.

Screening for Decisional Conflict in Patients

Decisional conflict is defined as a state of uncertainty as perceived by an individual about which course of action to take when the choice among competing actions involves risk, loss, regret, or a challenge to personal life values. Decisional conflict is multidimensional and influenced by the perception of being informed, the perception of being clear about personal values, opinions of significant others, and effectiveness of the decision outcomes. It can be used to assess decisional needs, to tailor interventions to these needs, and to evaluate their effects. It can be measured from a patient's perspective and from a health provider's perspective. Decisional conflict should not be confused with a direct measure of how much scientific uncertainty is involved in the decision to be made itself (e.g., conflicting results, absence of evidence). In routine clinical contexts, uncertainty refers to decisional uncertainty (one's own perception of not knowing which course of action to take when choosing among actions that involve risk, loss, regret, or challenge to personal life values) and outcome uncertainty (one's own perception of not knowing who is going to benefit or be harmed by the treatment that could be chosen). In lay terms, it is understood as the level of comfort that an individual faces when making a decision, that is, decisional comfort.

Decisional Conflict Scale

The Decisional Conflict Scale (DCS) consists of 16 items, grouped into five subscales (certainty, information, values clarification, support or pressure from others, and perception of the quality of the decision process). DCS scores correlate with knowledge assessment scores, intentions to accept influenza vaccine or breast cancer screening, delay with the decision to be immunized, decisional

regret, and intention to sue a physician. It is one of the very few existing measurements of the decision-making process applicable with both health providers and patients. The combination of the DCS score of the health provider with the DCS score of the patient can be used for assessing the agreement reached between both members of the dyad on decisions made during clinical encounters. This combination thus allows for the development of interventions designed to improve the clinical decision-making process between patients and their physicians in contexts of uncertainty. Consequently, screening of patients with the DCS to identify the existence and nature of their perceived uncertainty about the course of action to take is one of the key competencies of informed shared decision making. This helps to (a) acknowledge explicitly the presence of uncertainty in the decision-making process; (b) identify the areas that need to be addressed to provide decision support; and (c) engage in informed and shared decision making.

Clinical Algorithms and Practice Guidelines

Given their systematic approach to evidence, clinical practice guidelines are defined as systematically developed statements to assist practitioners and patients with decisions about appropriate healthcare for specific circumstances. Clinical practice guidelines have been the subject of many research initiatives. More specifically, a large number of studies have aimed at improving adherence of clinicians as well as patients to the recommendations of clinical practice guidelines but with very little success.

A large percentage of medical decisions largely occur in contexts of scientific uncertainty. These "grey-zone" (or preference-sensitive) decisions are characterized either by scientific evidence that points to a balance between harms and benefits within or between options, or by the absence or insufficiency of scientific evidence. Moreover, probabilities of risks and benefits in a population cannot be directly attributed at the individual level. Current clinical practice guidelines, however, are not adapted to grey-zone decisions and thus cannot help providers and their patients make informed and shared decisions. Clinical practice guidelines are still largely conceived as tools that should foster adherence to a best decision defined by the "expert health professional," rather than instruments that should support the best decision for a specific patient in a specific context. Thus it is no surprise that health professionals have criticized clinical practice guidelines for lacking relevant information to assist shared decision making with patients.

Patient Decision Aids

Patient decision aids are tools designed to help patients participate in clinical decision making. They provide information on the options and help patients clarify and communicate the personal values they associate with features of the different options. When compared with usual care or simple information leaflets, patient decision aids improve decision quality and the measures in the decision process of feeling informed and clear about values that are included in the overall measure of the DCS. Patient decision aids foster shared decision-making processes between health providers and their patients because they make explicit the uncertainty component of the decision to be made. Last, patient decision aids are known for reducing overuse of controversial medical procedures such as prostate-cancer-screening tests and improving underuse of beneficial public health measures such as childhood vaccination. Therefore, the use of patient decision aids in clinical practice is an effective way of managing uncertainty in routine clinical decision making.

Knowledge Gaps

Although many current initiatives focus on training health providers in improving their understanding of probability and thus of the uncertainty that is inherent in clinical decision making, it is not clear that this will be sufficient to affect favorably the decision quality of their patients. Therefore, patients will need to be able to access high-quality patient decision aids and be guided by competent individuals in risk communication and shared decision making. However, based on a systematic review of barriers to and facilitators of implementing shared decision making in clinical practice, time constraints remain the most often cited barrier across many different cultural and organizational contexts. Indeed, there is a general consensus that the growing demands and expectations of informed health

consumers and societies are putting a lot of pressure on limited resources, including human resources. Therefore, it remains essential that future studies investigate whether using patient decision aids in routine medical practice, engaging in shared decision making, and discussing the uncertainty that is inherent in clinical decision making actually take more time or not than usual care. This will require health service researchers and policy makers to be innovative and creative in elaborating the needed decision support intervention tools and thus "decision support care pathways" that can streamline the process of informed and shared decision making in overburdened healthcare clinical settings.

France Légaré

See also Clinical Algorithms and Practice Guidelines; Decisional Conflict; Evidence-Based Medicine; Patient Decision Aids; Risk Communication; Shared Decision Making

Further Readings

Boivin, A., Légaré, F., & Gagnon, M. P. (2008). Competing norms: Canadian rural family physicians' perceptions of clinical practice guidelines and shared decision-making. *Journal of Health Services Research & Policy, 13*(2), 79–84.

Edwards, A., Matthews, E., Pill, R., & Bloor, M. (1998). Communication about risk: Diversity among primary care professionals. *Family Practice, 15*(4), 296–300.

Folmer Andersen, T., & Mooney, G. (1990). *The challenges of medical practice variations*. Houndmills, UK: Macmillan.

Gerrity, M. S., White, K. P., DeVellis, R. F., & Dittus, R. S. (1995). Physicians' reactions to uncertainty: Refining the constructs and scales. *Motivation and Emotion, 19*(3), 175–191.

Gigerenzer, G. (2002). *How to know when numbers deceive you: Calculated risks*. New York: Simon & Schuster.

Hunink, M., Glasziou, P., Siegel, J., Weeks, J., Pliskin, J., Elstein, A. S., et al. (2004). Variablity and uncertainty. In M. Hunink, P. Glasziou, J. Siegel, J. Weeks, J. Pliskin, A. S. Elstein, et al. (Eds.), *Decision making in health and medicine: Integrating evidence and values* (3rd ed., pp. 339–364). Cambridge, UK: Cambridge University Press.

O'Connor, A. M. (1995). Validation of a decisional conflict scale. *Medical Decision Making, 15*(1), 25–30.

O'Connor, A. M., Bennett, C., Stacey, D., Barry, M. J., Col, N. F., Eden, K. B., et al. (2007). Do patient decision aids meet effectiveness criteria of the international patient decision aid standards collaboration? A systematic review and meta-analysis. *Medical Decision Making, 27*(4), 554–574.

Towle, A., & Godolphin, W. (2001). Education and training of health care professionals. In A. Edwards & G. Elwyn (Eds.), *Evidence-based patient choice inevitable or impossible?* (pp. 245–270). Oxford, UK: Oxford University Press.

Marginal or Incremental Analysis, Cost-Effectiveness Ratio

A marginal or incremental analysis focuses on the additional costs and additional outcomes associated with a change of some kind. For a marginal analysis, this change concerns a slight (marginal) increase or decrease in service. For an incremental analysis, this change generally concerns the introduction of a new intervention. The aim of both types of analysis is to establish the impact of the change on costs and outcomes relative to the situation prior to the change. This entry introduces the concepts of marginal or incremental analysis and reviews the importance of adopting such an approach to measuring the impact of changes for economic evaluation.

Marginal or Incremental Analysis

A marginal analysis is concerned with the additional costs and additional outcomes achieved from a marginal (unitary) change in service. For example, a marginal analysis of a mammography screening facility would consider the additional costs and outcomes associated with a one-unit increase in the number of mammographies undertaken within the department. It is important to note that the additional costs and outcomes associated with a one-unit change are not necessarily equal to the average cost and outcomes associated with the group. This is due to the existence of fixed costs and outcomes that are not affected by the number of units. As a result, marginal values are used in economics to determine the actual impact associated with changes in service.

In contrast, an incremental analysis is concerned with the additional costs and additional outcomes

associated with the introduction of a *new* service. For example, an incremental analysis would consider the additional costs and additional outcomes of a new approach to mammography screening (e.g., 2 view vs. 1 view).

Economic evaluation, irrespective of whether it is cost-effectiveness, cost-utility, or cost-benefit analysis, is interested in the impact of a change compared with the position before the change. As such, economic evaluation is concerned with marginal or incremental analyses, and, where the outcomes measured involve a single nonmonetary unit (e.g., quality-adjusted life years or life years), the results are presented as an incremental cost-effectiveness ratio.

Incremental Cost-Effectiveness Ratio

The incremental cost-effectiveness ratio (ICER) gives a measure of the additional cost per unit of health gain. It is estimated by comparing the additional costs and outcomes associated with the new service (or intervention) with those of the original service(s):

$$\text{ICER} = \frac{\text{Cost}_{\text{new intervention}} - \text{Cost}_{\text{current treatment}}}{\text{Outcome}_{\text{new intervention}} - \text{Outcome}_{\text{current treatment}}}.$$

When determining ICERs for a set of interventions, the interventions should be ranked in ascending order of outcome (or cost) and a ratio calculated for each intervention relative to the next best (more costly) viable intervention by dividing the additional cost by the additional outcome produced.

Incremental Analysis Versus Average Analysis

As with the mammography example above, a new intervention or service (2 view) is rarely the only option available (1 view). Even when the alternative is to "do nothing," this is rarely associated with zero costs and zero outcomes. For example, no active screening would involve the costs and outcomes of breast cancers found clinically. As such, an assessment of the impact of the addition of 2 view mammography should involve the additional costs and outcomes associated with 2 view compared with the costs and outcomes without 2 view (e.g., 1 view and no active screening).

Example

Consider a (hypothetical) situation where there are four methods available for mammography screening (A to D). These are characterized by the costs and outcomes given in the table below.

The average cost-effectiveness ratio suggests that all the screening methods provide outcomes at a price that is likely to be considered reasonable; even for Method D (which costs in excess of $250,000), the average cost-effectiveness is only $23,041 per unit of effect. However, as noted above, this average measure is misleading as it ignores the existence of the other methods of screening and the outcomes that can be generated from them for a lower cost. The appropriate way to measure the cost-effectiveness of Method D is to compare the additional (incremental) costs associated with the method and the additional outcomes generated by the method compared with the next most effective method. This analysis reveals that choosing Method D rather than Method C increases the costs by $146,495 but only increases the outcomes by 0.8—giving D a large incremental cost-effectiveness ratio of $183,119. In comparison with the relevant alternative, D no longer looks like good value for the money.

Elisabeth Fenwick

	Outcomes	*Costs ($)*	*Average Cost-Effectiveness ($)*	*Inc. Effect*	*Inc. Cost ($)*	*ICER ($)*
A	6.2	76,410	12,324	—	—	—
B	8.5	87,659	10,313	2.30	11,249	4,891
C	10.4	111,562	10,727	1.90	23,903	12,581
D	11.2	258,057	23,041	0.8	146,495	183,119

See also Cost-Effectiveness Analysis; Cost-Utility Analysis; Decision Rules

Further Readings

Drummond, M. F., O'Brien, B. J., Stoddart, G. L., & Torrance, G. W. (1997). *Methods for the economic evaluation of health care programmes* (2nd ed.). New York: Oxford University Press.

Torgenson, D. (1996). Authors should have used marginal analysis. *British Medical Journal, 312*, 909.

Torgenson, D., Donaldson, C., & Reid, D. (1996). Using economics to prioritize research: A case study of randomized trials for the prevention of hip fractures due to osteoporosis. *Journal of Health Services Research & Policy, 1*, 141–146.

Markov Models

A decision analysis problem is attacked using a formal process that begins with constructing a mathematical model. For more than 40 years, the decision tree has been the most common formalism, comprising choices, chances, and outcomes. The modeler arranges near-term events in a tree structure, and attempts to balance realism and attendant complexity with simplicity. In problems that lead to long-term differences in outcome, the decision model must have a definite time horizon, up to which the events are characterized explicitly. At the horizon, the future health of the patient or cohort must be summed and averaged into "subsequent prognosis." For problems involving quantity and quality of life, where the future natural history is well characterized, techniques such as the declining exponential approximation of life expectancy or calculus model may be used to generate outcome measures. Life tables may be used directly, or the results from clinical trials may be adopted to generate relevant values.

Some decision problems are less amenable to these summarizing techniques. In particular, clinical scenarios that involve a risk that is ongoing, or competing risks that occur at different rates, lead to either rapidly branching decision trees or unrealistic pruning of possible outcomes for the sake of simplicity. In these cases, a probabilistic model of natural history can substitute for the outcome node of the decision tree. Beck and Pauker introduced the Markov process as a solution for the natural history modeling problem in 1983, building on their work and others' work with stochastic models over the previous 6 years. During the ensuing 25 years, more than 1,000 articles have directly cited either this paper or a tutorial published a decade later, and more than 1,700 records in PubMed can be retrieved using "(Markov decision model) OR (Markov cost-effectiveness)" as a search criterion. This entry defines the Markov process model by its properties and discusses two aspects of Markov modeling: transition probabilities and regular versus absorbing models.

The Markov Process and Transition Probabilities

A Markov process is a special type of stochastic model. A stochastic process is a mathematical system that evolves over time with some element of uncertainty. This contrasts with a deterministic system, in which the model and its parameters specify the outcomes completely. The simplest example of a stochastic process is coin flipping. If a fair coin is flipped a number of times, a sequence of results such as TTHHTHTHHHHTTHTHHTT might arise. At each flip (or trial), either tails (T) or heads (H) would result with an equal probability of 1/2. Dice rolling is another example of this type of stochastic system, known as an independent trial experiment. Each flip or roll is independent of all that have come before because dice and coins have no memory of prior results.

The Markov process relaxes this assumption a bit. In a Markov model, the probability of a trial outcome varies depending on the current result (generally known as a "state"). It is easy to see how this model works via a simple example. Consider a medical scheduling clerk who assigns new patients to three doctors: Adams, Baker, and Chou. The clerk randomly assigns patients to these practitioners but has a few idiosyncrasies.

Never is Adams assigned two patients in a row, but after assigning a patient to Adams, the clerk randomly gives the next person to either Baker or Chou by flipping a coin. After assigning to Baker, the clerk randomly gives the next patient to any of the three doctors with probability 1/3. After assigning to Chou, the clerk assigns the next patient to Adams with probability 1/2 and to Baker or Chou with probability 1/4. Thus, the last assignment (Adams, Baker, or Chou) must be known to determine the probability of the next assignment.

Table 1 shows this behavior as a matrix of *transition probabilities*. Each cell of Table 1 shows the probability of a patient being assigned to the doctor named at the head of the column if the last patient was assigned to the doctor named at the head of the row. An $n \times n$ matrix is a probability matrix if each row element is nonnegative and each row sums to 1. Since the row headings and column headings refer to states of the process, Table 1 is a special form of probability matrix: a transition probability matrix.

Table 1 Doctor assignment probability table

	Next		
Prior	*Adams*	*Baker*	*Chou*
Adams	.000	.500	.500
Baker	.333	.333	.333
Chou	.500	.250	.250

Table 2 Two-step Markov probabilities

	Two Later		
Prior	*Adams*	*Baker*	*Chou*
Adams	.417	.292	.292
Baker	.278	.361	.361
Chou	.208	.396	.396

This stochastic model differs from independent trials because of the *Markov property*: The distribution of the probability of future states of a stochastic process depends on the current state (and only on the current state, not the prior natural history). That is, one does not need to know what has happened with scheduling in the past, only who was most recently assigned a patient. For example, if Baker got the last patient, the next one will be assigned to any of the three physicians with equal probability.

The Markov property leads to some interesting results. What is the likelihood, if Adams is assigned a patient, that Adams will get the patient after next? This can be calculated as follows:

After Adams, the probability of
Baker is 1/2 and Chou 1/2.

After Baker, the probability of Adams is 1/3,
and after Chou it is 1/2.

So the probability of Adams-(anyone)-Adams is $(1/2) \times (1/3) + (1/2) \times (1/2)$ or .417. A complete table of probabilities at two assignments after a known one is shown in Table 2. This table is obtained using matrix multiplication, treating Table 1 as a 3×3 matrix and multiplying it by itself. Note that the probability of Adams going to Adams in two steps is found in the corresponding cell of Table 2.

This process can be continued because Table 2 is also a probability matrix, in that the rows all sum to 1. In fact, after two more multiplications by Table 1, the table is as shown in Table 3.

The probabilities in each row are converging, and by the 10th cycle after a known assignment, the probability matrix is as shown in Table 4.

This is also a probability matrix, and it has a straightforward interpretation. Ten or more cycles after a known assignment, the probability that the next assignment will be to Adams is .294, and to Baker or Chou .353. Or, if someone has no knowledge of what has been happening recently, the likelihood of the next patient going to Adams is .294, and so on. This is the limiting Markov matrix or the steady state of the process. This particular scheduler, despite the idiosyncratic behavior, gives around 30% of the patients to Adams over time and an equal number of the remainder to each of the other two doctors.

Table 3 Assignment model after four cycles

	Four Later		
Prior	*Adams*	*Baker*	*Chou*
Adams	.315	.342	.342
Baker	.291	.354	.354
Chou	.279	.360	.360

Table 4 Steady-state or limiting Markov matrix

	Cycle 10 and Later		
Prior	*Adams*	*Baker*	*Chou*
Adams	.294	.353	.353
Baker	.294	.353	.353
Chou	.294	.353	.353

Absorbing Markov Models

The patient scheduling example is known as a *regular* Markov chain. The probabilities are constant and depend only on the state of the process. Any state can be reached from any other state, although not necessarily in one step. Regular chains converge to a limiting set of probabilities. Another category of Markov models is *absorbing*. In these systems, the process has a state that is possible to enter, in a finite set of moves, from any other state but from which no movement is possible. Once the process enters the absorbing state, it terminates (or stays in that state forever). The analogy with clinical decision models is obvious: An absorbing Markov model has a "dead" state.

This is shown in Figure 1, a simplified three-state absorbing clinical Markov model. In a clinical model, the notion of time appears naturally. Assume that a clinical process is modeled that has clinic visits month to month. At any given month, the patient may be in a Well state, shown in the upper left of Figure 1, the Sick state in the upper right, or Dead in the lower center. If in the Well state, the most likely result for the patient is that he or she will remain well for the ensuing month and next be found still in the Well state. Alternatively, the patient could become Sick or die and move to the Dead state. If sick, the patient will most likely stay sick, but a small chance of returning to the clinically Well state is possible. The patient could also die from the Sick state, at a higher probability than from the Well state.

A possible transition probability matrix for this model is shown in Table 5.

Thus, a Well patient remains so with probability .9, has a 9% chance of becoming Sick in 1 month, and a 1% chance of dying in the cycle. A Sick patient has a 2% chance of returning to the Well state, a 7% chance of dying in 1 month, and the remainder (91%) of remaining in the Sick state. Of course, the Dead state is *absorbing*, reflected by a 100% chance of staying Dead.

Table 5 is a probability matrix, and so it can be multiplied as in the prior example. After two cycles, the matrix is as shown in Table 6.

After 2 months, someone who started in the Well state has slightly more than 81% chance of staying well and a 16% chance of being Sick. By the 12th cycle, or 1 year, the top row of the transition matrix is as follows:

.326 .375 .299

So someone starting well has a 37.5% chance of being in the Sick state after 12 months and a nearly 30% chance of being dead. This matrix converges

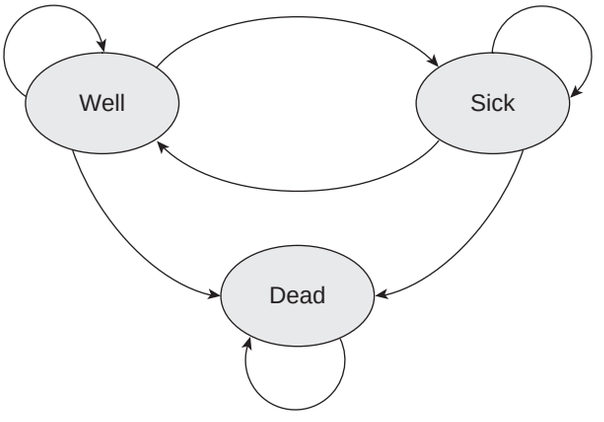


Figure 1 Three-state absorbing clinical Markov model

Table 5 Transition probability matrix for clinical example

	Next		
Prior	*Well*	*Sick*	*Dead*
Well	.9	.09	.01
Sick	.02	.91	.07
Dead	0	0	1

Table 6 Two-month state matrix

.812	.163	.025
.036	.830	.134
.000	.000	1.000

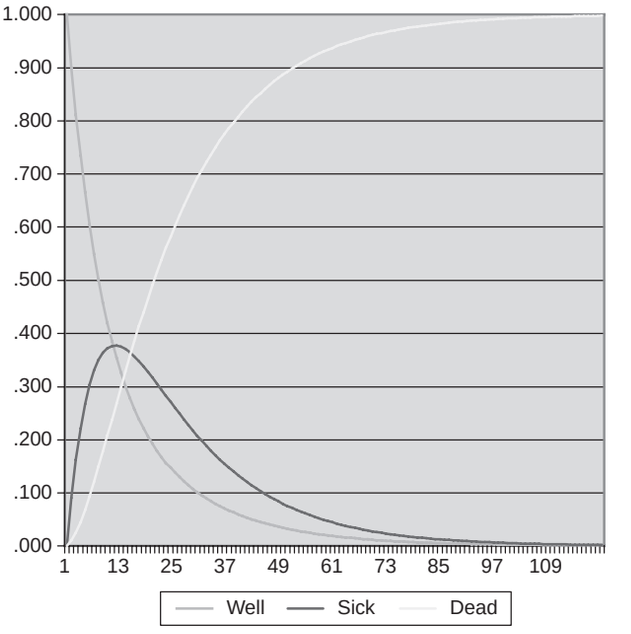


Figure 2 Likelihood of being in each state at specific cycles

slowly because of the small probability of death in any one cycle, but eventually this matrix will end up as a set of rows:

0 0 1

Everyone eventually dies. Death is the ultimate absorbing state.

Clinical Markov models offer interesting insights into the natural history of a process. If the top row of the transition matrix is taken at each cycle and graphed, Figure 2 results. This graph can be interpreted as the fate of a cohort of patients beginning together at Well. The membership of the Well state decreases rapidly as the forward transitions to Sick and Dead overwhelm the back transition from Sick to Well. The Sick state grows at first, as it collects patients transitioning from Well, but soon the transitions to Dead, which of course are permanent, cause the Sick state to lose members. The Sick state peaks at Month 12, with 37.5% of the cohort. The Dead state is actually a sigmoid (S-shaped) curve, rising moderately for a few cycles because most people are Well, but as soon as the 7% mortality from the Sick state kicks in, the curve gets steeper. Finally it flattens as few people remain alive. This graph is typical of absorbing Markov process models.

Use of Absorbing Markov Models in Clinical Decision Analysis

Other entries in this encyclopedia focus on the applications of this model; here, the principal uses of absorbing Markov processes are noted. Theoretically, the Markov formalism can substitute for an outcome in a typical decision tree. Whereas a traditional outcome node is assigned a value, or *utility*, a Markov model is used to calculate the value. For this to work, each Markov state is assigned an incremental utility for being in that state for one model cycle. In the example above, the Well state might be given a value of 1, the Sick state a value of .8. That is, the utility for being in the Sick state is 80% of the value of the Well state, for each cycle in it. In most models, Dead is worth 0. Incremental costs can also be applied for Markov cost-effectiveness analysis.

Two enhancements to the Markov model render the formalism more realistic for clinical studies; both involve adding a time element. First, although the Markov property requires no memory of prior states, it is possible to superimpose a time function on a transition probability. The most obvious example of this is the risk of death, which rises over time regardless of other clinical conditions. This can be handled in a Markov model by modifying the transition probability to death using a function: $p(\text{Well} \rightarrow \text{Dead} - 0.01 + G(\text{AGE})$, where G represents the Gompertz mortality function or another well-characterized actuarial model.

Second, standard practice in decision modeling discounts future costs and benefits to incorporate risk aversion and the decreasing value of assets and events in the future. Discounting may be incorporated in Markov models as simply another function that can modify the state-dependent incremental utilities.

A number of refinements to the basic Markov model have been developed over the past 25 years, and more modern techniques of operations research can handle realistic and complex clinical problems that violate the Markov property to some extent. Nevertheless, this stochastic formalism remains valuable for clinical studies and teaching medical decision making.

J. Robert Beck

See also Applied Decision Analysis; Decision Tree: Introduction; Markov Models, Applications to Medical Decision Making; Markov Models, Cycles; Markov Processes; Stochastic Medical Informatics

Further Readings

Beck, J. R., & Pauker, S. G. (1983). The Markov process in medical prognosis. *Medical Decision Making, 3,* 419–458.

Sonnenberg, F. A., & Beck, J. R. (1993). Markov models in medical decision making: A practical guide. *Medical Decision Making, 13,* 322–339.

Markov Models, Applications to Medical Decision Making

The basic purpose of a decision model is to estimate the prognosis of a patient or a population of patients subsequent to each alternative choice of management strategy being compared. For practical reasons, the analysis must be restricted to a time frame, often referred to as the *time horizon* of the analysis. The time horizon may be finite (e.g., 5 years from the time of decision) or may be indefinite or defined in operational terms such as "for the remainder of the patient's life" or "until all patients in the population are dead." The choice of time horizon is determined by many factors, including the time frame of events of interest, the availability of data, and the perspective of the analysis. When the time horizon is the remaining lifetime, then the model must represent the prognosis following each management strategy that incorporates all future events in the patients' lives.

There are various ways in which a decision analyst can assign values to these terminal nodes of the decision tree. In some cases, the outcome measure is a simple life expectancy. One method for estimating life expectancy is the declining exponential approximation of life expectancy (DEALE), which calculates a patient-specific mortality rate for a given combination of patient characteristics and comorbid diseases. Life expectancies may also be obtained from Gompertz models. In the reference case recommended by the Panel on Cost-Effectiveness Analysis, prognosis is modeled as quality-adjusted life expectancy, in which the prognosis incorporates both quantity and quality of life. For health economic analyses, economic costs must also be assigned to each strategy being compared.

Various modeling techniques can be used to estimate prognosis. This entry introduces the Markov model as an alternative to simple decision tree models and discusses the assumptions inherent in Markov models, how they are evaluated, and how they are used in decision models to determine prognosis.

Limitations of Simple Trees

A simple tree is one consisting only of decision, chance, and terminal nodes (Figure 1). In this example, the tree models an unspecified disease that can have the outcomes of Disabled or Well and Death for either Well or Disabled patients. The terminal nodes represent outcomes that must be assigned a utility, typically a quality-adjusted life-expectancy. Each pathway from the root of the tree to the terminal node represents a unique combination of events that can be factored into

the utility. This structure is acceptable if the events modeled in the tree occur within a short time span, such as over the treatment of a short-term disease (e.g., pneumonia) or surgery. Simple trees have the following limitations:

- They cannot easily specify when events occur or differentiate between earlier and later events.
- They cannot easily model continuous risk when the timing of events is uncertain.
- They cannot model situations for which events may occur more than once.

Most realistic clinical decision problems involve all these factors. Even when the clinical situation fits within these limitations, the analyst still has to assign utilities for terminal outcomes. In the case of Disabled or Well, these utilities must represent all subsequent prognosis, whether for a finite time frame or for the remainder of the patient's life.

Markov models avoid all the above limitations by modeling clinical conditions as discrete *health states* and modeling all events as transitions among states.

History of the Markov Model in Medical Decision Making

The Markov process is named after the Russian mathematician Andrey Markov (1856–1922). The Markov model was introduced into medical decision making by Beck and Pauker in their seminal 1983 paper in *Medical Decision Making* to overcome the limitations of simple decision models to represent risks over time. At the time of Beck and Pauker's original paper, the Markov model could be solved only by matrix algebra or by simulations using a spreadsheet or by writing a dedicated computer program. With any of these methods, models were difficult to construct and modify, and sensitivity analyses were laborious and time-consuming. In 1984, Hollenberg developed the Markov cycle tree, a formalism that represents the health states and events in a Markov model. The tree-based formalism made it far easier to construct and modify models and provided a convenient mechanism for automated evaluation and sensitivity analysis. With the more convenient representation, analysts were quick to adopt Markov models, and within a few years of its introduction to medical decision making, the Markov model became the method of choice for most medical decision problems.

Definitions

Markov Process

A Markov model is a decision model representing a *Markov process*. The Markov process is a *stochastic process* (subject to random variation in outcome) with the following characteristics:

- A finite set of health states (the *state space*) referred to as the *Markov states*
- Transitions between pairs of states with a defined probability for each transition
- The Markov ("no memory") property

Markov processes inherently represent the passage of time. Most medical decision problems

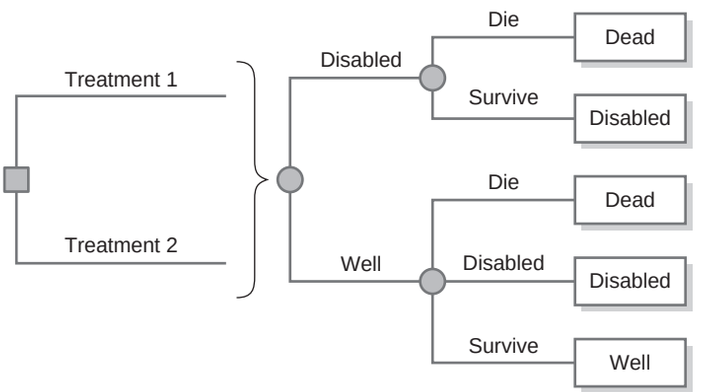


Figure 1 Simple tree

are represented as *discrete-time Markov processes,* which means that time is modeled in discrete, uniform steps. The remainder of this discussion applies only to discrete-time Markov processes

Markov Cyc les

Time is represented in discrete-time Markov processes as discrete, uniform intervals referred to as *cycles*. The *cycle length* is selected to be appropriate to the model. For models representing an entire life span, a cycle length of 1 year may be appropriate. When events occurring over shorter periods of time are important, a cycle length of 1 month or even 1 week may be more appropriate. The choice of a cycle length is a compromise between clinical realism and evaluation time. Short cycles result in a finer granularity of time in the model and greater precision but may result in simulations that take a long time to run.

Conceptually, Markov models represent all events as transitions from one state to another during Markov cycles.

States and Transitions

Markov models represent the universe of health outcomes as a finite set of mutually exclusive, collectively exhaustive *health states* (the state space of the Markov process). All events are represented as transitions from one state to another. A simple representation of the Markov model is a *directed graph* called a *state transition diagram* as shown in Figure 2. States are represented by circles, each labeled with the name of the state. Arrows indicate *allowed transitions*. If there is no arrow pointing from one state to another, a transition from the first state to the second is not allowed. For example, in Figure 2, a transition is allowed from WELL to DISABLED, or from WELL to DEAD, but not from DISABLED to WELL, from DEAD to WELL, or from DEAD to DISABLED. States can have transitions to themselves, as indicated by circular arrows in Figure 2, meaning that a patient can remain in the same state for consecutive cycles. *Temporary states* are states that can make transitions only to other states but not to themselves. *Absorbing states* are states that cannot make transitions to any other states. In Figures 2 and 3, DEAD is an absorbing state. Each transition is characterized by a probability, associated with the arrows (edges) of the state transition diagram. Sometimes state transition diagrams represent two consecutive cycles, as in Figure 3.

The *state* of a Markov process is defined by the following:

- The distribution of a cohort among the Markov states in the case of a cohort simulation or the state in which a single subject resides in the case of a Monte Carlo simulation
- The set of transition probabilities for all state transitions

The distribution can be represented as a probability vector ($1 \times n$ matrix, where n is the number of states) with an entry for the probability of membership of each state. The set of transition probabilities is conveniently represented as an $n \times n$ matrix referred to as the *transition probability matrix* (see Table 1). The entries in the rows of the probability matrix must sum to 1.

Markov Property: "No Memory" Assumption

Markov models are defined by the Markov property, which is that the future state of the model in any cycle is determined only by the current state and is not affected by the prior history of the Markov process. In other words, it matters

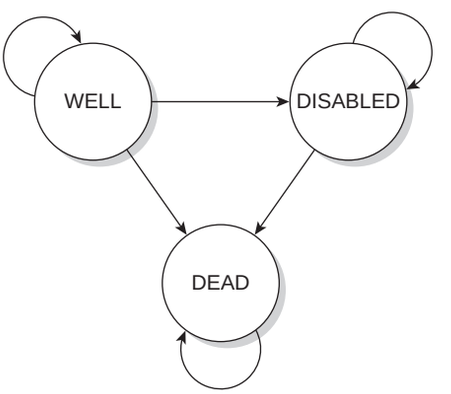


Figure 2 Markov state transition diagram

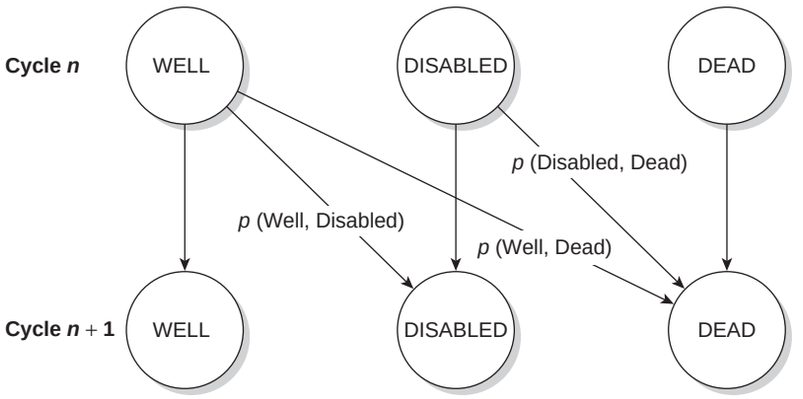


Figure 3 Markov state transition diagram with consecutive cycles

only what the current state is, not how it got there. The Markov property is referred to as the "no memory" property of Markov processes.

What the Markov property means in practical terms is that the transition probability matrix during any cycle of the Markov process does not depend on the history of the Markov process. Therefore, two Markov processes with identical states will have identical transition probability matrices even if the previous histories of the processes are different. It is possible to construct stochastic models that do depend on prior history, but they are no longer Markov processes.

Markov Chains

Markov chains are Markov processes for which the transition probabilities are constant over time. This results in a very simple evaluation method using matrix algebra. While the evaluation is simpler and faster and results in greater precision, the assumption of constant transition probabilities is unrealistic for most clinical decision problems except for models representing very short time horizons. Many probabilities in clinical decision problems do change over time, most notably the background probability of death.

Table 1 Transition probability matrix

		TO		
		WELL	*DISABLED*	*DEAD*
FROM	WELL	.6	.2	.2
	DISABLED	0	.6	.4
	DEAD	0	0	1

Evaluation Methods

Incremental Utility

In an ordinary decision model, utilities are the values assigned to terminal nodes of a tree and represent both quality of life and duration of the health state. In a Markov model, utilities are associated with Markov states, so only the quality of life component is represented. The *incremental utility* is a number between 0 and 1 that represents the amount of utility accrued by spending one cycle in a given state. As with ordinary decision models, incremental utilities may also include a cost component, and the incremental cost of a health state is the cost accrued by a patient being in that state for one cycle.

Matrix Algebra

When the transition probability matrix is constant over time, the model is a Markov chain. Markov chains may be solved with matrix algebra. The matrix algebra solution provides an exact solution to the Markov process, whereas the simulations are approximations whose accuracy depends on cycle length. Although the matrix algebra solution is simple and elegant, it is rarely used in clinical applications of decision analysis because the assumption of constant probabilities over time

is not appropriate for most clinical problems. The details of the matrix algebra solution are beyond the scope of this entry but may be found in the 1983 work of Beck and Pauker.

Cohort Simulation

When the transition probabilities are not constant (the more general case), there is no closed form solution, and a Markov process must be evaluated by simulation. There are two simulation methods, cohort simulation and Monte Carlo simulation. In *cohort simulation*, a hypothetical cohort of identical patients is simulated passing through the process. During each cycle, the membership of each state is determined from the membership of the previous state and the transition probabilities from each state to every other state.

The cohort simulation is represented graphically in Figure 4. The starting state (Figure 4a) shows all members of the cohort in the starting (WELL) state. At an intermediate time, the cohort is distributed among the three states (Figure 4b), and eventually (Figure 4c), all members are in the absorbing DEAD state.

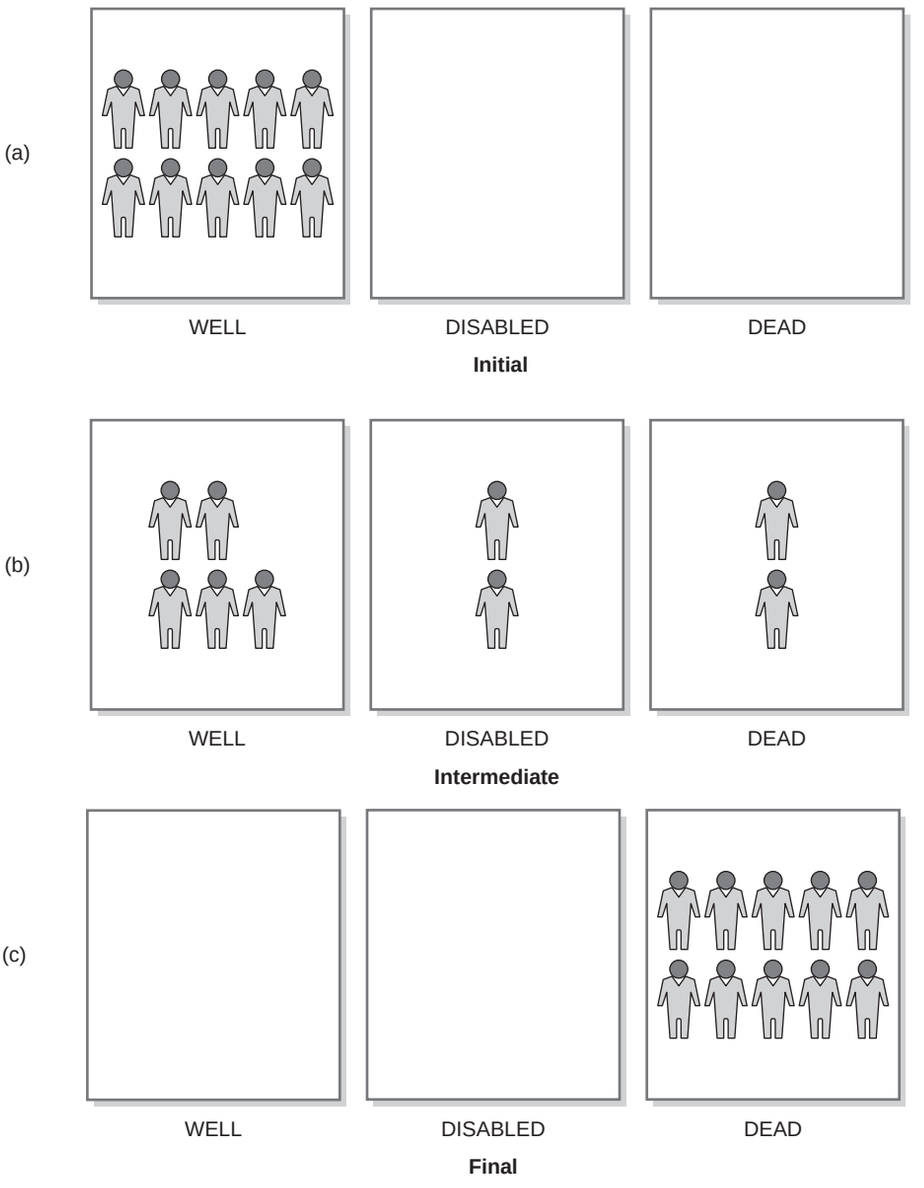


Figure 4 Cohort simulation: (a) initial state, (b) intermediate state, and (c) final state

Table 2 Markov cohort simulation

Cycle	*WELL*	*DISABLED*	*DEAD*	*Cycle Sum*	*Cumulative Sum*
Start	10,000	0	0	0	0
1	6,000	2,000	2,000	7,400.0	**7,400**
2	3,600	2,400	4,000	5,280.0	**12,680**
3	2,160	2,160	5,680	3,672.0	**16,352**
4	1,296	1,728	6,976	2,505.6	**18,858**
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
23	0	1	9,999	0.7	**23,751**
24	0	1	9,999	0.7	**23,751.7**
25	**0**	**0**	**10,000**	**0**	**23,751.7**

The *cycle sum* for cycle *i*, representing the contribution of each cycle to the expected utility of the simulation, is

$$\text{Cycle sum} = \sum_{j=1}^{s} \text{uINCR}_{ij} \times p_{ij},$$

where uINCR_{ij} is the incremental utility of state *j* during cycle *i* and p_{ij} is the probability of being in state *j* during cycle *i*.

The net expected utility for the Markov process is the sum of cycle sums over all cycles of the process:

$$\text{Cumulative sum} = \sum_{i=1}^{c} \sum_{j=1}^{s} \text{uINCR}_{ij} \times p_{ij},$$

where *c* is the number of cycles in the Markov process.

Stopping Criteria

A simulation must involve a finite number of cycles to be tractable. Therefore, a stopping criterion for the simulation must be specified. Often, the simulation is run until the entire cohort is "absorbed" (all patients are in an absorbing state). In practical terms, this means running the simulation until all patients are dead. The expected utility is then the quality-adjusted life expectancy (and/or expected cost) of the cohort. Usually, the stopping criterion is defined as the point at which the cycle sum falls below some predetermined threshold (e.g., .0001 quality-adjusted cycles) or when the proportion of the cohort in the DEAD state exceeds a threshold (e.g., .9999). Simulations may also be run for a predetermined number of cycles (e.g., 6 months or 5 years) to represent short-term events.

Spreadsheet Simulation

A simple method of carrying out a cohort simulation uses standard spreadsheet software, as illustrated in Table 2.

The simulation starts with 10,000 subjects in the starting state. This number is arbitrary for illustration purposes. The method would be no different if 100,000 or 1,000 subjects were used. There is a column for each Markov state. The first row shows the starting distribution of the Markov process. In this case, all 10,000 subjects begin in the WELL state. Each cell calculates the membership of the corresponding state based on the membership of the states in the previous cycle and the transition probabilities. For example, the membership of the DISABLED state at the end of Cycle 1 is

$$\begin{aligned}&\text{Membership}(\text{WELL}_{\text{start}}) \times p(\text{WELL,DISABLED})\\ &+ \text{Membership}(\text{DISABLED}_{\text{start}}) \times (1 - p(\text{DISABLED,DEAD})).\end{aligned}$$

The first term represents members entering the DISABLED state from the WELL state. The second term represents members remaining in the DISABLED state, which is the probability of DISABLED members not dying during the cycle. By filling in each cell of the spreadsheet appropriately, the membership of each cycle is determined.

The cycle sum is calculated by multiplying the membership of each state in each cycle by the incremental utility of the state and summing the products. The cumulative sum for each cycle is calculated by adding the cycle sum to the cumulative sum from the previous cycle. When the membership of the WELL and DISABLED states falls below 1 in 10,000, the simulation is terminated. In this case, the expected utility is the cumulative sum of 23,752 quality-adjusted cycles divided by the number of subjects in the cohort, or 2.38 quality-adjusted cycles per subject.

Markov Cycle Trees

Although the spreadsheet method is conceptually simple, it would become unwieldy with a larger number of states. Moreover, it relies on knowing the transition probability matrix during each cycle. The matrix is straightforward when transitions are simple between states. However, in most realistic decision problems, the transitions between states may result from complex sequences of events, and there may be more than one path between pairs of states. For example, the transition from WELL to DEAD may occur by developing an illness and dying, undergoing a treatment and dying from the treatment (e.g., surgery), or dying from a complication of treatment (e.g., a pulmonary embolism) or from unrelated causes. This can make determination of the net transition probabilities complicated and very difficult to recalculate if the model is modified.

To address this issue, Hollenberg devised the *Markov cycle tree*. The Markov cycle tree represents all the possible events occurring during each Markov cycle as a tree, as illustrated in Figure 5. The root is a special node called a *Markov node*. There are several differences from an ordinary decision tree:

- The branches of the Markov node represent the Markov states.
- Each branch of the Markov node is associated with an incremental utility.
- The terminal nodes represent the state in which a subject reaching that terminal node will begin the next cycle.

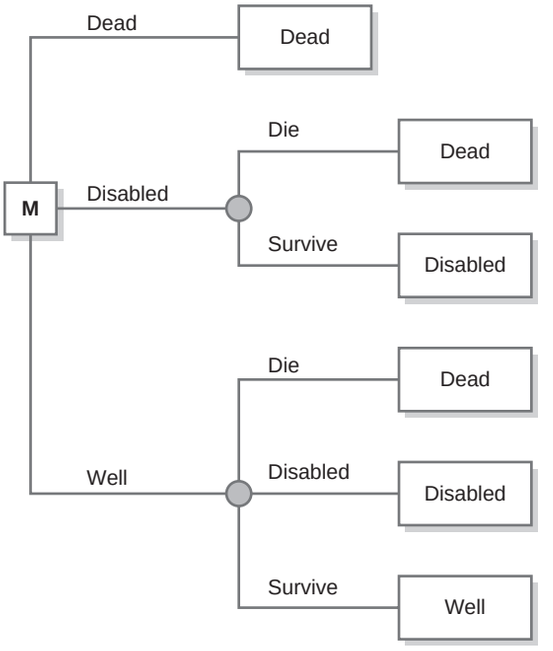


Figure 5 Markov cycle tree

When evaluated with specialized software, the cycle tree mechanism makes it unnecessary to calculate the overall transition probabilities from one state to another. The analyst needs only to assign the correct probabilities at each branch, and the software will determine the membership of each state at the end of each cycle.

Monte Carlo Simulation

Another way of evaluating Markov processes (those with time-variant transition probabilities) is Monte Carlo simulation. Instead of simulating a cohort simultaneously passing through the Markov process, Monte Carlo simulation considers a traverse of a single subject at a time through the simulation. This is illustrated in Figure 6. A single subject is pictured entering the simulation in the WELL state. In each cycle, there is a probability that the subject will make a transition from the WELL state to the DISABLED or DEAD state. During the third cycle, the subject makes a transition to the DISABLED state, and during the sixth cycle, the subject makes a transition to the DEAD state and the trial is ended. Each trial is continued either until the subject is absorbed (dies) or until a prespecified number of cycles is reached. As with the cohort simulation, each cycle spent in a state is credited with the incremental utility for that state.

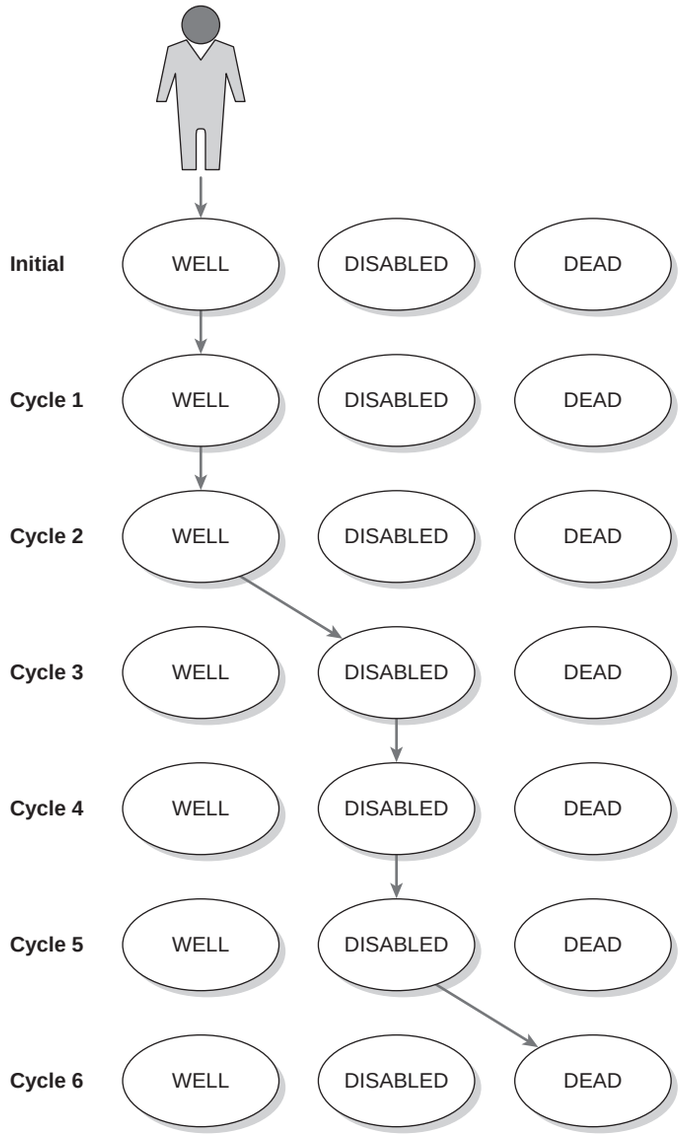


Figure 6 Monte Carlo simulation

Monte Carlo simulations are carried out with a large number of trials, typically 10,000 to 100,000. The exact sequence of states visited varies randomly from one trial to the next but is determined by the transition probabilities. The cumulative sums for all the trials form a distribution with a mean and variance.

The characteristics of the individual subjects entering the distribution may also be drawn from distributions. For example, the age of subjects may be drawn from a distribution that represents the age distribution in a population, and other parameters, such as background mortality rate, can then be calculated from that age for each trial.

Frank A. Sonnenberg

See also Decision Trees, Evaluation With Monte Carlo; Markov Models, Cycles; Markov Processes; Quality-Adjusted Life Years (QALYs)

Further Readings

Beck, J. R., & Pauker, S. G. (1983). The Markov process in medical prognosis. *Medical Decision Making, 3,* 419–458.

Doubilet, P., Begg, C. P., et al. (1985). Probabilistic sensitivity analysis using Monte Carlo simulation: A practical approach. *Medical Decision Making, 5*(2), 157.

Hollenberg, J. P. (1984). Markov cycle trees: A new representation for complex Markov processes. *Medical Decision Making, 4,* 529.

Sonnenberg, F. A., & Beck, J. R. (1993). Markov models in medical decision making: A practical guide. *Medical Decision Making, 13*(4), 322.

Torrance, G. W., & Feeny, D. (1989). Utilities and quality-adjusted life years. *International Journal of Technology Assessment in Health Care, 5*(4), 559–575.

Weinstein, M. C., Siegel, J. E., et al. (1996). Recommendations of the panel on cost-effectiveness in health and medicine. *Journal of the American Medical Association, 276*(15), 1253–1258.

Markov Models, Cycles

The Markov model provides a means of representing clinical situations in which risk is continuous, probabilities may change over time, or events may occur more than once. Markov models represent clinical events as transitions between health states, known as Markov states. Several advanced techniques extend the versatility of this modeling method.

Appropriate Use of Rates and Probabilities

Rates Versus Probabilities

Because Markov models inherently represent the passage of time, the transition probabilities must reflect a specific time frame. The Markov model shown as a state transition diagram in Figure 1 and as a Markov cycle tree in Figure 2 represents a person in the WELL state becoming either DISABLED or DEAD or a person in the DISABLED state becoming DEAD. Since the events represented in the cycle tree may happen during every cycle, the transition probabilities must reflect the length of time of one cycle. For a given degree of risk, the probability that any event will occur is greater for a longer period of time than for a short period.

The management of probabilities and appropriate matching to time in Markov models is accomplished by expressing risks as *rates* and converting rates to probabilities for the appropriate time frame. A probability is the likelihood (on a scale of 0 to 1) that an event will happen. The key distinction between a rate and a probability is that a probability is a dimensionless quantity that may *apply* to a particular time frame. On the other hand, a *rate* is a *ratio* that intrinsically includes time. Rates are analogous to instantaneous velocity.

$$\text{Rate} = \frac{p\text{Transition}}{\text{Unit time}}.$$

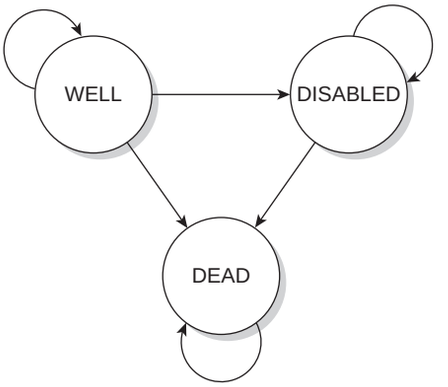


Figure 1 State transition diagram

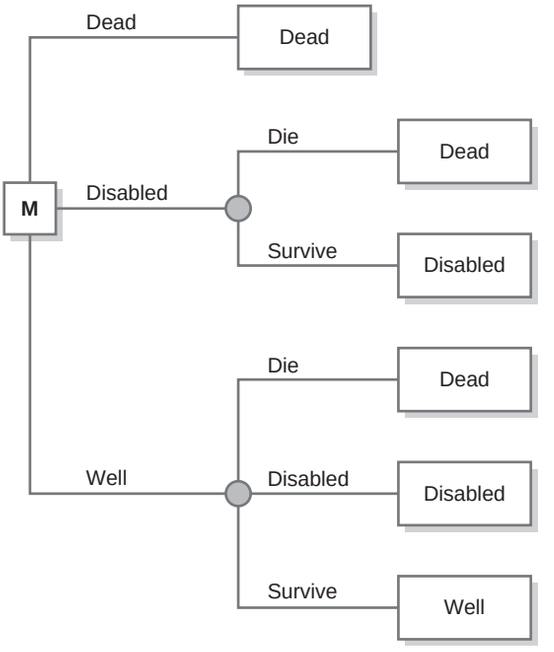


Figure 2 Markov cycle tree

The declining exponential approximation of life expectancy (DEALE) described by J. Robert Beck and Stephen G. Pauker provides a framework in which constant mortality rates can be manipulated. Consider Figure 3, which depicts a cumulative mortality curve for a constant mortality rate. The cumulative probability of death having occurred at time t is described by the equation

$$\text{Cumulative Mortality} = 1 - e^{-\text{rate} \times t},$$

where t is the time and *rate* is the mortality rate expressed in the same units as t.

The probability of death (or the fraction of a cohort that dies) in a given length of time is always slightly less than the rate expressed in the same length of time. This is because for each increment of time, a certain fraction of the cohort dies and consequently the size of the surviving cohort is smaller than the original cohort. In the next increment of time, the deaths represent a smaller fraction of the original cohort than the deaths in the first increment. In the above example, if the mortality rate is .05/year (.05 death per person per year) and time is 1 year, then the fraction of the original cohort that will be dead in 1 year is

$$\text{Cumulative Death} = 1 - e^{-0.05} = .0488,$$

which is slightly (2.4%) less than the rate of .05. The larger the rate, the more the cumulative probability differs from the rate. For example, if the rate is .2, then the cumulative death probability is .181, with a difference of 9.5%.

Standardized mortality figures published by the United States National Center for Health Statistics (NCHS) are published as "life tables." Each table gives the probability that a person of a given age will die between age x and age $x + 1$. The mortality *rates* are not provided in these tables but can easily be calculated by inverting the above formula:

$$\text{Mortality Rate} = \ln(1 - P).$$

Scaling Rates and Converting to Probabilities

The denominator (length of time) for published rates and probabilities (such as the NCHS life tables) is most commonly 1 year. In a Markov model, the probability needed is the probability that an event occurs during one Markov cycle. If the cycle length is not equal to the length of time in the published probability, then the probability must be scaled appropriately. Probabilities cannot be simply divided and multiplied. For example, if there were a 10% chance of an event occurring in 1 month, it would be incorrect to say that the probability of an event occurring in 2 months would be 10 × 2% or 20%. The reason is the same as that for the difference between rates and probabilities. After 1 month, only 90% of the original cohort is left, so the fraction of cohort experiencing the event in the second month is 10 × .9 or 9%. Thus after 2 months, the cumulative mortality probability is .1 + .09 or .19, rather than .20. The formula for cumulative death given in the foregoing produces the same result.

To model situations with various cycle times and with various time horizons, it is most convenient to convert the original probability to a rate, multiply

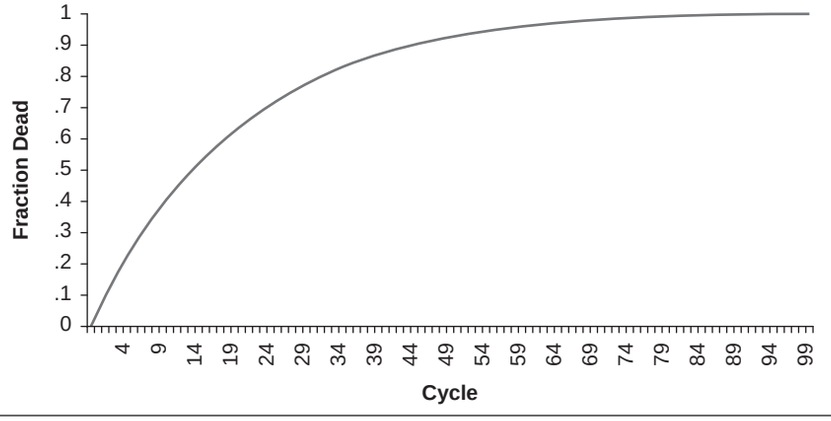


Figure 3 Cumulative mortality curve

the rate by the time interval needed, and then calculate the probability based on the new rate. Thus to determine the monthly rate represented by an annual mortality probability of .05, the rate corresponding to the annual mortality probability of .05 is

$$\text{Annual Rate} = \ln(1 - .05) = .0513.$$

The monthly rate then is .0513 / 12 = .004274.

The probability of death in 1 month (the parameter actually used in the Markov model) is therefore

$$\text{Mortality Probability} = 1 - e^{-0.004274} = .04265.$$

These calculations assume that the event rate is constant over each 1-year interval. However, it does not require assuming that the event rate is constant over the time horizon of the Markov model, because the monthly probability can be calculated separately for each year of the simulation.

Baseline muASR Versus Excess Mortality Rates

The mortality rate calculated from life tables, as described above, is referred to as the age, sex, and race-adjusted (ASR) mortality rate because the NCHS reports age-specific mortality separately for males and females and for white and nonwhite races. (NCHS also reports mortality as a composite for all races.) The corresponding population mortality rate ("baseline mortality"), representing death from all causes, is known as "muASR" because the Greek letter "mu" (μ) is common shorthand for "mortality rate."

In addition to baseline mortality, decision models consider mortality from specific causes. The mortality rates contributed by these causes are referred to as *excess mortality rates*. Excess mortality rates can be added to muASR to get a total or net mortality rate for persons with specific combinations of comorbidities. Thus,

$$\text{muTOTAL} = \text{muASR} + \text{muExcess},$$

where muExcess may consist of more than one component. For example, if someone has both coronary artery disease (CAD) and diabetes, we would have

$$\text{muTOTAL} = \text{muASR} + \text{muCAD} + \text{muDiabetes}.$$

Then, muTOTAL can be used to calculate the necessary probability of death in each cycle.

Avoiding Double Counting: Subtracting Mortality From Comorbidities

When constructing a model that includes comorbidities that constitute major causes of mortality in the population (e.g., CAD), adding the excess mortality for CAD to muASR would result in double counting because muASR already includes a component from CAD. To avoid this double counting, the tables of muASR must be adjusted by subtracting the population mortality rate for CAD. Note that this population mortality rate from CAD is not the same as the CAD mortality rate in *a patient with CAD* because the population includes many people who do not have CAD. Similarly, any major comorbid conditions considered in the model should have their corresponding population mortality rates subtracted from muASR before it is used to calculate background mortality in a model.

Incremental Utility

In a Markov model, the utility (quality of life) of each state is multiplied by the membership of each state during each cycle. The sum of these products is the cycle sum, and the sum of all cycle sums is the cumulative sum, which represents the quality-adjusted life-expectancy for the Markov simulation. Thus, a cohort that is 50% in the WELL state (utility = 1.0), 25% in the SICK state (utility = .8), and 25% in the DEAD state (utility = 0) will have a cycle sum of .5 + .25 × .8 + 0 = .7. The utility quantity for each state in a given cycle is known as the *incremental utility* because it is the increment of utility applied for spending one cycle in that state. It is analogous to the quality of life for a state.

The Half-Cycle Correction

The evaluation method for Markov cohort simulations or Monte Carlo simulations is a discrete process, meaning it is carried out as a series of discrete steps with a fixed cycle time. Therefore, accounting of state membership must be done either at the beginning of each cycle or at the end.

The necessity of the half-cycle correction derives from the fact that transitions occur throughout a cycle, so the membership of the state is overestimated when counted at the beginning of a cycle and underestimated when counted at the end of the cycle. Transitions occur, on average, in the middle of each cycle.

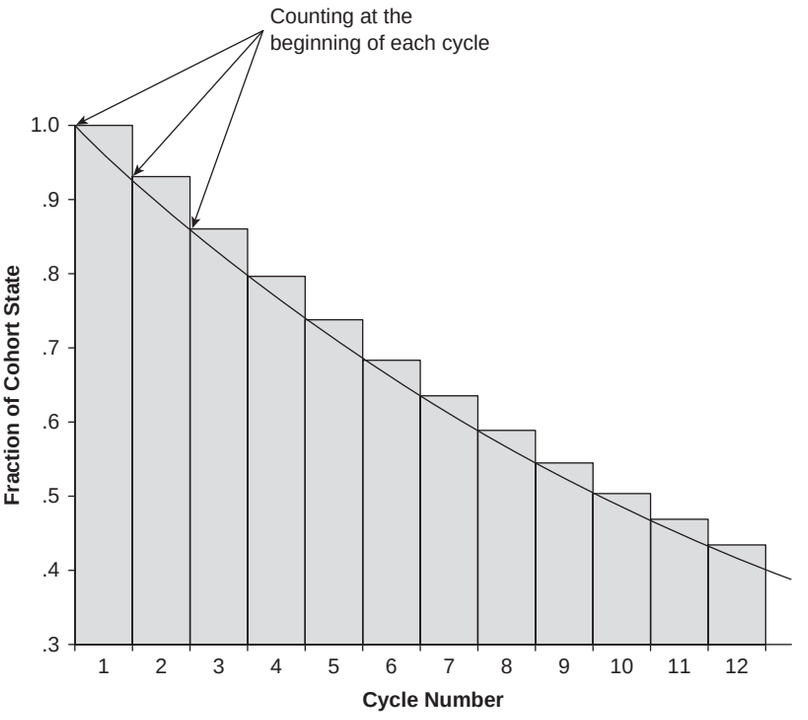


Figure 4 Overcounting

Figure 4 illustrates the membership in a Markov state over several cycles. The curve represents the actual membership of the state over time. Each rectangle represents the calculated contribution of the state to the cycle sum. If counting is done at the beginning of each cycle, as in Figure 4, then the area of the rectangle overestimates the contribution of the state to each cycle by the area of the rectangle above the curve. On the other hand, if the counting is done at the end of each cycle as in Figure 5 (the convention in Markov simulations), the area of each rectangle underestimates the contribution of the state to each cycle by the area of each rectangle falling below the curve but above the rectangle.

By adding a half cycle at the beginning of the simulation (represented by a rectangle half as wide as the others, with height equal to the starting membership of the state) but above the rectangle, and counting membership at the end of each cycle, the net effect is to count membership in the middle of each cycle, thus closely counterbalancing the over- and undercounting as illustrated in Figure 6. The additional half cycle is added by specifying a parameter referred to as the *initial utility* for each state, which is equal to half the starting membership of the state multiplied by the incremental utility of that state. It is apparent from Figure 6 that any mismatch between over- and undercounting will be less if a shorter cycle time is used.

Tail Utility. The simple counting method illustrated in Figures 4 to 6 works when the model is run until the entire cohort is absorbed (dead). However, when the simulation is stopped after a finite number of cycles (e.g., after 5 years), the remainder of the life expectancy must be accounted for. Suppose that the simulation is stopped after 12 cycles. As shown in Figure 7, the simple counting method would result in overcounting one half cycle beyond the end of the 12th cycle. To correct this, the last cycle sum must be reduced by half as illustrated by the hatched rectangle in Figure 8. This adjustment is not necessary when the simulation is run to absorption because the cycle sum (area of the rectangle) for the last cycle is close to zero.

In some cases, the Markov model may be run for a finite number of cycles to model detailed events

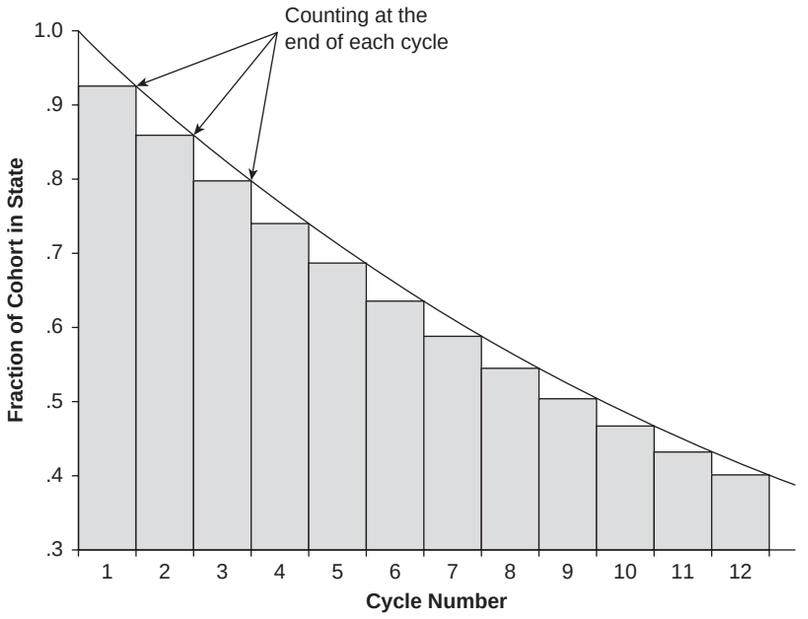


Figure 5 Undercounting

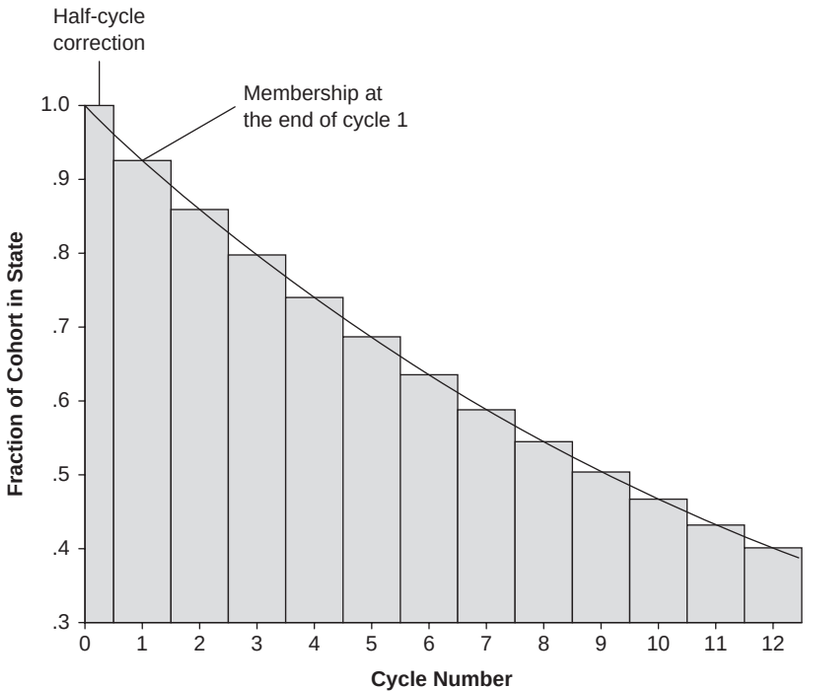


Figure 6 Half-cycle correction

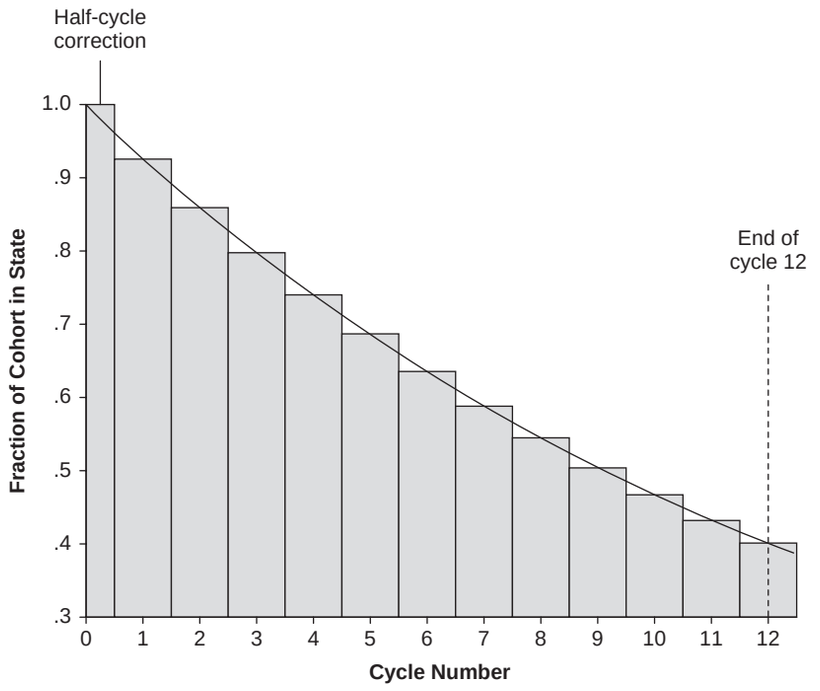


Figure 7 Stopping after a finite number of cycles

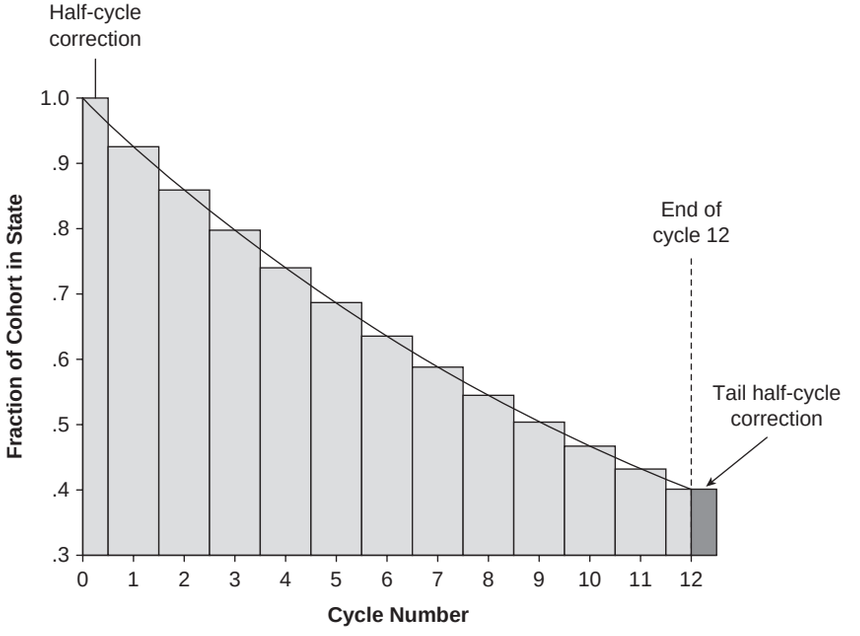


Figure 8 Tail utility correction

over a short time horizon, but the subsequent prognosis may still be of interest. For example, the 12 cycles may represent the events occurring during the first 12 months after an intervention, after which detailed data about event rates are no longer available. In that case, the remaining prognosis can be modeled by adding a component to the tail utility of each cycle at the end of the simulation.

Dual (Cost and Health) Increments

As in simple trees, Markov models may be adapted for cost-utility models by creating a dual-utility structure with one component for the cost and one for the health outcome. The state-specific utility components correspond to the *incremental utility* of each state; there is an incremental cost utility and an incremental health utility. The cost component must also be added to the *initial utility* for each state (to implement the half-cycle correction) and to the tail utility if it is used in the analysis.

Determining Cycle Length

Cycle length is usually determined so that it best matches the events in the model and the available data. For example, a model that considers the long-term outcome of a chronic illness, such as hypertension, may have a cycle length of a year. A model that considers detailed events occurring over a short period of time (e.g., postoperative complications or the phases of a pregnancy) may require a monthly, or even a weekly, cycle length.

For Markov cohort simulations and Monte Carlo simulations, cycle length also determines the precision of the model. Long cycle lengths are subject to a greater discrepancy between the simulated expected utility and the actual expected utility. The ultimate in precision is the fundamental matrix solution for Markov chains, which represents an exact solution. There is a trade-off between precision and efficiency; the shorter the cycle length, the more cycles must be simulated for a given time horizon and, therefore, the longer the simulation will take.

Tolls

In the simple examples shown here, the utilities of the model are values or expressions attached to terminal nodes of the trees. Under certain circumstances, it is convenient to apply utility adjustments to events represented as branches in a tree rather than terminal nodes. For example, a model may contain a variable "Cost" that keeps track of the cost of a path through the tree. When the path includes a branch representing an event that has an associated cost (e.g., surgery), the cost can be added using a *toll*. This adds the cost of the event to the cumulative cost of the path through the tree. By attaching the cost only to the paths, including that branch, the contribution of the cost to the overall cost of a decision strategy is weighted appropriately according to the likelihood of following that path. Tolls may be used for financial costs and may also be used to represent a short-term-utility adjustment representing the morbidity of an associated clinical event.

Tunnel States

Temporary Markov states are states that have allowed transitions only to other states. Thus, it is impossible to remain in a temporary state for more than one cycle. In certain circumstances, it is important to model temporary situations that a subject can remain in for more than one cycle. This can be accomplished by the use of *tunnel states*. Tunnel states are a set of temporary states that must be visited sequentially as illustrated in Figure 9. This state transition diagram represents a pregnancy. The starting state is "Not Pregnant." Subjects can remain in the Not Pregnant state or can become pregnant moving to the Month 1 state. States "Month 1" through "Month 9" are tunnel states. They can be visited only sequentially. State "Month 9" can make a transition only to the "Not Pregnant" state. This structure allows modeling events that occur for a short length of time and also permits modifying the occurrence and probabilities of events differently for each tunnel state. For example, the risk of miscarriage could be defined for each month of the pregnancy by specifying it separately for each tunnel state.

Frank A. Sonnenberg

See also Decision Trees, Evaluation With Monte Carlo; Markov Models, Applications to Medical Decision Making; Markov Processes; Quality-Adjusted Life Years (QALYs)

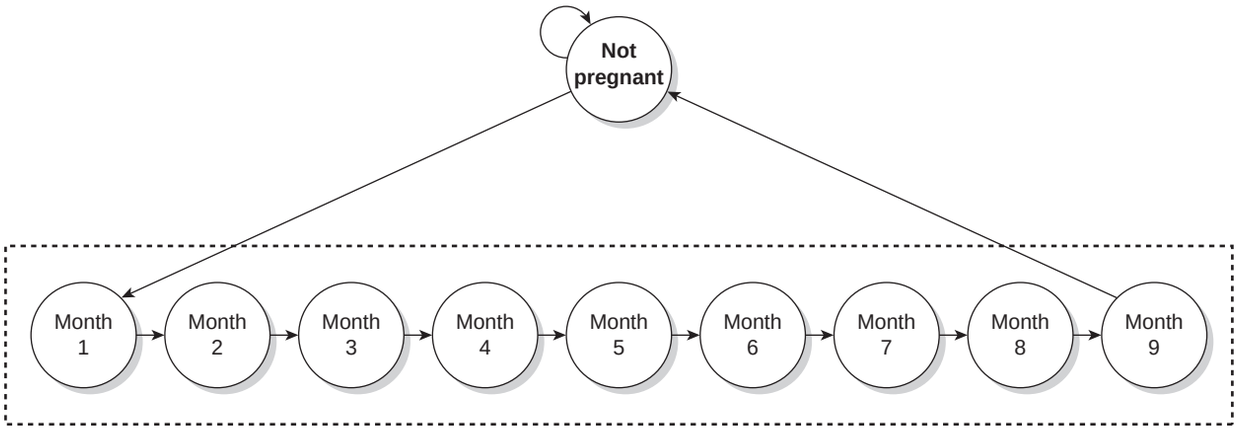


Figure 9 Tunnel states

Further Readings

Beck, J. R., Kassirer, J. P., et al. (1982). A convenient approximation of life expectancy (the "DEALE"): I. Validation of the method. *American Journal of Medicine, 73*(6), 883.

Beck, J. R., & Pauker, S. G. (1983). The Markov process in medical prognosis. *Medical Decision Making, 3,* 419–458.

Beck, J. R., Pauker, S. G., Gottlieb, J. E., Klein, K., & Kassirer, J. P. (1982). A convenient approximation of life expectancy (the "DEALE"): II. Use in medical decision-making. *American Journal of Medicine, 73*(6), 889.

Hollenberg, J. P. (1984). Markov cycle trees: A new representation for complex Markov processes. *Medical Decision Making, 4,* 529.

National Center for Health Statistics. (2008). *Life tables.* Retrieved December 30, 2008, from http://www.cdc.gov/nchs/products/pubs/pubd/lftbls/lftbls.htm

Sonnenberg, F. A., & Beck, J. R. (1993). Markov models in medical decision making: A practical guide. *Medical Decision Making, 13*(4), 322.

Markov Processes

Markov processes are mathematical processes in which, given the present state of the process, the future is independent of the past. They are named after the Russian mathematician Andrei Markov (1856–1922), who provided the first theoretical results for this type of process. They offer a flexible and tractable framework for medical modeling and are typically used to analyze processes that evolve over time. They can be used to aggregate information from different sources and to extrapolate short-term study results into the future.

A Simple Two-State Example

Markov processes can be used to model lifetime duration, for humans as well as devices. For example, of a group of hearing aids, some may fail early on, whereas others will last a long time before they eventually break down. If the probability to fail increases with time, then a graph of the failure times might look like a bell-shaped curve. However, for hearing aids, breakdowns will often be due to an accident, so a constant breakdown rate may be more realistic. In that case, in each period of time a certain proportion of the hearing aids will break down, and the distribution of the life duration follows an exponential distribution (Figure 1).

If a hearing aid has a constant breakdown rate that is equal to μ, then the mean life duration is $1/\mu$ and the lifetime, T, of this hearing aid follows an exponential probability distribution:

$$\mathrm{PR}\{T \geq t\} = \exp(-\mu t).$$

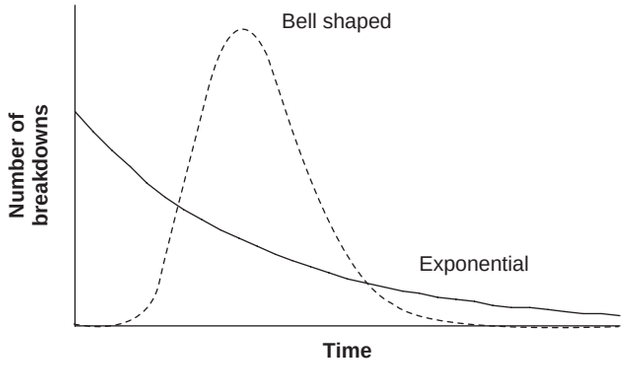


Figure 1 Graph of hearing aid failure times

For example, if the average life duration is half a year, then the annual breakdown rate is $\mu = 1/.5 = 2$, and the probability that the hearing aid survives the first year is equal to $\exp(-2 \times 1) = 14\%$. Because of the constant breakdown rate, the lifetime duration is memory-less: If the aid hasn't broken down yet after a year, then the aid's breakdown rate is still the same constant rate, μ, so the probability that the aid survives one more year is again 14%. In other words, the remaining lifetime is independent of the time already spent; the future is independent of the past. Markov processes are basically the extension of this memory-less property to more complicated processes, by introducing a state space.

The life of the hearing aid can be modeled as a Markov process with two states, indicating whether the aid has broken down or not, and a rate of transition, μ, from one state to the other (Figure 2).

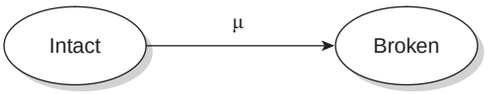


Figure 2 Markov process with two states

In this description of the process, the hearing aid can break down at any point in time. Instead of constantly looking at the hearing aid, one could observe its state only at the beginning of every week. This changes the continuous-time Markov process to a discrete-time Markov process (Figure 3).

From one week to the next, the hearing aid breaks down with probability p. The continuous-time and discrete-time models describe the same process, so their parameters μ and p are related. If the mean life duration of the hearing aid is half a year ($\mu = 1/.5 = 2$), then the probability that the hearing aid breaks down during any particular week is equal to

$$p = \Pr\left\{T \le \frac{1}{52}\right\} = 1 - \Pr\left\{T > \frac{1}{52}\right\}$$
$$= 1 - \exp\left(-\frac{\mu}{52}\right) \approx .038.$$

Compared with continuous-time Markov processes, discrete-time processes are considerably

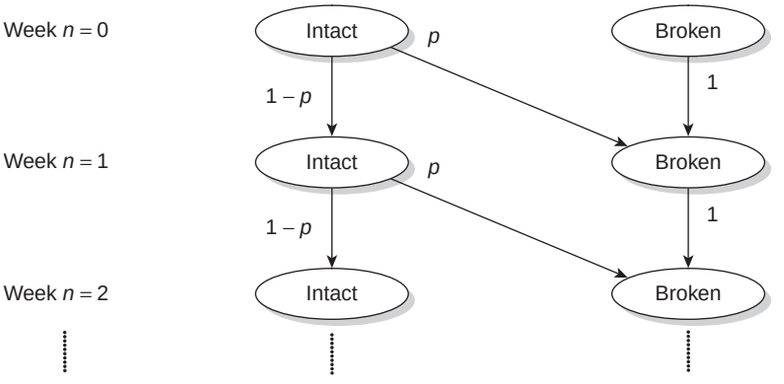


Figure 3 Discrete-time Markov process

easier to analyze. The probability that the hearing aid is intact (or broken) in the next week, $n + 1$, can be calculated from the probabilities in the current week, n:

$$\begin{aligned}\Pr\{\text{Intact in week } n+1\} &\quad (1 \quad p) \times \Pr\{\text{Intact in week } n\},\\ \Pr\{\text{Broken in week } n+1\} &\quad p \times \Pr\{\text{Intact in week } n\}\\ &\quad +1 \times \Pr\{\text{Broken in week } n\}.\end{aligned}$$

The first of these recursive formulae states that the hearing aid can only be intact if it was also intact in the previous week and remained intact (with probability $1 - p$). The second formula states that the hearing aid can be broken either if it was intact and broke down (with probability p) or if it was already broken. Suppose that initially the hearing aid was intact:

$$\begin{aligned}\Pr\{\text{Intact in week } 0\} &= 1,\\ \Pr\{\text{Broken in week } 0\} &= 0.\end{aligned}$$

Starting from these initial probabilities, the recursive formulae can be used to calculate the transient distribution, that is, the probability distribution through time ($n \geq 0$). The main advantage of the discrete-time process is that the recurrence formulae remain valid when the constant breakdown probability, p, is replaced by a nonhomogeneous, that is, time-dependent, probability p_n. For example, for humans, a nonhomogeneous breakdown rate is more realistic because the mortality rate increases with age.

The broken state is an absorbing state: Once broken, the hearing aid remains broken. As a result, in the long run, the aid will certainly be broken. The nature of the process changes considerably when the hearing aid can be repaired. Suppose that it takes some time to repair the aid, so each week a broken hearing aid may or may not be repaired. And suppose that, each week, it is repaired with probability $q = .8$. The introduction of repair makes this Markov process recurrent: Regardless of the initial state, the process will continue to alternate between the intact and the broken states. The new process has a unique long-run probability distribution. The long-run probability that the hearing aid is broken is equal to

$$\lim_{n\to\infty} \Pr\{\text{Broken in week } n\} = \frac{p}{p+q} \approx .045.$$

In the long run, the hearing aid will be intact for about 95% of the time. If the hearing aid is initially intact, then after the first week the probability of a broken aid is equal to $p \approx .038$, which is already quite close to the long-run probability (Figure 4).

Markov Chain Models

In general, mathematical models can be used to aggregate information from different sources and to extrapolate study results to other settings. For example, short-term results can be extrapolated into the future and intermediary outcome measures can be translated to measures of disease burden. Markov processes provide a flexible and tractable framework for medical modeling and are typically used to analyze processes that evolve over time. Most medical applications of Markov models use so-called Markov chains, that is, discrete-time Markov models that describe the state of the process at regular time intervals. Compared with continuous-time Markov models, discrete-time models are more easily analyzed and closely resemble medical research in cohorts of patients. They allow for transition probabilities that change with time, which is important, for example, to model age-dependent mortality.

A Markov chain model is defined by three types of parameters: the state space, the initial probability distribution, and transition probabilities. It can be graphically represented as shown in Figure 5.

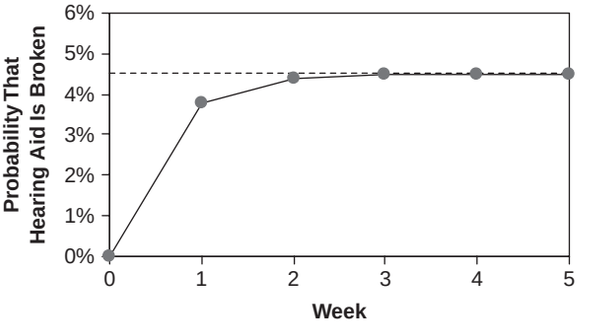


Figure 4 Long-run probability

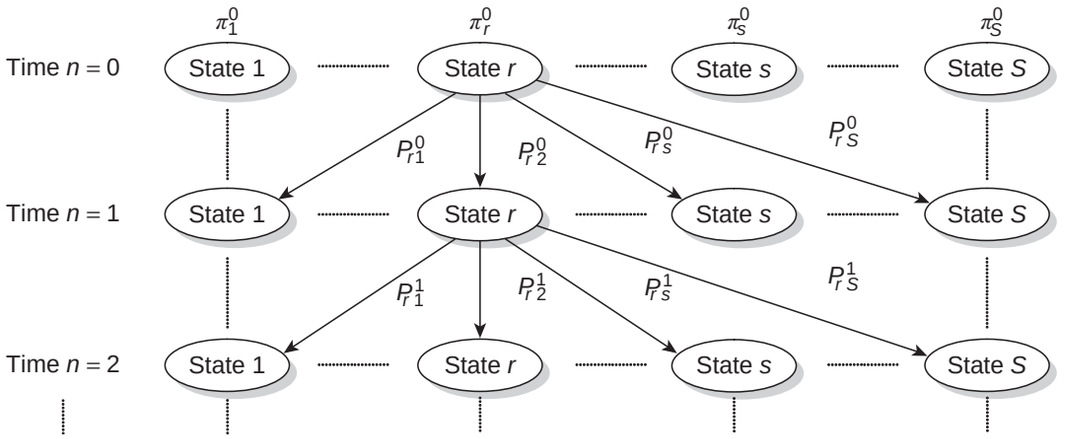


Figure 5 Markov chain model

The state space describes the states that the process can be in

$$\{1, \ldots, S\}, \text{ with } S < \infty. \tag{1}$$

For example, the states can describe the current severity of disease or the kind of treatments a patient has received in the past or whether the patient is or has been a smoker. The state space must be a finite countable set of exhaustive and mutually exclusive states. In addition, it must be such that it contains all information that is relevant for how the process will further evolve, because the transition probability matrix is only allowed to depend on the current state and on time. The initial probability distribution over the state space,

$$\pi_s^0 \text{ for } s \in \{1, \ldots, S\}, \tag{2}$$

describes the starting point of the process: The process starts in state s with probability $\pi^0{}_s$. Often, the process starts in one particular state with probability 1, for example, by starting with a patient who is initially healthy or has just received a particular diagnosis or treatment. In other cases, the initial probability distribution may reflect the fact that the actual initial state is unknown. For example, in genetic and diagnostic models, the initial probabilities can be distributed over two states that represent whether or not the patient has a particular mutation or disease. The third type of parameter of a Markov model is the matrix of transition probabilities,

$$P_{rs}^n \text{ for } n \geq 0 \text{ and } r, s \in \{1, \ldots, S\}. \tag{3}$$

The probabilities determine how the process evolves from one period to the next: Probability $P^n{}_{rs}$ is the probability that at time $n + 1$ the process will be in state s, if at time n the process is in state r. Either nothing changes ($s = r$) or a transition is made to a new state ($s \neq r$). Transitions to other states can, for example, represent the fact that a patient receives a particular treatment, that quality of life is improved, or that the patient dies. The transition probabilities are allowed to be nonhomogeneous, that is, time dependent.

With time, the Markov chain continues to jump from state to state. Apart from the actual process, its probability distribution over the state space evolves with time also. Conditional on the present, the future probability distribution of a Markov chain is independent of the past. As a result, the probability distribution at time $n + 1$ can be calculated recursively, by conditioning on the state of the process at the preceding time n:

$$\begin{aligned}
\pi_s^{n+1} &= \Pr\{\text{Process in state } s \text{ at time } n+1\} \\
&= \sum_{r=1}^{S} \Pr\{\text{Process in state } r \text{ at time } n\} \\
&\quad \times \Pr\{\text{Process in states at time} \\
&\quad \times n+1 | \text{Process in state } r \text{ at time } n\} \\
&= \sum_{r=1}^{S} \pi_r^n P_{rs}^n,
\end{aligned} \tag{4}$$

for $n \geq 0$ and $s \in \{1, \ldots, S\}$. This formula states that the probability to be in state s at time $n + 1$ (probability π_s^{n+1}) is equal to the probability to be in any state r at time n (with probability π_r^n) and then jump from state r to state s (with probability P_{rs}^n). Starting from the initial probability distribution, the transient distribution, that is, the distribution through time, can be calculated this way.

Rewards

Apart from the transient distribution, the purpose of modeling is often to estimate particular performance measures, such as how often a particular event occurs, discounted costs, or quality-adjusted survival. Rewards, or tolls, can be used to model such performance measures. This adds two types of parameters to the Markov model:

$$\text{state rewards } C_s^n, \text{ for } n \geq 0 \text{ and } s \in \{1, \ldots, S\}, \tag{5}$$

and

$$\text{transition rewards } C_{rs}^n, \text{ for } n \geq 0 \text{ and } r, s \in \{1, \ldots, S\}. \tag{6}$$

State rewards are associated with being in a state, whereas transition rewards are associated with going from one state to another. For example, disease-free survival can be calculated by assigning state reward 1 to all healthy states and 0 to the other states. Transition rewards can be used to estimate the frequency and costs associated with particular events. For example, if a state denotes that a patient is receiving treatment, then a transition reward can be used to model costs associated with acute reactions to that therapy. The probability that, in period n, the process starts in state r and then jumps to state s is equal to $\pi_r^n P_{rs}^n$. Therefore, summing over all possible states r and s, and assuming that transitions are made on average in the middle of each time interval, the expected total reward accrued in period n can be calculated as

$$C^n = \sum_{r=1}^{S} \sum_{s=1}^{S} \pi_r^n P_{rs}^n \left(\frac{1}{2} C_r^n + C_{rs}^n + \frac{1}{2} C_s^n \right). \tag{7}$$

Calculating the probability distribution (Equation 4) and the reward (Equation 7) only requires simple arithmetic operations. They can be implemented in general-purpose spreadsheet software, but special-purpose software is commercially available as well.

Limitations of Markov Chain Models

Discrete-time Markov models provide a flexible and tractable framework for many medical models. Still, Markov models do have a number of limitations that may or may not be resolved.

- In general, the applicability of mathematical models is limited by the availability and quality of data. Unreliable estimates of model structure and parameters inevitably lead to unreliable conclusions. Sensitivity analysis can be used to analyze the impact of such uncertainty.

- In Markov chains, the next state only depends on the present state of the process. This may not always be sufficient for a valid model. For example, excess mortality from smoking depends on whether someone has smoked in the past, and the probability of cancer recurrence tends to decrease with time. In these examples, the observable present is insufficient to model mortality. This can often be resolved by including the past in the state space (by distinguishing past smokers from never smokers) or by including unobservable characteristics (whether a patient is cured or not). For the real world, it is difficult to imagine how the future could be influenced by past events that are not somehow reflected in the present. Still, the information required to make the past irrelevant may be too large to fit in a finite state space.

- Calculating the summation in Equation 4 requires that the state space $\{1, \ldots, S\}$ be discrete and finite, which may sometimes be too restrictive. First, the natural state space may be continuous instead of discrete, such as blood pressure. This can usually be resolved by dividing the possible continuous values into a number of ranges of values. Second, a finite state space may be insufficient when, in theory, the number of states should be infinite. For example, the number of patients waiting to be treated is

potentially infinite. Also, difficulties may arise in models with more than one timescale. For example, excess mortality due to a particular disease is mostly reported in the literature as a function of the time since diagnosis, but different patients are diagnosed at different ages. The Markov time parameter, n, cannot be used to model both age and time since diagnosis, but adding either of the timescales to the state space would make the state space theoretically infinite. However, a finite state space can usually be obtained by setting reasonable bounds to the process.

- From one period to the next, Markov chains can only model a single transition. A continuous-time model might have shown more than one transition in that same period. For example, a model in which the time parameter, n, denotes the number of years may be appropriate for cancer models but will overlook many cases of flu. These problems can usually be resolved by choosing a shorter time interval. Still, shorter cycle times will lead to longer computation times, which may limit the use of large-scale models.

In cases where these limitations provide insoluble problems, simulation is likely to be the alternative method of choice. Simulation models provide an even more flexible framework than Markov models, but their statistical nature complicates the calculations and analysis.

Steady-State Analysis

Whereas most medical models use transient Markov chain analysis, most theoretical results for Markov processes concern their long-run behavior. Although the process itself will continue to jump from state to state, its long-run probability distribution will under certain conditions stabilize and converge toward a stationary or steady-state distribution. A homogeneous Markov chain on a finite countable state space has a unique steady-state probability distribution if the state space is irreducible (all states are accessible from each other) and aperiodic (for any state, a return to that state need not only occur in multiples of k (≥ 2 periods). This steady-state distribution is independent of the initial distribution of the process. It can be calculated by repeatedly applying Equation 4 until the probabilities stabilize. An alternative method follows from the observation that, after stabilization, the time parameter, n, in Equation 4 can be omitted:

$$\lim_{n\to\infty} \pi_s^n = \pi_s = \sum_{r=1}^{s} \pi_r P_{rs}, \text{ for } s \in \{1, \ldots, S\}. \qquad (8)$$

This formula is a balance equation: in the long run, the number of jumps from state s on the left-hand side (i.e., the probability to be in state s) must be balanced with the number of jumps into state s on the right-hand side. Together with the condition that the probabilities add up to 1, Equation 8 provides a set of equations with a unique solution that is equal to the steady-state distribution. The steady-state probability, π_s, can be interpreted as the probability that an outside observer, entering the system after the process has been in operation for a long time, will find the process in state s. Also, π_s is the long-run fraction of time the process spends in state s. Long-run average rewards can easily be calculated from the steady-state distribution.

For medical applications with a lifetime horizon, the steady-state distribution is irrelevant because in a steady state all patients are dead. Steady-state analysis is only relevant for stable processes, with homogeneous transition probabilities, for example, if patients with a chronic condition can recover from the more severe states and mortality can be ignored. Also, queuing processes such as a waiting room or a transplantation waiting list may be stable enough to have a steady-state distribution.

Semi-Markov Models

In Markov chain models, given the present state of the process, the future is independent of the past. In nonhomogeneous Markov chains, transition probabilities can depend on the time since the process started, but they cannot depend on the time spent in a particular state. This may be unrealistic. For example, for most types of cancer, recurrence after treatment becomes less and less likely with time. Several modeling tricks are available to incorporate time dependency into transition probabilities, by extending the state space. For example, a cure model can be used to model right after treatment whether a patient can have a recurrence in the future: Since only part of the

patients have a risk of recurrence, the overall risk of recurrence decreases with time. Alternatively, so-called tunnel states can be used that basically add the time spent in a state to the state space description.

For example, consider a patient who suffers from episodes of depression. The patient can be either depressed (D) or nondepressed (ND), which could be modeled by a two-state Markov chain. However, early after an episode, the risk of a new episode may be larger than later on. This cannot be modeled using nonhomogeneous transition probabilities because the pattern depends on the time since the previous episode, not on calendar time. A solution is to add the time since the previous episode to the state space: Denote by ND_i the state that the patient is in the ith week since the previous episode and by p_i the probability that a period in state ND_i is followed by an episode of depression. If $p_1 > p_2 > ... > p_m$, then the risk of a new episode decreases with the time since the previous episode. The nondepressed states are called tunnel states because from one period to the next, the patient cannot remain in the same state: Either a new episode of depression starts or the patient continues in the tunnel.

Time dependency can also be incorporated using semi-Markov models. Standard Markov chains have fixed cycle times between jumps. In semi-Markov models (see Figure 6), the state of the process changes in accordance with a Markov chain, but the time between jumps can have any distribution that can depend on both the state that is jumped from and the state that is jumped to. Semi-Markov models can be implemented using multidimensional transition matrices; the implementation is relatively close to Markov chain methodology. More complex methodology is required if general continuous-time distributions are assumed between jumps, but results remain tractable because the analysis of the imbedded Markov chain at transition moments can be separated from the analysis of the sojourn times between jumps. For example, a patient with depression episodes can be modeled as a two-state semi-Markov process that switches between depressed and nondepressed, with decreasing-failure-rate Weibull distributions to model that a transition from either state becomes less and less likely with time.

Wilbert van den Hout

See also Decision Trees: Sensitivity Analysis, Basic and Probabilistic; Markov Models; Markov Models, Applications to Medical Decision Making; Markov Models, Cycles

Further Readings

Briggs, A., Claxton, K., & Sculpher, M. (2006). *Decision modelling for health economic evaluation.* Oxford, UK: Oxford University Press.

Hawkins, N., Sculpher, M., & Epstein, D. (2005). Cost-effectiveness analysis of treatments for chronic disease: Using R to incorporate time dependency of treatment response. *Medical Decision Making, 25*, 511–519.

Naimark, D., Krahn, M. D., Naglie, G., Redelmeier, D. A., & Detsky, A. S. (1997). Primer on medical decision analysis: Part 5. Working with Markov processes. *Medical Decision Making, 17*, 152–159.

Sonnenberg, F. A., & Beck, J. R. (1993). Markov models in medical decision making: A practical guide. *Medical Decision Making, 13*, 322–338.

TreeAge Software, Inc. (2007). *TreeAge Pro user's manual.* Williamstown, MA: Author. Retrieved from http://www.treeage.com/files/pro2007/pdfs/TreeAgePro2007.pdf

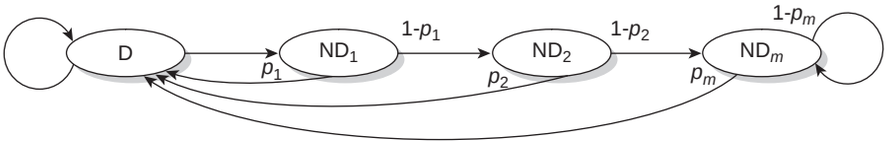


Figure 6 Semi-Markov model

Maximum Likelihood Estimation Methods

In medical decision making, statistical modeling plays a prominent role. The likelihood theory provides a generally applicable method to estimate and test parameters in a statistical model. The method goes back to one of the most famous statisticians from history, Sir Ronald A. Fisher, who worked on the method between 1912 and 1922. Most statistical methods make use of likelihood methods. This entry begins by explaining the likelihood function and then maximum likelihood estimation. Next, it discusses the properties of maximum likelihood estimation. This entry closes with a discussion of the application of likelihood methods for testing a null hypothesis.

The Likelihood Function

Consider a random sample of size n from a population where on each individual in the sample the value of an outcome variable Y is observed. Suppose a statistical model is available that specifies the distribution of Y up to an unknown parameter θ, which can be a single parameter or a vector of more parameters. If the outcome variable, Y, is discrete, its distribution is specified by the probability function, which gives the probability of each possible outcome value, y, given the parameter(s) θ. If Y is continuous, its distribution is described by the probability density function, which is a function such that the probability of Y taking a value between a and b corresponds with the area under its graph between a and b. The probability function or probability density function of Y is denoted by $f(y|\theta)$. It might depend on other observed variables X ("covariates") such as sex, age, and so on, but this dependence is suppressed in the notation. The observations are denoted by $y_1, y_2, \ldots, y_n$. The probability (density) function of one observation, say from individual i, is $f(y_i|\theta)$. The simultaneous probability of all observations in the sample is the product of $f(y_i|\theta)$ over all individuals in the sample. Given the observations, this is a function of the unknown parameter(s) and is called the *likelihood function,* $L(\theta)$:

$$L(\theta) = \prod_{i=1}^{n} f(y_i|\theta).$$

Example

Suppose one is interested in the unknown prevalence, θ, of type II diabetes in a certain population with age above 65 years. To estimate this prevalence, n individuals are randomly drawn (with replacement) from the population, and outcome Y is observed, $Y = 1$ if the individual in the sample has type II diabetes and $Y = 0$ if not. The probability of a random individual having the disease is θ, so his contribution to the likelihood function is θ. The probability of a random individual not having the disease is $1 - \theta$, so the contribution of a healthy individual to the likelihood function is $1 - \theta$. Suppose m individuals with the disease are observed in the sample. Then the likelihood function is

$$L(\theta) = \theta^m (1-\theta)^{n \quad m}.$$

Thus, if the sample size is $n = 300$, and 21 individuals with type II diabetes are observed, the likelihood function is

$$L(\theta) = \theta^{21}(1-\theta)^{279}.$$

Maximum Likelihood Estimation

According to the likelihood theory, the best estimate, $\hat{\theta}$ is that value of θ for which the likelihood function takes its largest value. $\hat{\theta}$ is called the *maximum likelihood estimate* (MLE). Thus the MLE is the parameter value under which the observed data have maximal probability. To calculate it in practice, mostly the natural logarithm of the likelihood,

$$l(\theta) = \log L(\theta) = \sum_{i=1}^{n} \log f(y_i|\theta),$$

is maximized. The reason is that sums are easier to handle mathematically than products. To determine $\hat{\theta}$, the derivative of the log likelihood, $l'(\theta)$, which is called the score statistic, is calculated and equated to zero:

$$l'(\theta) = \frac{dl(\theta)}{d\theta} = 0.$$

The solution gives the MLE. In the case of a p-dimensional parameter, the score statistic is a vector, and equating it to zero leads to a system of p

equations with p unknowns. Mostly, in practice this system of equations does not have an analytic solution, and iterative numerical methods have to be used.

Example

The log likelihood function in the above example is

$$\begin{aligned} l(\theta) &= \mathrm{m}\log\theta + (\mathrm{n}-\mathrm{m})\log(1-\theta) \\ &= 21\log\theta + 279\log(1-\theta). \end{aligned}$$

The score statistic is

$$l'(\theta) = \frac{dl(\theta)}{d\theta} = \frac{m}{\theta} - \frac{n-m}{1-\theta} = \frac{21}{\theta} - \frac{279}{1-\theta}.$$

Equating the score statistic to zero gives the MLE $\hat{\theta} = 21/300 = .07$, not surprisingly just equal to the sample prevalence.

Properties

Invariance

Suppose one is interested in estimating some function of the parameter(s), say $g(\theta)$. Then the MLE of the new parameter $\omega = g(\theta)$ is calculated by maximizing its likelihood function $L(\omega)$. An alternative way of getting an estimate for ω would be to calculate $\hat{\omega} = g(\hat{\theta})$ The *invariance* property of the MLE says that both estimates are identical.

Example

Suppose in the above example, one is interested in estimating the prevalence odds, ω, which is the ratio between the number of diseased and healthy individuals in the population. Thus $\omega = \theta/(1-\theta)$, and $\theta = \omega/(1+\omega)$. The log likelihood function of ω is

$$l(\omega) = 21\log\left(\frac{\omega}{1+\omega}\right) + 279\log\left(\frac{1}{1+\omega}\right).$$

Maximizing this yields the MLE $\hat{\omega} = 21/279$, which is, in accordance with the invariance property, identical to $\hat{\omega} = \hat{\theta}/(1-\hat{\theta}) = (21/300)/(1-21/300) = 21/279$.

Consistency

Under some weak assumptions that are in practice almost always fulfilled, the MLE is mathematically proved to be *consistent*, which means that if the sample size n tends to infinity, the difference between the MLE and the true value of θ tends to zero.

Bias

Often a MLE is unbiased, which means that the expected value of $\hat{\theta}$ is equal to θ, but not necessarily. However, under some weak regularity conditions, it can be proved mathematically that the MLE is asymptotically unbiased, that is, its bias tends to zero if the sample size tends to infinity. In the above example, $\hat{\theta}$ is unbiased, but $\hat{\omega}$ is not unbiased. However, the theory ensures that the bias in $\hat{\omega}$ tends to zero if the sample size tends to infinity.

Efficiency

The MLE is proved mathematically to be asymptotically efficient, which means that for large samples, no other estimator has a lower mean squared error. This is the main rationale for using the MLE.

Asymptotically Normal

The MLE of a single parameter $\hat{\theta}$ has asymptotically a normal distribution with mean θ and variance equal to the inverse of the Fisher information. This means that for large sample sizes, $\hat{\theta}$ is approximately normally distributed, the approximation becoming better if the sample size increases. Fisher's information, $I(\theta)$, is defined as the expected value of the squared score statistic:

$$I(\theta) = E_\theta l'(\theta)^2.$$

Since the score statistic has expectation zero, $I(\theta)$ is the variance of the score statistic. It can be shown that under some weak regularity conditions, $I(\theta)$ is equal to the negative of the expected value of the second derivative of the log likelihood:

$$I(\theta) = -E_\theta \frac{d^2 l(\theta)}{d\theta^2}.$$

Fisher's information can be estimated by substituting the MLE in the negative of the second derivative of the log likelihood:

$$\hat{I} = -\frac{d^2 l(\hat{\theta})}{d\theta^2}.$$

If θ is multidimensional, $I(\theta)$ is the negative of the matrix of the expected values of all second partial derivatives. Its estimate, $\hat{I}$, is obtained by substituting the MLE in the matrix of all second partial derivatives of the log likelihood. $\hat{I}$ is called the observed Fisher information.

Example

In the above example, the second derivative of the log likelihood is

$$\frac{d^2\, l(\theta)}{d\theta^2} = \frac{d}{d\theta}\left(\frac{m}{\theta} - \frac{n-m}{1-\theta}\right) = -\left(\frac{m}{\theta^2} + \frac{n-m}{(1-\theta)^2}\right).$$

Since the expected number of disease cases in a sample of size n is $n\theta$, the Fisher information is

$$I(\theta) = \frac{n\theta}{\theta^2} + \frac{n(1-\theta)}{(1-\theta)^2} = \frac{n}{\theta(1-\theta)}.$$

According to the likelihood theory, the distribution of the MLE $\hat{\theta} = m/n$ is approximately normal with mean θ and variance $\theta(1-\theta)/n$. In this special case, the exact distribution is binomial with the same mean and variance. It is well known that a binomial distribution is well approximated by a normal distribution if $n\theta$ and $n(1-\theta)$ are at least 5.

Standard Errors

A measure of precision of a single parameter estimate is provided through its standard error, which is defined as an estimate of its standard deviation. Thus the standard error of the MLE, $\hat{\theta}$, of a single parameter θ is

$$SE(\hat{\theta}) = \frac{1}{\sqrt{-\hat{I}}} = \frac{1}{\sqrt{-\frac{d^2 l(\hat{\theta})}{d\theta^2}}}.$$

Example

In the above example, the standard error of $\hat{\theta} = m/n$ is equal to

$$SE(\hat{\theta}) = \frac{1}{\sqrt{\left(\frac{m}{\hat{\theta}^2} + \frac{n \quad m}{(1 \quad \hat{\theta})^2}\right)}} = \sqrt{\frac{\hat{\theta}(1-\hat{\theta})}{n}}.$$

This is the well-known expression for the standard error of a proportion.

In general, the covariance matrix of a p-dimensional parameter is estimated by substituting the MLEs in the inverse of the negative of the matrix of second derivatives of the log likelihood:

$$\mathrm{Var}(\hat{\theta}) = -\begin{bmatrix} \frac{d^2 l(\hat{\theta})}{d\theta_1^2} & \cdots & \frac{d^2 l(\hat{\theta})}{d\theta_1 d\theta_p} \\ \vdots & \ddots & \vdots \\ \frac{d^2 l(\hat{\theta})}{d\theta_1 d\theta_p} & \cdots & \frac{d^2 l(\hat{\theta})}{d\theta_p^2} \end{bmatrix}^{1}.$$

Likelihood-Based Methods for Hypothesis Testing and Confidence Intervals

The likelihood theory provides three different generally applicable approximate methods for testing null hypotheses. The hypotheses may be quite general, single, or composite and may concern one or more parameters. All three methods lead to approximate confidence intervals constructed by inverting the test, that is, a confidence interval for a parameter consists of all values of the parameter that are not rejected when tested. In practice, the three methods mostly yield similar results, and the choice of the method is often just a matter of convenience.

The most well known is the method of Wald. It is based on the fact that the MLE follows approximately a normal distribution. For null hypotheses concerning a single parameter, this method leads to approximately standard normal test statistics of the form Z = estimate/standard error, and to approximately 95% confidence intervals of the form estimate $\pm 1.96 \times$ standard error.

The second method is the likelihood ratio method. It is based on the fact that two times the difference between the maximized log likelihood under the alternative and under the null hypothesis follows approximately a chi-square distribution under the null hypothesis. In the one-parameter case, the likelihood ratio confidence interval is the interval of θ_0 values for which $|\chi^2| < 3.84$.

The third method is the score method. It is based on the fact that the score statistic is asymptotically zero mean normally distributed with variance equal to the Fisher information. Again, in the one-parameter case, the 95% confidence interval consists of all θ_0 values for which $|Z| < 1.96$.

Example

Suppose that in the above example one wishes to test the null hypothesis that the type II diabetes prevalence is equal to some specific value θ_0 that for some reason is of interest. Wald's test statistic is

$$Z = \frac{\hat{\theta} - \theta_0}{\sqrt{\frac{\hat{\theta}(1 \quad \hat{\theta})}{n}}},$$

which has under the null hypothesis an approximate standard normal distribution. If $\theta_0 = .04$, $Z = 2.037$ and the corresponding p value is .042.

The approximate 95% confidence interval according to Wald's method consists of all values θ_0 that are not rejected at level $\alpha = .05$, that is, all values θ_0 for which $|Z| < 1.96$, leading to

$$\hat{\theta} \pm 1.96 \cdot \sqrt{\frac{\hat{\theta}(1-\hat{\theta})}{n}}.$$

In our example, Wald's approximate 95% confidence interval is (.041, .099).

The likelihood ratio test statistic is given by

$$\begin{aligned}\chi^2 &= 2\left[l(\hat{\theta}) - l(\theta_0)\right] \\ &= 2\Big[(m\log\hat{\theta} + (n-m)\log(1-\hat{\theta})) - \\ &\quad (m\log\theta_0 + (n-m)\log(1-\theta_0))\Big],\end{aligned}$$

which under the null hypothesis has an approximate chi-square distribution with one degree of freedom. If $\theta_0 = .04$, $\chi^2 = 5.79$ and the corresponding p value is .016.

The likelihood ratio 95% confidence interval is given by all θ_0 values for which $\chi^2 < 3.84$. In the example, this is (.045, .103).

The score test is given by

$$Z = \frac{\hat{\theta} - \theta_0}{\sqrt{\frac{\theta_0(1 \quad \theta_0)}{n}}},$$

which has under the null hypothesis an approximate standard normal distribution. In the example, $Z = 2.65$ with p value .008. The corresponding 95% confidence interval is obtained by calculating the interval of θ_0 values for which $|Z| < 1.96$. This confidence interval is known as the Wilson score confidence interval. In the example, it is (.046, .105).

Theo Stijnen

See also Distributions: Overview; Statistical Notations; Statistical Testing: Overview

Further Readings

Aldrich, J. (1997). R. A. Fisher and the making of maximum likelihood 1912–1922. *Statistical Science, 12,* 162–176.

Cox, D. R., & Hinkley, D. V. (1979). *Theoretical statistics*. Boca Raton, FL: CRC Press.

Measures of Central Tendency

When physicians or clinical researchers encounter a sample of data, they often try to get an overall picture about the data before proceeding with any analysis. For example, suppose a physician has a group of senior patients who would like their weights under control so that health problems developed by obesity could be minimized. Before the doctor prescribes any medications or gives dietary administration solutions to the patients, she or he might want to know the typical values of these patients' body mass index (BMI), which presents a reliable indicator of body fatness and is commonly used to monitor weight. These single summarized values are called measures of central tendency. A measure of central tendency attempts to describe the distribution of the data by identifying the most typical or representative individual score. This typical value can then be used to describe the entire data sample. Measures of central tendency are useful for making comparisons among different groups of individuals since they reduce a large number of measurements into a single typical value that makes comparisons easier. For instance, after the physician gets the typical BMI score of the patients, it is easy for her or him to know if most of the patients have healthy weight status or make a comparison to see whether this group of senior patients has better or worse weight status than the nationwide senior population. Measures of central tendency are

simple and useful. However, there is not a single standard procedure for determining a measure of central tendency in every situation. Mainly there are three most commonly used measures of central tendency: *mean, median,* and *mode.*

Mean

The most common measure of central tendency is the arithmetic mean, which is defined as the sum of all scores divided by the number of observations in the data sample. It can be denoted by the formula $(1/n)\sum_{i=1}^{n} x_i$, where n is the number of observations and X_i is the ith individual observation . The mean is easy to interpret and compute. It is an averaged value, not depending on the order of the data. It should lie in the range of the data, neither less than the smallest value nor greater than the largest one. In the above BMI example, if the physician wants to know the distribution of her or his patients' BMIs, the arithmetic mean of the BMIs might be considered by the physician as a proper representation. To ease the calculation, suppose there are 10 patients in total with their individual BMI values described as 25, 19, 23, 31, 21, 25, 20, 23, 26, and 35. The calculation of the mean is shown as follows:

$$\text{Mean} = \frac{25+19+23+31+21+25+20+23+26+35}{10} = 24.8.$$

Based on the mean value of BMIs, 24.8, the physician would think those patients have normal weight status since generally BMI values of 18.5 to 24.9 for adults 20 years old or older are considered healthy and no medications are needed for weight control. Note that as previously stated, the mean, 24.8, is less than the maximum value of BMIs (35) and greater than the minimum value (19). The mean has two important algebraic properties that make it the most widely used measure of central tendency. First, the deviations of each value from the mean sum up to zero. Second, the sum of the squared deviations will be less than the sum of squared deviations from any other constants. That is, the mean minimizes the squared deviations, a characteristic taken advantage of in many inferential statistics such as in the construction of confidence intervals. To illustrate this, let's go back to the BMI example. The BMI deviations from the arithmetic mean for each patient are calculated as 0.2, –5.8, –1.8, 6.2, –3.8, 0.2, –4.8, –1.8, 1.2, and 10.2, respectively. Apparently, the deviations will be summed up to 0 as shown in the equation $0.2-5.8-1.8+6.2-3.8+0.2-4.8-1.8+1.2+10.2=0$. Moreover, the sum of the squared deviations from the mean is equal to

$$0.2^2+5.8^2+(-1.8)^2+6.2^2+(-3.8)^2+0.2^2+(-4.8)^2+(-1.8)^2+1.2^2+10.2^2=221.6,$$

which is the smallest if any other constants are used to calculate deviations rather than the mean. For example, the deviations from another constant (say 25) will be 0, –6, –2, 6, –4, 0, –5, –2, 1, and 10, respectively. It is easy to verify that the sum of the deviations from 25 is –2 rather than 0, and the sum of the squared deviations is 226, which is greater than the sum of deviations from the mean, 221.6.

Median

The median is a value that separates the set of data into two groups, the highest half and the lowest half. Since 50% of the observations are less than the median, it is also called the 50th percentile or the second quartile. The median can be obtained by sorting the data into either ascending or descending order by their magnitudes, then repeatedly removing pairs of the currently largest and smallest values until one or two values are left. The left solo value is the median, but if there are two values left, the median is taken as the arithmetic mean of the two. In the BMI example, the original data sorted by ascending order are 19, 20, 21, 23, 23, 25, 25, 26, 31, and 35. Since the number of observations is 10, an even number, there will be two scores left after removing four pairs of the currently largest and smallest scores. The median is the arithmetic mean of the fifth and sixth smallest values (23 + 25)/2 = 24. This approach is typically applicable to small sets of data. A more general approach to calculate the median is described as follows. Let $\boldsymbol{n}$ be the number of observations in a sorted data series in ascending order. The median will be the $(\frac{n+1}{2})$th largest value

in the sorted data series if n is an odd number, and the arithmetic mean of the$(\frac{n}{2})$th and $(\frac{n}{2}+1)$th largest values if n is an even number. In our BMIs example, n is 10, an even number, so the median is the arithmetic mean of the fifth and sixth largest values, 23 and 25, respectively.

Mode

The most frequently occurring value in the data set is called the mode. Graphically, it is the highest point in a bar plot or histogram. The mode is fairly easy to calculate. However, it is not necessarily well defined, so it is possible to have more than one mode (bimodal or multimodal) or no mode in a data set. In the previous BMI example, frequency for each unique value of BMIs was calculated as below:

Values of BMI	19	20	21	23	25	26	31	35
Frequency	1	1	1	2	2	1	1	1

Since two values, 23 and 25, occur the most frequently, it is a bimodal data set with mode 23 and 25.

Choosing a Measure of Central Tendency

Which measure of central tendency should be used depends on the type and symmetry of the data. In terms of interval (e.g., BMI) or ratio data (e.g., donor to recipient weight ratio in kidney transplant), all three measures could be used, but the mode is least often used in this case. For ordinal data (e.g., pain scales), median and mode may be used, but the median takes the ranking of categories into account, providing more information about the data series than does the mode. Usually, the mean of an ordinal data series makes no sense, but if an assumption of equal metric distances between categories was made, the mean might also be appropriate for ordinal data. The mode is the only meaningful measure of central tendency for nominal data such as gender or ethnicity.

The symmetrical property is another important issue that needs to be considered when choosing a measure of central tendency. Generally, if the distribution of a data series is perfectly symmetric, as is the Gaussian (normal) distribution, the mean, median, and mode are the same. However, if the distribution is right (positively) skewed, as is a chi-square distribution with degrees of freedom of 3, they are in different locations, and the relative magnitudes would be mode < median < mean, whereas for a left (negatively) skewed distribution, the order would be mode > median > mean. Figure 1 illustrates the relationships.

The mean is the most appropriate representative for a symmetric distribution as it can be easily used in further statistical inference. However, the arithmetic mean can be easily influenced by extreme values as all given data are involved in the calculation of the mean. Let's revisit the BMI example. Suppose the last patient happened to be very obese, with a BMI value of 100 instead of the original 35; then, the arithmetic mean of BMI would soar to

$$\frac{25+19+23+31+21+25+20+23+26+100}{10}=31.3,$$

which is greater than any other BMI values except the obese patient. In this case, the mean departs from the majority of the data and loses its role as the typical value of the data series. Based on the mean of BMI, 31.3, the physician might get a biased impression that the majority of the patients are obese. Consequently, the physician might make a wrong decision to prescribe weight control medication to those patients and add unnecessary risk to the patients. In contrast with the arithmetic mean, the median is resistant to outliers as the computation of the medians does not get all the given data involved, whereby changing a small number of values to extreme magnitudes will not result in a large change of the median. In the BMI example, the median did not change when an outlier was arbitrarily introduced, and is still a good representative for the data series. Hence, the median is a commonly used measure of central tendency for a largely skewed data set. In practice, when a researcher reports a characteristic of the data, he or she needs to check the symmetry first to select a proper measure of central tendency. A histogram or bar plot is an effective visual tool for this purpose.

Changhong Yu

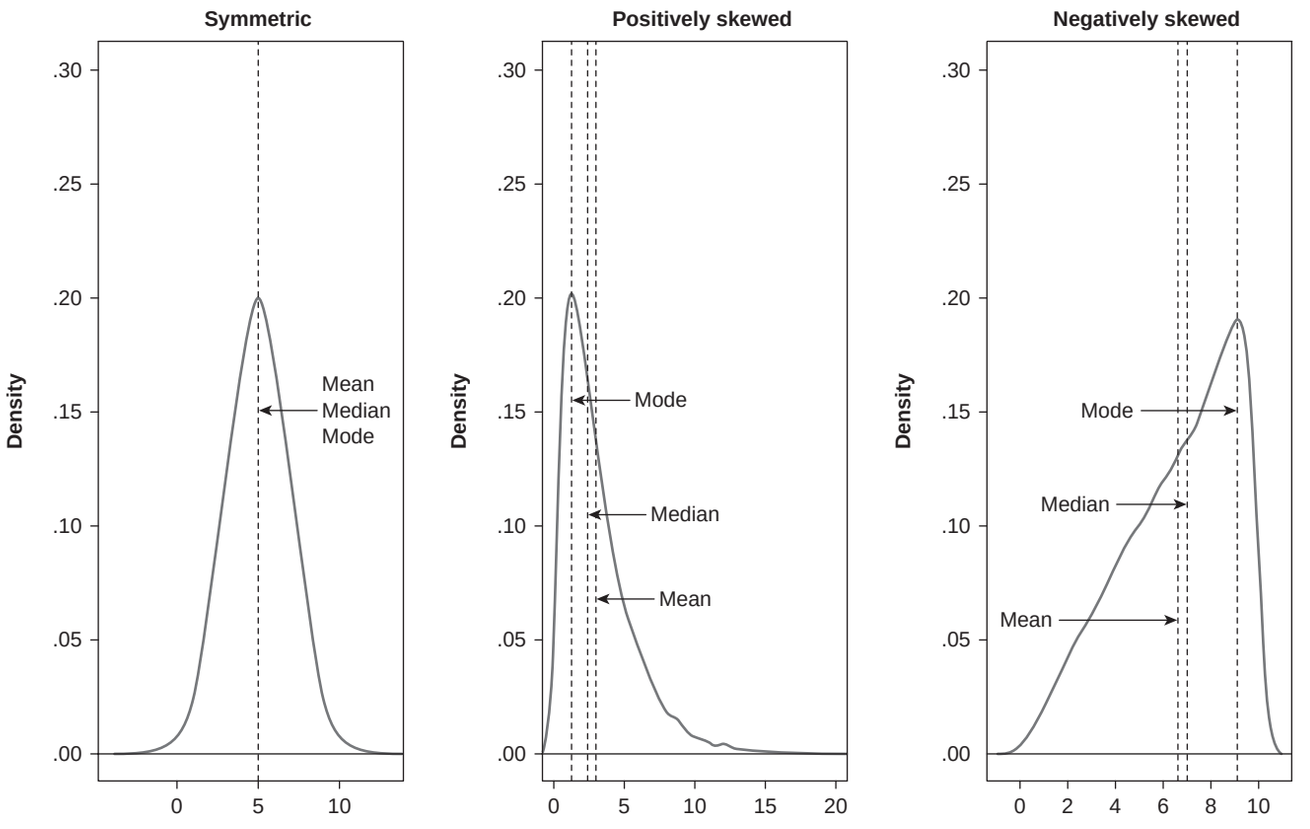


Figure 1 Skewness of data distribution

See also Frequency Estimation; Measures of Variability; Statistical Testing: Overview; Variance and Covariance

Further Readings

Hays, W. L. (1981). *Statistics* (3rd ed., chap. 4). New York: Holt, Rinehart, & Winston.

Kaas, R., & Buhrman, J. M. (1980). Mean, median and mode in binomial distributions. *Statistica Neerlandica, 34*(1), 13–18.

Kotz, S., & Johnson, L. N. (Eds.). (1985). *Encyclopedia of statistical sciences* (Vol. 5). New York: Wiley.

Rosner, B. (1995). *Fundamentals of biostatistics* (4th ed., chap. 2). Belmont, CA: Wadsworth.

Rumsey, D. (2003). *Statistics for dummies* (chap. 5). Hoboken, NJ: Wiley.

Swinscow, T. D. V. (2008, January). *Statistics at square one* (6th ed., chap. 1). Retrieved February 9, 2009, from http://www.bmj.com/statsbk

Measures of Frequency and Summary

Improving health for individuals and populations requires an understanding of patterns of disease occurrence. This is in large part a matter of counting—researchers count the number of times a disease occurs, how many deaths occur, and so on. Although simple at first glance, a proper understanding of the terminology and attention to potential errors in this work are important if the numbers are to accurately reflect the underlying truth. This entry describes the essentials of measuring the incidence and prevalence of disease.

Two key concepts form the foundation for this material: disease *risk* and disease *burden*. Risk is the probability of developing the disease of interest and can be assessed with measures of *incidence*.

When measuring incidence, researchers are assessing frequency, meaning they are counting events, that is, disease occurrences. At times their attention is at the population level, such as when they are assessing the number of new cases of HIV in a country over a particular calendar year. At other times they are interested in an individual who is free of disease and would like to make decisions based on the probability that this individual will develop the disease, such as when estimating risk of heart attack for a patient in a doctor's office. In both instances, researchers are interested in *new* events, that is, the development of disease. An individual with disease is called a *case*.

Disease burden, on the other hand, is a public health concept that is a function of the number of existing cases at a particular time and is assessed with measures of *prevalence*. With this perspective, researchers are not confining their attention to new cases but rather the number of individuals with a particular disease status. Prevalence measures are not used to estimate risk.

Before getting to healthcare examples, consider the following analogy. You are standing at the entrance of a local café, watching as customers enter, hang out for a while, and then leave. Each customer who enters is analogous to a new case. Each customer leaving is analogous to either a death or a cure. Therefore, counting up the number of individuals entering during a specific time period (e.g., 1 day) is analogous to incidence. Looking through the café window and counting the number of customers inside at a particular time (e.g., 12:00 noon) is analogous to prevalence. The number inside is a function both of how many are entering and of how long each customer stays. As we shall see, deaths and cures among cases affect disease duration and therefore prevalence.

Assessing Risk With Incidence Measures

This entry now discusses two variations on the concept of incidence: cumulative incidence (also known as incidence proportion) and incidence rate (also known as incidence density). Cumulative incidence is the proportion of healthy (disease-free) individuals who develop disease among a population at risk for the disease over a specified time period:

$$\text{Cumulative incidence} \quad \frac{\begin{array}{c}\text{Number of new cases of}\\ \text{disease in a population}\\ \text{at risk over}\\ \text{a specific period}\end{array}}{\begin{array}{c}\text{Number of individuals}\\ \text{in the population}\\ \text{during that period}\end{array}}.$$

Cumulative incidence is therefore expressed as a proportion, that is, a number between 0 and 1, although it may be multiplied by 100% if one wishes to express it as a percentage. Alternatively, one can express this per 1,000 persons or some other arbitrary number that fits the situation. So, for example, the cumulative incidence of diabetes over 10 years among a population of overweight individuals might be

$$\frac{18000}{100000} = .18 \text{ or } 18\%$$

or

$$.18 \times 1000 = 180 \text{ cases per } 1{,}000 \text{ persons.}$$

Figure 1 is a graphic representation of cumulative incidence, in which four subjects are observed, each for an entire 3-year period. The cumulative incidence is .25.

The denominator must include only disease-free individuals at the beginning of the time period because existing cases are not at risk for becoming new cases.

The specification of a time period is essential to understanding risk. For example, the great majority of individuals who live to age 80 develop hypertension (cumulative incidence approaches 1). However, if one's interest is in the short-term risk of development of hypertension among a population of teenagers, this number is not useful. In such a case, the 10-year cumulative incidence might be of more interest.

With cumulative incidence, all individuals in the denominator must be observed for occurrence of the disease for the entire time period. Often, however, a study will involve patients in which the periods of observation vary between subjects. This may be because some entered the study later than others or left the study earlier than others due to death or being lost to follow-up.

When each individual in a study has not been observed for the entire time period, the

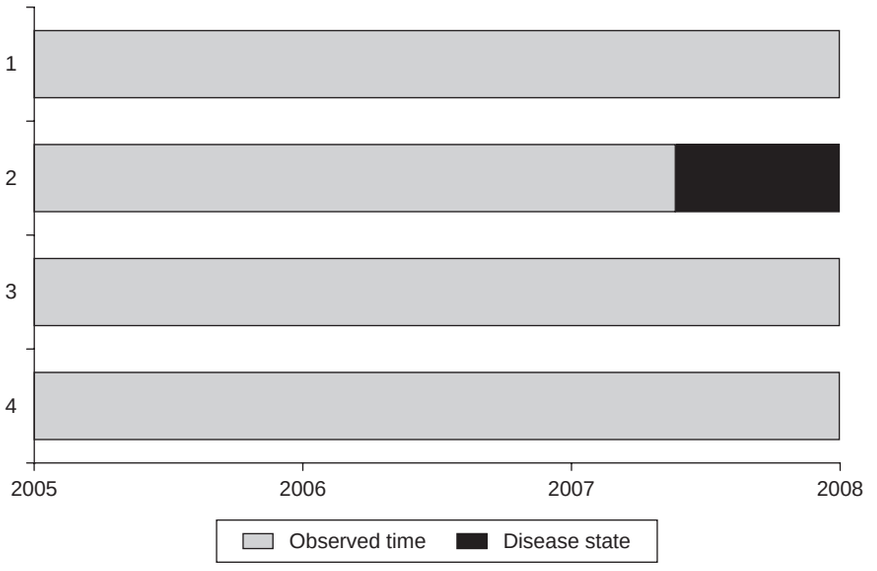


Figure 1 Cumulative incidence

Note: All four subjects have been observed for the entire time period. Subject 2 developed the disease.

appropriate measure is the *incidence rate*, also known as *incidence density*. Like cumulative incidence, this is a measure of new cases and therefore a measure of risk. Incidence rate is reported as the number of cases per person-time. Person-time is the sum of the time intervals of observation for each individual. For example, if four subjects have been observed for 3, 3, 2, and 2 years, respectively, then the researchers have 10 person-years of observation. One of these individuals may have developed the disease, so the researchers have

$$\text{Incidence rate} = \frac{\text{Number of new cases of disease in a population at risk}}{\text{Sum of person = time of observation}},$$

$$1/10 = .1 \text{ cases per person-year}$$

or

$$.1 \times 100 = 10 \text{ per } 100 \text{ person-years.}$$

Here the value is a rate, it has the units of cases per person-time, and, unlike a proportion, the value can exceed 1. Incidence rate is sometimes referred to as the *force of morbidity*.

Figure 2 demonstrates the idea of incident rate graphically. Subject 4 left the study, due either to death or being lost to follow-up, so the researchers do not count time after this point in the denominator. Subject 2 developed the disease after 2.5 years of observation and then remained under observation until the study concluded, so that the total observation time for this subject is 3 years.

The risk of developing disease in a disease-free individual can be estimated by the cumulative incidence or incidence rate in a population made up of subjects who are similar to the individual in question. The more precisely the individual matches up with the characteristics of the population, the more accurate the estimate of risk is. For example, the researchers may estimate that a 74-year-old man has a 10-year risk of stroke of 8%, based on an observational study of people in their 70s. However, if the man is known to have important risk factors for stroke, such as high blood pressure, diabetes, and smoking, then a more precise estimate will be higher if it is based on the incidence of a more narrowly defined population made up of individuals who share these same risk factors.

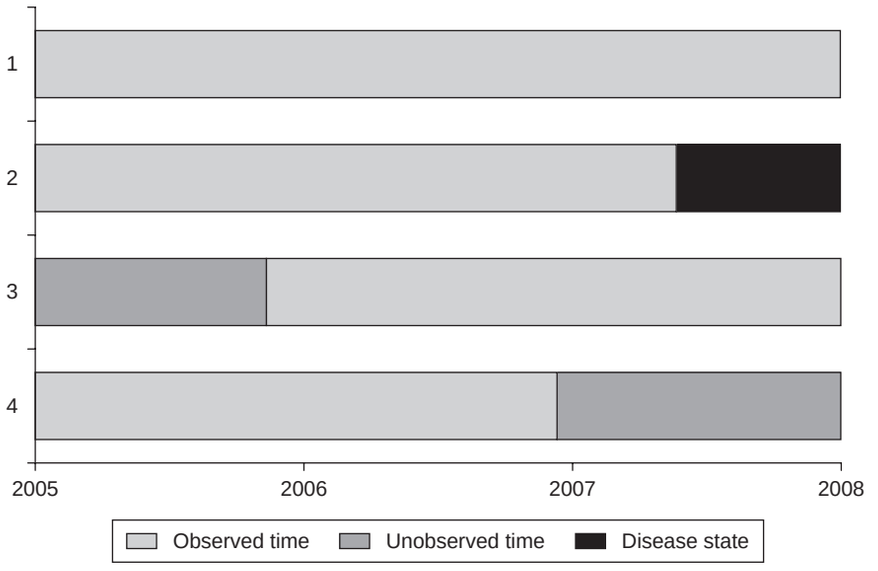


Figure 2 Incident rate

Note: Subjects 3 and 4 have been observed for only part of the time period. One new case per 10 total person-years; .1 case per person-year.

Assessing Disease Burden With Prevalence Measures

As discussed in the foregoing, cumulative incidence and incidence rate are measures of risk and deal with new cases. Measures of prevalence, on the other hand, reflect disease burden in a population and deal with existing cases. Rather than counting events, prevalence measures involve counting the number of individuals with a particular disease state.

Usually a prevalence measure reflects *point prevalence*, meaning the proportion of individuals within a population that has the disease state of interest at a particular time:

$$\text{Point prevalence} \quad \frac{\text{Number of individuals within a population with the disease state at a particular time}}{\text{Number of individuals in the population at that time}},$$

$$\frac{1200}{10000} = .12 \text{ or } 12\%,$$

or

$$.12 \times 1000 = 120 \text{ per } 1{,}000 \text{ persons.}$$

Figure 3 demonstrates point prevalence graphically.

The researchers can also calculate a *period prevalence*, in which the numerator is the total number of subjects who had the disease during a specified time span. In Figure 3, the 3-year period prevalence is 2/4 = .5.

The essential difference between incidence and prevalence is that in the latter researchers do not determine when a disease begins. Since they are not measuring new cases, they cannot use prevalence to indicate risk of developing the disease. An increasing prevalence in a population may be attributed to a longer duration of disease (due to, for example, improved survival or decreased cure rates).

Patterns of Occurrence

There are three important terms that are often used to describe the degree to which disease is present in a population. *Endemic* indicates that a disease is chronically present or a usual occurrence within a geographic area. For example, malaria is endemic in sub-Saharan Africa. An *epidemic* is said to exist when the existence of disease in a particular region is clearly above and

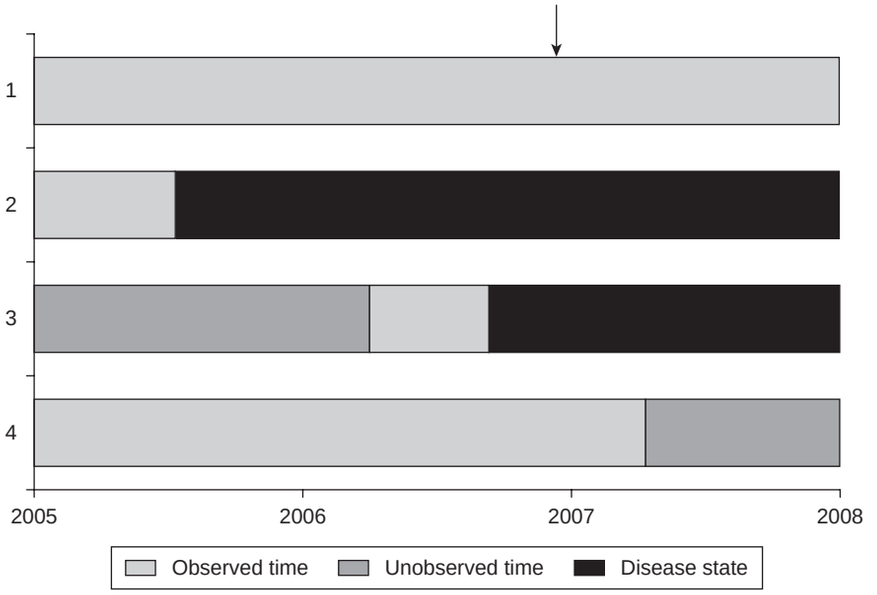


Figure 3 Point prevalence

Note: In January 2007, 2 of 4 subjects had the disease. Point prevalence is 5.

beyond what is expected from historical information. The large increase in the prevalence of obesity in the United States over the past 20 years is an example of an epidemic. Finally, *pandemic* is used to signify a worldwide epidemic. Influenza is a disease, for example, with the potential to become pandemic.

Competing Risks

Consideration of the concept of *competing risks* helps one understand changes in patterns of disease over time. A competing risk is present if there is an event other than the event of interest that may occur and change the probability of the event of interest. For example, consider a patient with diabetes. Such an individual is at risk for several important health events, but here two are considered: (1) development of end-stage renal disease (ESRD) and (2) cardiovascular death. Individuals with diabetes generally experience a slow decline in kidney function over a period of years. An important minority of such individuals eventually reach a severely low level of function whereby either dialysis or kidney transplantation is necessary to sustain life. Such individuals have ESRD. An individual cannot develop ESRD, however, if he dies from cardiovascular disease before reaching this severely low level of kidney function, which is in fact a common circumstance.

In recent decades, there has been a steady climb in both the incidence and prevalence of ESRD. Important in the understanding of this phenomenon is the fact that during this time period, there have been great advances in the treatments for cardiovascular disease. These advances have allowed many individuals who would otherwise have died to live on to experience ESRD. The incidence of cardiovascular death, an event that is competing with ESRD development, has decreased, allowing more cases of ESRD to develop.

Taking this example one step further gives some appreciation for the growing importance of competing risks. Individuals who do live on to develop ESRD and who are successfully treated will live, thereby having the opportunity to experience other significant events such as heart attack and stroke. Where a 60-year-old diabetic

individual in 1960 might have had one fatal heart attack, a similar 60-year-old individual in 2008 may survive the heart attack, live on to begin dialysis treatment for ESRD, and in the ensuing years experience a stroke and two more heart attacks before dying at age 80. As a general rule, as treatments improve, particularly for chronic diseases, the concept of competing risks will be increasingly important to our understanding of disease occurrence.

Potential Errors in Measuring Incidence and Prevalence

There are several potential sources of error that are important when trying to measure incidence and prevalence. In each of the following examples, the true incidence and prevalence may be overestimated.

Increase in the Frequency or Effectiveness of Screening

Screening is testing for disease in the absence of signs or symptoms of the disease and is central to preventive medicine. A positive screening test that is later confirmed will indicate disease and therefore identify a case. An increase in the number of known cases may be a function of better screening even if the true, unknowable prevalence is actually stable over time. For example, an increase in the apparent prevalence in depression in a population over time may be due to the fact that more physicians are looking for it through screening questions and questionnaires.

Enhanced Diagnosis

As technology advances, medical professionals are able to uncover disease states that in the past would remain occult. For example, an individual with chest pain and a negative treadmill stress EKG might in years past have been told that heart disease was not evident. Today, the physician has several methods, both invasive and noninvasive, to detect coronary disease accurately when it is in fact present.

Changes in Diagnostic Criteria

For example, in 2003 the American Diabetes Association lowered the cutoff value of impaired fasting glucose from 110 mg/dl to 100 mg/dl. The calculated prevalence of this disorder increased dramatically.

Enhanced Documentation

For studies based on data from medical records, for example, one must consider changes in documentation, perhaps attributable to conversion from paper to electronic records or in response to regulatory or financial demands. Such changes might account for apparent increases in disease occurrence.

The foundation of understanding disease in populations is simply counting new or existing cases. But one must use terms such as *incidence* and *prevalence* properly and understand the nuances that may introduce error.

Christopher Hebert

See also Frequency Estimation

Further Readings

Gordis, L. (2000). *Epidemiology* (2nd ed.). Philadelphia: W. B. Saunders.

Greenberg, R. S., Daniels, S. R., Flanders, W. D., Eley, J. W., & Boring, J. R. (2005). *Medical epidemiology* (4th ed.). New York: McGraw-Hill.

Rothman, J. R., & Greenland, S. (1998). *Modern epidemiology* (2nd ed.). Philadelphia: Lippincott-Raven.

Measures of Variability

Variability is the extent to which measurements differ from one another. Understanding the variability in a sample or population is important to evaluating whether an observed outcome is meaningful in a statistical analysis. Using the variability, researchers can identify whether a change in a measure is larger than what would be expected by chance. In addition, when reviewing data about a treatment or intervention, the average patient outcome may

be less important than the range of likely outcomes. Frequently reported measures of variability include variance, standard deviation, standard error, coefficient of variation, and interquartile range. Graphs and plots of data may be useful for illustrating variability and guiding statistical analyses.

Variance and Standard Deviation

Among the most common measures of variability is the variance, σ^2, which is a function of the differences between each data point and the average (mean) of the data. Larger variances indicate more variability (see Figure 1, which shows the difference in variability for two groups with identical means when $\sigma^2 = 1$ and $\sigma^2 = 9$). The variance is always greater than or equal to 0: If all the values are identical, $\sigma^2 = 0$. While the variance is affected if each measurement in the data is multiplied by the same number (change in the scale), it is not changed if the same number is added to each measurement (shift in location). Another useful property of the variance is that the variance of a sum of uncorrelated measures is equal to the sum of their variances.

The variance of a population may be estimated from a sample of size n using the unbiased estimator $s^2 = 1/(n - 1) \sum_{i=1}^{n} (X_i - \bar{X})^2$, where $X_1, \ldots, X_n$ is a random sample and $\bar{X}$ is the sample mean. The biased version, σ^2 which replaces the denominator, $n - 1$, with n, is less commonly used. As an example, Table 1 gives the birth weights of 12 male and 12 female infants in kilograms. The mean birth weight for males in this sample is $\bar{X} = 3.48$, and the sample variance is $s^2 = .26$. For females, the sample mean is $\bar{X} = 3.30$, and the sample variance is $s^2 = .15$.

The standard deviation, σ or SD, is the square root of the variance and is often used with the average to describe the distribution of a measure. It is greater than or equal to 0 and is measured in the same units as the measure of interest, which is useful for interpretation. The sample standard deviation, s, the square root of the sample variance, s^2, is used in many formulas for confidence intervals and hypothesis testing. In Table 1, $s = .51$

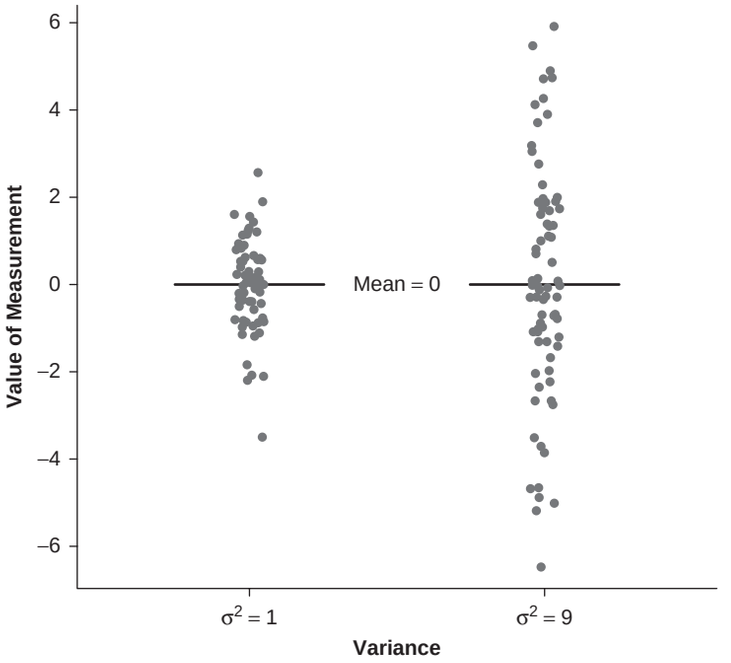


Figure 1 Comparison of different variability measures

Note: Scatterplot of random samples from normal distributions with mean 0 with variance equal to 1, and mean 0 with variance equal to 9.

Table 1 Birth weight (kg) of full-term male and female infants

Males		*Females*	
4.1	3.4	3.8	3.1
3.5	3.2	2.8	3.6
3.3	3.0	3.8	3.8
4.0	3.2	3.4	3.0
3.7	2.7	3.4	2.7
4.5	3.2	3.0	3.2

for males and .39 for females. When estimating the combined (pooled) standard deviation from more than one group or sample, the following formula is often used:

$$s_{\text{pooled}} = \sqrt{[((n_1 - 1)s_1^2 + (n_2 - 1)\ s_2^2 + \ldots + (n_k - 1)s_k^2)/(n_1 + n_2 + \cdots + n_k - k)]},$$

when k is the number of samples. The pooled sample standard deviation for males and females in Table 1, where $k = 2$, is $s_{\text{pooled}} = .46$.

Chebyshev's inequality illustrates the relationship between the standard deviation and the distribution of the data by indicating the proportion of the values that fall beyond a given number of standard deviations from the mean: For any distribution, the fraction of values at least d standard deviations away from the mean is less than or equal to $1/d^2$. In the case of the normal distribution, approximately 68.3%, 95.4%, and 99.7% of values are within one, two, and three standard deviations from the mean, respectively (see Figure 2).

Measures Based on the Sample Mean and Standard Deviation

The term *standard error* (*SE*) is sometimes used in place of the standard deviation, although, in general, the two terms are not interchangeable. The standard error is usually viewed as the standard deviation of an estimated quantity, for example the standard error of the sample mean (*SEM*). The SEM reflects the variability of the mean calculated from a sample of observations. The standard error for the mean is calculated by dividing the standard deviation for the sample by the square root of the sample size, n. The standard deviation is typically used to describe the variability of the data, while the standard error is used to describe how precisely the mean of the data was estimated and is used in the construction of confidence intervals. In Table 1, the SEM is .15 for boys and .11 for girls.

The coefficient of variation (CV) is another variability measure, commonly used in situations where variability is dependent on the size of the mean. The coefficient of variation is defined as the ratio of the standard deviation to the mean. This value is multiplied by 100 to reflect the measure as a percentage. In Table 1, the CV for males is 14.7%, and the CV for females is 11.8%, reflecting the larger variability among the males in the sample. A related measure is the effect size, or treatment effect, which provides a standardized measure of the difference between groups. The effect size is the difference in two group means divided by the standard deviation from either group, usually the control group, or by the pooled standard deviation.

Nonparametric Measures of Variability

The above variability measures, which are based on differences from the mean and are called parametric, are most meaningful when the data have a symmetric distribution such as the normal distribution's bell-shaped curve. However, when the distribution is not symmetric (skewed), differences from the mean hold less value because it is not the center of the distribution (see Figure 3 for examples of symmetric and nonsymmetric distributions). In these cases, the variability can be described using the percentiles of the observed sample, which are considered to be nonparametric and are valid for any distribution of data.

Two common nonparametric measures of variability include the range, defined as the minimum and maximum observed values, and the interquartile range, which is the first quartile (25th percentile or Q1) and third quartile (75th percentile or Q3). While the range provides the extremes of the distribution, the minimum and maximum are greatly affected by outlying observations. The interquartile range provides a more stable (robust) measure of variability because it is less influenced by outlying measurements. In Table 1, the interquartile

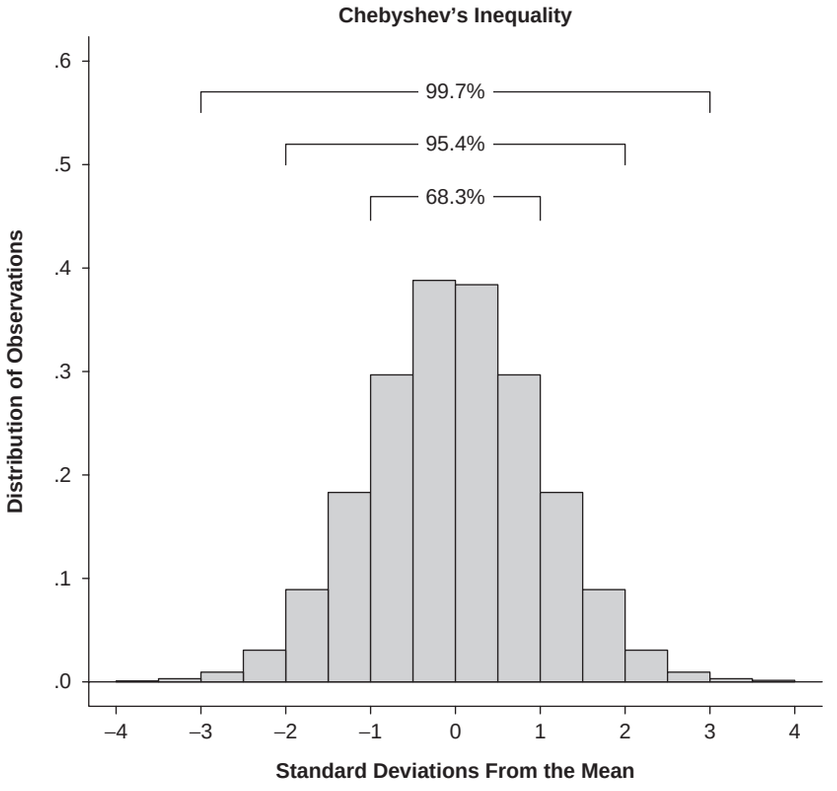


Figure 2 Chebyshev's Inequality

Note: The percentages of data that fall within a given number of standard deviations from the mean are shown. The histogram is a sample from a normal distribution with mean 0 and standard deviation 1.

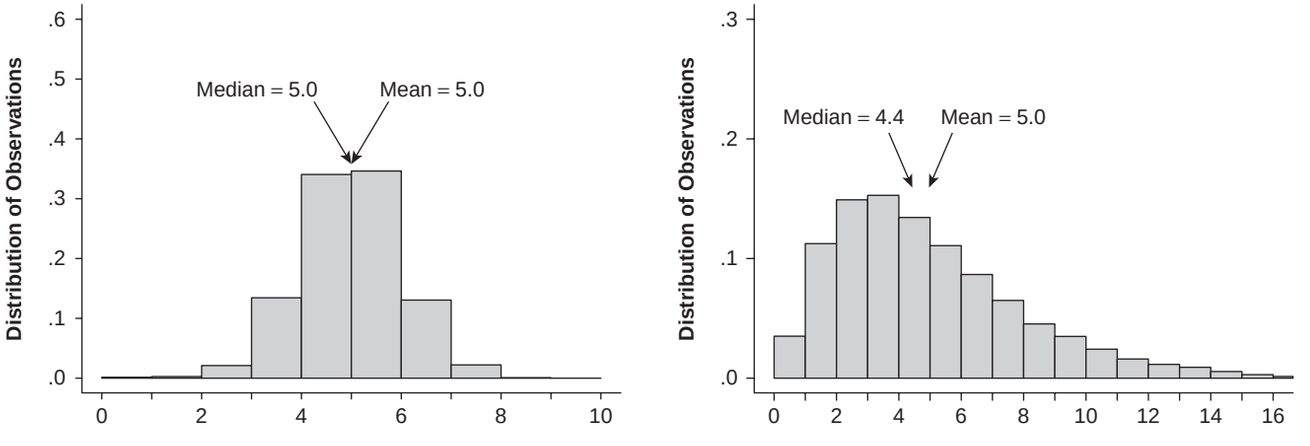


Figure 3 Symmetric and nonsymmetric distributions

Note: The symmetric plot shows data drawn from a normal distribution with mean 0 and standard deviation 1. The nonsymmetric distribution is a sample from a chi-square distribution with 5 degrees of freedom.

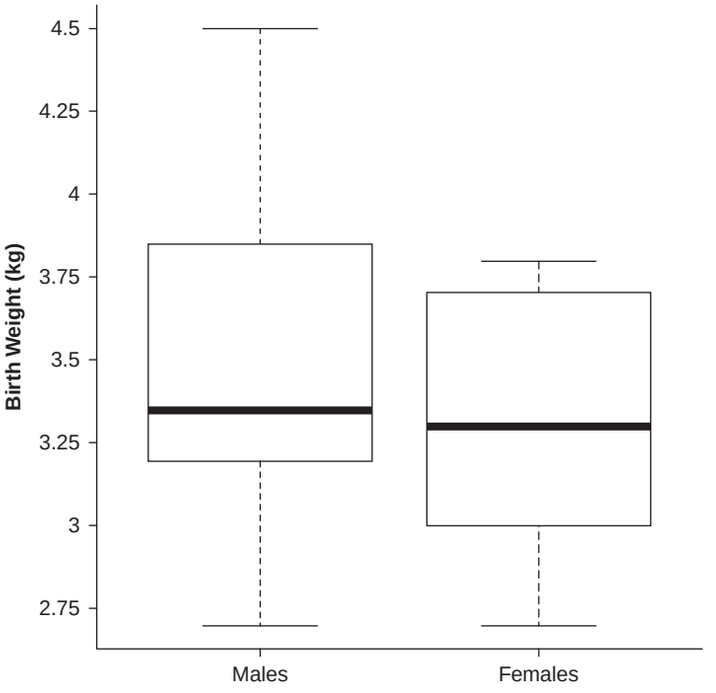


Figure 4 Boxplot of the birth weight data from Table 1

Note: The lower and upper lines of the box indicate the 25th and 75th percentiles of the data, and the center line in the box indicates the median. The whiskers extend to the minimum and maximum data values.

ranges for males and females are very similar (3.20–3.78 for males and 3.00–3.65 for females), while the ranges reflect the greater variability and greater maximum measurement in the males (2.70–4.50 for males and 2.70–3.80 for females).

Graphical Assessment of Variability

To assess the variability observed in a given sample, graphical methods are often very useful. Scatterplots (Figure 1) and histograms (Figures 2 and 3) can provide a general shape of the observed data distribution as well as identify the mean or median. Boxplots that display the quartiles of the data are also beneficial because they can reveal information about the data, including whether the data are not symmetric or if outlying observations exist. Boxplots stratified by treatment group or other factors can be used to assess empirically whether assumptions of statistical models hold. In the boxplots in Figure 4, male subjects have more variability in birth weight than do females.

Statistical analyses often begin with these graphs to uncover patterns that may be present in the data. Using these pictures along with the numerical summaries described above, researchers can understand the variability of their data and therefore understand their measurements better.

James F. Bena and Sarah E. Worley

See also Analysis of Variance (ANOVA); Confidence Intervals; Hypothesis Testing; Variance and Covariance

Further Readings

Casella, G., & Berger, R. L. (2001). *Statistical inference* (2nd ed.). Belmont, CA: Duxbury.

Rosner, B. (2005). *Fundamentals of biostatistics* (6th ed.). Belmont, CA: Duxbury.

Medicaid

Medicaid is a U.S. governmental program that is funded by federal and state (and, in some cases,

local) governments, is administered by the states under federal guidance, and is intended to cover the costs of medical and other healthcare-related services for the poorest of America's citizens. It was enacted in 1965, exemplifying the government's concern about access to medical care for two large segments of the U.S. population—the elderly and the poor. Buried within legislation enacting Medicare (as Title XVIII of the Social Security Act), through which federally supported health insurance was extended to the country's elderly, Medicaid (as Title XIX of the Social Security Act) was enacted as part of the continuing efforts of the federal and state governments' limited efforts to fund medical care for various categories of needy people. Originally enacted as an optional program (i.e., there was no mandate by the federal government requiring that individual states implement Medicaid), it was not until Arizona implemented its Medicaid program in 1982 that all states (and the District of Columbia and other U.S. territories) provided Medicaid coverage to eligible individuals. As of 2006, 43 million low-income individuals were recipients of Medicaid coverage.

Medicaid, like other public and private health insurance programs in the United States, plays an important role in U.S. health policy. The policy decisions inherent in the successful administration of the program can be informed by appropriate application of decision-making tools such as cost-effectiveness analysis and other statistical techniques, especially in the face of limiting resources. This entry provides the reader with an overview of the Medicaid program, including the federal and state governments' role in its administration; the eligibility requirements for receiving benefits and what benefits recipients are eligible to receive; the costs of Medicaid as part of the nation's health expenditures; and the expansion of the Medicaid program through the State Child Health Insurance Program (SCHIP). Last, this entry considers the effect of Medicaid on states' budgets and states' attempts to control the program's expenditures, including the health reforms introduced by the state of Oregon, the centerpiece of which was a prioritized list of services that ranked medical conditions and treatment on the basis of cost-utility analysis and that sparked a national debate on rationing.

Program Administration and Financing

Each state administers its own Medicaid program, following broad requirements and guidance from the federal Department of Health and Human Services' Centers for Medicare and Medicaid Services (CMS) and within which states have considerable discretion concerning which groups of low-income people are eligible for coverage, which benefits will be covered, and what mechanisms will be used to reimburse service providers. As a result, Medicaid programs vary considerably from state to state, with some programs providing very comprehensive coverage for large numbers of eligible individuals, while others cover more limited and basic sets of services.

Medicaid is supported by federal and state funds; some states require local government support as well, offsetting some portion of the state government's share of the funding. States participating in Medicaid receive federal matching funds on a percentage basis that can range from 50% to as high as 85% of their medical service expenditures, depending on the state's per capita income. (Most state administrative costs are matched at 50% for all states.) In this way, high-income states such as Connecticut, Maryland, Massachusetts, and New York, all of which have federal medical assistance percentages (FMAPs) of 50%, are required to spend more of their own funds to cover Medicaid recipients than is required of low-income states such as Arkansas, the District of Columbia, Mississippi, and West Virginia, all of which have FMAPs exceeding 70%. The average federal match is about 57% of the cost of the programs; the federal government spends $1.14 for every $2 spent by the states. However, despite the higher matching percentages for states with low per capita incomes, average per capita federal spending is higher in high-income states such as New York than in low-income states such as Alabama due to state-level differences in Medicaid policy choices, healthcare costs, and population demographics.

Eligibility

Generally speaking, Medicaid is a means-tested program, and individuals need to meet certain eligibility requirements to qualify for coverage. While some eligibility requirements vary from

state to state, states are required to cover certain populations to receive federal matching funds. These populations can be grouped into five broad eligibility requirements—categorical, income, resource, immigration status, and residency—all of which must be met by an individual to qualify for Medicaid coverage.

Categorical

Federal statute has outlined more than two dozen eligibility categories, which can be broadly grouped as children, pregnant women, adults in families with dependent children, individuals with disabilities, and the elderly. Individuals falling within these categories are considered "categorically needy."

Income

Income standards are tied to a dollar amount that is a specified percentage of the federal poverty level. This, as well as the methodology by which a person's income is calculated to determine his or her eligibility, can vary both from state to state and also by eligibility category—what counts as income, what income can be disregarded from the calculation, and the amount in healthcare costs a person needs to incur to reduce or "spend down" his or her income to meet the income-based eligibility requirement. (The categorically needy, however, do not need to spend down to qualify for Medicaid.)

Resource

In most states, individuals in most eligibility categories must have resources (e.g., savings accounts, personal property such as an automobile [above a specified value], and real estate other than one's home) that have a total value less than some specified amount to qualify for Medicaid.

Immigration Status

Citizens who meet their state Medicaid program's financial and other nonfinancial eligibility requirements are entitled to Medicaid coverage. Immigrants who have entered the United States illegally but meet other eligibility requirements to qualify for Medicaid are allowed to receive only emergency medical care. Immigrants legally residing in the United States may be eligible for the full range of Medicaid services, depending on the date on which they entered the country.

Residency

In addition to being a citizen of the United States or a legal immigrant who entered the country prior to August 22, 1996, an individual must be a resident of the state offering the Medicaid coverage for which the individual is applying.

There are other pathways to becoming eligible for Medicaid. For example, an optional Medicaid eligibility group is the "medically needy"—individuals who have incurred high medical expenses and who meet Medicaid's categorical requirements (as noted above) but whose income is too high to qualify for coverage. These individuals can qualify for eligibility by spending down their income by paying for medical expenses. However, for these and all other Medicaid recipients, eligibility for Medicaid coverage is not indefinite. A recipient's eligibility must be reviewed and redetermined at least once every 12 months. A person who no longer meets the eligibility requirements of his or her state loses his or her entitlement to Medicaid coverage.

Services Covered

To receive federal matching funds, each state's program is required to provide certain specific health services in its Medicaid benefit package. These mandatory services include the following:

- Inpatient and outpatient hospital care
- Physician, midwife, and certified nurse practitioner services
- Laboratory and X-ray services
- Nursing home and home healthcare for individuals of age 21 and older
- Early and periodic screening, diagnosis, and treatment services for children below age 21
- Family planning services and supplies
- Ambulatory care provided by entities designated as federally qualified health centers (outpatient care facilities that receive federal grants as community health centers) or as Medicare-certified rural health clinics

In addition, states may offer optional services, and for these too they will receive federal matching funds. Such services include prescription drugs, physical therapy, hospice care, vision care, including glasses, and dental care. In addition, states may apply to the federal government for waivers to continue receiving matching funds for services for which matching is not otherwise available. For example, states have received waivers to offer special home- and community-based services such as case management, personal care, and home health aid services to recipients at risk of being institutionalized in nursing facilities. Section 1115 waivers permit the continuation of matching funds for state-level demonstration projects implementing different approaches to administering Medicaid, under the condition that expenditures should not exceed the amount that would be spent in the absence of the waiver. One such waiver program, the Oregon Health Plan, will be described in more detail below.

State Children's Health Insurance Program

During the late 1990s, 40 million Americans, or 14% of the U.S. population, lacked health insurance. In 1997, SCHIP was enacted (as Title XX of the Social Security Act) as a way to expand health insurance coverage for uninsured low-income children who were not eligible for Medicaid. (These were children living in families with incomes at or below twice the federal poverty level or 50% higher than the state's Medicaid eligibility level, whichever was higher.) Between October 1997 and September 1998, the first year of SCHIP's implementation, more than 660,000 children had been enrolled at some point during that 12-month period; through 2007, more than 7.1 million children had been enrolled in all. However, about 20% of the 13 million children in poverty remain uninsured.

As with Medicaid, SCHIP is administered at the state level but with guidance from the federal CMS, within which each state determines the design of its program, eligibility groups, benefit packages, payment levels for coverage (including cost-sharing arrangements), and administrative and operating procedures. However, unlike Medicaid, which is a federal legal entitlement to states and for which no specific level of funding is appropriated in advance, the federal government appropriated $24 billion in matching funds to the SCHIP program as grants-in-aid for the program's initial 10-year period (1998–2007). States participating in SCHIP under its Section 1115 waiver meet that authority's budget-neutrality standard by not exceeding the annual SCHIP allocation.

Like the Medicaid program, SCHIP is jointly financed by the federal and state governments, with the federal government providing funds to states on a matching basis. An individual state's share of the federal funding appropriation was calculated on the basis of a formula that blended estimates of the number of low-income children and the number of uninsured low-income children in the state, adjusted by a health cost factor to account for differences in healthcare costs between the states. The SCHIP matching rate ranges from 65% to 85%, somewhat higher than the Medicaid matching rate of 50% to 85%, and as such was intended to provide an incentive for states with the greatest need to invest state resources in new SCHIP programs. However, poorer states received lower SCHIP enhancements (ranging from 6% to 8%) than did richer states with their lower Medicaid match rates; states with the lowest Medicaid match rate received SCHIP match rates of 65%, 15% higher than their 50% Medicaid match rate.

In addition, states had the option of implementing SCHIP as a discrete program, an expansion to their existing Medicaid program, or a combination program. As of 2008, 18 states have implemented a separate SCHIP program, while 9 states (including the District of Columbia) and 24 states have implemented SCHIP as Medicaid expansions and as combined SCHIP-Medicaid programs, respectively.

SCHIP was authorized for an initial 10-year period, which was to expire in 2007. Efforts to reauthorize the program during 2007 failed, but the program was given an 18-month extension through March 2009, although there is concern that despite the extension and the funding it provides, additional growth in the availability of government health insurance coverage will be necessary to meet the growth in the numbers of uninsured children, whose numbers have increased by 1 million between 2006 and 2008.

Medicaid Spending and Enrollment

In 2006, U.S. expenditures for health services and supplies (exclusive of certain research and

infrastructure investment) were almost $1.97 trillion or $6,561 per person. Medicaid expenditures, including those for the SCHIP expansion, amounted to about $320.2 billion (16% of the nation's expenditures for health services and supplies), or $5,164 per recipient. Almost 95% of Medicaid expenditures fund medical services for Medicaid program recipients, with the remaining 5% supporting program administration. Of medical expenditures for recipients, about 59% funds acute care services (inpatient hospital, physician, laboratory, outpatient, and drugs), and almost 36% funds long-term care services (home health, nursing home and other residential facilities, and mental health).

A breakdown of Medicaid expenditures by type of recipient shows that while children represent the largest group of Medicaid recipients, the elderly and disabled account for a larger proportion of expenditures. In 2005, children constituted 50% of Medicaid recipients and accounted for only 17% of Medicaid expenditures, costing on average about $1,700 per child. Ten percent of recipients are elderly and 14% of recipients are disabled, but these groups account for 26% and 41% of Medicaid expenditures, respectively. Medicaid expenditures are about $12,000 per elderly recipient and about $14,000 per recipient who is disabled. About 17% of Medicare beneficiaries also qualify for enrollment in Medicaid ("dual eligibles"), and these individuals, making up only 14% of Medicaid enrollees, account for 40% of Medicaid expenditures.

Medicaid experienced spending growth each year since its inception, until 2006, when the program experienced its first drop in spending (a decrease of 0.9%) even as national health expenditures rose 6.7% from 2005. However, this decrease was primarily due to the implementation of the Medicare Part D benefit, which provides prescription drug coverage to Medicare beneficiaries and which relieved the Medicaid program of the costs for drugs for Medicare beneficiaries who were also eligible for Medicaid coverage. Aside from this decline in drug spending, however, other personal healthcare Medicaid expenditures increased by 5.6%, a smaller increase than in previous years and reflecting the combined effect of weaker growth in enrollment and states' implementation of cost containment initiatives. Medicaid's expenditure growth has created pressures on state and local government budgets for several years, with Medicaid accounting for an increasing proportion of state and local spending. In 2006, Medicaid expenditures accounted for 21.5% of state spending. States implemented initiatives to control costs, such as managed care programs, alternative programs for long-term care, and more restrictive eligibility criteria.

An innovative initiative was implemented by the state of Oregon. The legislature of that state, in an effort to both control the cost of the state's Medicaid program and expand access to health insurance in the state, passed legislation in 1989 and 1991 to extend eligibility to the state's Medicaid program to all legal residents of the state with incomes up to the federal poverty level; expanded the use of prepaid managed care for the Medicaid population; and created a high-risk pool, with state-subsidized premiums, for residents who had been denied coverage due to preexisting conditions. In 1991, the state also applied to the federal government for approval to implement the plan, eventually receiving approval in 1993 as a Section 1115 demonstration project.

The centerpiece of the plan was a prioritized list of health and medical services that the state's Medicaid-eligible population would be entitled to receive. In addition, the extent to which these services would be covered by the state's Medicaid program depended on the amount of funding available within the state's budget to cover the cost of services. This list was intended to be a basic healthcare package that was more limited than what was covered under the state's traditional Medicaid program. Clinical information on thousands of conditions and their associated treatments, including the effectiveness of treatments; claims data to provide information on the cost of services; and the results of public forums, focus groups, and a telephone survey, all of which were intended to gauge the public's preferences and values concerning healthcare, were all applied to deriving the cost-utility of pairs of conditions and their associated treatments. Once this prioritized list of "condition/treatment pairs" was derived, a line would be drawn at the point on the list that the state's Medicaid budget could support; condition/treatment pairs falling above the line would be covered, while condition/treatment pairs falling

below the line would not. This initial list, however, was widely criticized because the resulting rank ordering put some procedures (such as treatment for thumb sucking and acute headaches) higher on the list than others (such as treatment for AIDS or cystic fibrosis). Not only was this initial list withdrawn from the demonstration waiver application to the federal government, but the state also abandoned the use of cost-utility analysis to derive the prioritized list and applied other algorithms to construct the list.

Oregon's list is revised every 2 years as part of the state's biennial budget process. Considerations for revision include improved outcomes information, medical advancements, and even the inclusion of services related to physician-assisted suicide, which became legal in Oregon in 1997. In 2006, the list underwent a complete reprioritization to incorporate a revised methodology, one that placed greater emphasis on preventive services, chronic disease management, and the effect of treatment on the health of the individual and the population. However, recent years' benefit reductions, the introduction of a reduced benefit package for persons with somewhat higher incomes that also included some cost-sharing provisions for certain populations, and an economic downturn in the state that resulted in increased unemployment and decreased tax revenues and eroding political support for the plan have all combined to challenge the sustainability of the program.

Franklin N. Laufer

See also Cost-Effectiveness Analysis; Cost-Utility Analysis; Decisions Faced by Nongovernmental Payers of Healthcare: Managed Care; Economics, Health Economics; Government Perspective, General Healthcare; Government Perspective, Informed Policy Choice; Medicare

Further Readings

Catlin, A., Cowen C., et al. (2008). National health spending in 2006: A year of change for prescription drugs. *Health Affairs, 27,* 14–29.

DiPrete, B., & Coffman, D. (2007). *A brief history of health services prioritization in Oregon*. Retrieved July 6, 2008, from http://www.Oregon.gov/OHPPR/HSC/docs/PrioritizationHistory.pdf

Engel, J. (2006). *Poor people's medicine: Medicaid and American charity care since 1965*. Durham, NC: Duke University.

Holahan, J., & Weil, A. (2007). Toward real Medicaid reform. *Health Affairs, 26*, w254–w270. Retrieved July 1, 2008, from http://content.healthaffairs.org/cgi/reprint/26/2/w254

Kaiser Commission on Medicaid and the Uninsured. (2002). *The Medicaid resource book* (Pub. #2236). Retrieved June 25, 2008, from http://www.kff.org/Medicaid/2236-index.cfm

Kaiser Commission on Medicaid and the Uninsured. (2007). *Health coverage and children: The role of Medicaid and SCHIP* (Pub. #7698). Retrieved June 25, 2008, from http://www.kff.org/Medicaid/upload/7698.pdf

Kaiser Commission on Medicaid and the Uninsured. (2007). *Medicaid enrollment and spending trends* (Pub. #7523-02). Retrieved June 25, 2008, from http://www.kff.org/Medicaid/upload/7523-02.pdf

Kaiser Commission on Medicaid and the Uninsured. (2007). *The Medicaid program at a glance* (Pub. #7235-02). Retrieved June 25, 2008, from http://www.kff.org/Medicaid/upload/7235-02.pdf

National Association of State Budget Officers. (2008). *The fiscal survey of states*. Retrieved July 7, 2008, from http://www.nasbo.org/publications.php#fss2007

Oberlander, J. (2006). Health reform interrupted: The unraveling of the Oregon Health Plan. *Health Affairs, 26*, w96–w105. Retrieved July 1, 2008, from http://content.healthaffairs.org/cgi/reprint/26/1/w96

Oregon Department of Human Services. (2006). *Oregon Health Plan: An historical overview*. Retrieved July 6, 2008, from http://www.Oregon.gov/DHS/healthplan/data_pubs/ohpoverview0706.pdf

Tengs, T. O. (1996). An evaluation of Oregon's Medicaid rationing algorithm. *Health Economics, 5,* 171–181.

Medical Decisions and Ethics in the Military Context

In either a military or civilian context, medical decisions can be divided into three general categories: (1) the decision whether to treat, (2) the decision when to treat, and (3) the decision how to treat. This entry looks at how the difference in the character and practice of military medicine influences these decisions. For medical decision making in a military context, patients are generally

divided into two broad categories: (1) military personnel (of one's own nation) and (2) everyone else. This second category includes allied military personnel, prisoners of war, and civilians. This distinction becomes important because there are different considerations governing the decision-making process regarding the treatment of these two groups. For example, international law requires that prisoners of war receive medical care and that decisions regarding their care be based solely on their medical condition. Because of the sheer breadth of issues that relate to the decision-making process in military medicine, this entry is restricted to the first category—members of one's own military establishment.

Among the various militaries, there may be some minor differences in the way in which each service practices medicine. For the sake of brevity and consistency, this entry focuses on the practices of the U.S. Army Medical Department (AMEDD). Even after restricting this to practices of the AMEDD, the sheer breadth of issues relating to medical decisions in a military context precludes discussing more than a few key points. These include the goals of military medicine, the decision whether to treat military personnel, the decision when to treat military personnel, and the decision how to treat military personnel.

Goals of Military Medicine

Medical decisions are best understood in terms of the goals that the decisions are intended to achieve. In a civilian setting, the goals of healthcare primarily revolve around the interests of an individual patient or, in cases involving scarce resources, such as transplantable organs, around a relatively small set of patients. In civilian medicine, respecting the patients' autonomy—their right to make their own medical decisions—governs most decisions. In such an environment, the goals of healthcare include the medical interest of the patient, respecting the wishes and values of the patient, and in rare cases a just allocation of scarce medical resources.

To understand military medical decisions, one must understand the goals of military medicine. Because of the integrated role military medicine plays in the planning and execution of military operations, its goals are inexorably tied to those of the military in general. This is evident in the AMEDD's mission statement: "Preserve the Fighting Force." The goals of the military can be understood at a variety of levels, but for the purposes of this discussion, there are two goals that are particularly relevant. The first is the successful completion of military operations assigned to it by the civilian authority, and the second is to protect the lives and health of the military personnel. The military has a contractual and moral obligation to protect to the extent possible the lives of its soldiers. Every military operation needs to balance the risks to personnel against the value of achieving the military objective. "Needs of the military" will be understood to include both the successful achievement of military objectives and protecting to the extent possible the lives of military personnel.

The Decision Whether to Treat

There are two major concerns when looking at the medical treatment of military personnel. The first is that soldiers may be treated without their consent. Such treatment may take the form of failing to obtain consent prior to treatment or treating soldiers in opposition to their stated wishes. Second, because the medical treatment of soldiers can be viewed by some as being subservient to the needs of the military, some soldiers may be perceived as being denied or receiving substandard treatment. While both concerns can manifest themselves in all three categories of decision making (whether to treat, when to treat, and how to treat), this section focuses on the first of these concerns: that a solider may be treated without his or her consent. The second concern will be addressed later in this entry.

Military medicine has an obligation to treat its own soldiers. This is true regardless of whether the soldier can or cannot be returned to active service. Military healthcare professionals have an obligation to the patient before them, but because of the nature of the service, they must balance this obligation with their obligation to the needs of the military. The interests of any individual patient must be weighed against what is needed for successfully completing military operations and protecting the lives and health of other soldiers. As a consequence, military members' autonomy may be overridden, and they may be treated against their

wishes. United States Code, Title 10, Subtitle B, Part II, Chapter 355, Section 3723 allows the Secretary of the Army to "order the hospitalization, medical and surgical treatment, and domiciliary care, for as long as necessary, of any member of the Army on active duty." An example of this occurred during the first Gulf War when U.S. military members were required to be inoculated with pyridostigmine bromide as a protective measure against nerve agent exposure.

In the case of battlefield medicine, soldiers are seldom given the opportunity to provide informed consent or refusal for treatment. This is due in part to the traumatic nature of battlefield injuries. Life preserving treatment must be administered immediately, and patients, because of their traumatic injuries, often lack the ability to make such decisions. In this, military medicine mirrors the practices of civilian emergency room procedures. The second factor that makes respecting a soldier's right to refuse treatment in a combat zone difficult is the fact that the needs of the military require that those soldiers who can be returned to duty do so as soon as possible. Military concerns prohibit allowing soldiers to avoid continued service by refusing treatment that is necessary to perform their assigned duties. It should be noted that avoiding service is not the only reason why a soldier might refuse treatment. Many combat soldiers feel that being evacuated out of the theater of operations because of their medical condition constitutes "abandoning" their buddies and, given the choice, would prefer to remain with their units despite their injuries. Military medicine has an obligation to ensure that military personnel are medically fit to perform their assigned duties. This is for their protection as well as the protection of those with whom they serve. Such decisions to refuse treatment cannot be respected.

The fact that members of the military may have their autonomy overridden by itself does not mean that doing so must be the norm in a military setting. The needs of the military as they relate to the healthcare of soldiers should be understood in terms of a spectrum. In combat situations, medical choices are often limited and the needs of military operations immediate. Most military medical treatment does not occur within the context of combat operations, however. In a noncombat setting, military medicine attempts to mirror its respect for a patient's autonomy, to the extent possible, with that of its civilian counterpart. The need of a standing army to have its soldiers medically fit at all times requires that soldiers receive periodic medical examinations and that any medical condition that would impair their ability to perform their assigned duties be treated. It also requires that at times potential medical problems be prophylactically treated. This was the justification for requiring soldiers to take pyridostigmine bromide and, currently, anthrax vaccination. There are many medical conditions that do not affect a soldier's ability to perform his or her assigned duty. While military medical personnel have an obligation to offer treatment for any medical condition as healthcare professionals, they are not required, as agents of the military, to require compliance with any treatment that does not impair a soldier's ability to perform his or her duties. In such cases, the patient should be allowed to make decisions regarding treatment based on his or her own values and wishes.

In a noncombat setting, the military is often willing to allow soldiers to refuse treatment that is necessary for their continued service. In such situations, the soldier is informed that a refusal to accept treatment will result in separation from the military. While this allows greater autonomy in military medicine, it also proves to be a source of coercion that can interfere with genuine informed consent. The threat of separation for a soldier who is heavily invested in his or her military career may cause him or her to feel forced to accept the treatment. It should be noted that the need for military personnel to be ready to fight is critical to the military being able to accomplish its goals. As a result, refusal to accept necessary medical treatment cannot be allowed to be used as a mechanism to avoid service. The decision of a soldier to refuse necessary treatment should only be accepted if there is reason to believe that the soldier's motive is to refuse unwanted medical interventions and not to avoid participating in some military operation.

The Decision When to Treat

In nonbattlefield medicine/nonemergency cases, there is some flexibility when a soldier is treated. Soldiers as a rule should have some input into

when they receive treatment, keeping in mind the fact that soldiers may need to be mobilized at any time with little or no notice. There are going to be exceptions to this rule, though. There are certain critical career fields in the military that must be protected. These career fields reach critical status because they are essential to the military's ability to achieve its mission and/or there are a limited number of soldiers qualified to perform those duties. In such cases, it is particularly important that these solders be medically fit at all times.

There is a concern that some soldiers may be denied care or receive substandard care because of the needs of the military. If this is an issue, it presents itself most prominently in the context of battlefield medicine. In a nonbattlefield setting, soldiers are rarely denied treatment. The treatment of certain minor medical conditions that do not impair a soldier's ability to perform his or her task and present no long-term risk to the soldier may be delayed until a soldier has completed his or her current duty assignment if treatment is not available at the soldier's current location but will be provided at some point. Even if a soldier is deemed unfit for continued military service, any medical condition that is service related or discovered while the soldier is on active duty will be treated when the soldier is transitioned into the care of the Veterans Affairs medical system.

Ideally, the decision of when to treat a patient should be based solely on the medical needs of that patient. This is true for both civilian and military medicine. When medical resources are not limited and when there are no overriding military interests, every soldier should be treated according to his or her need. Unfortunately, even in a civilian setting, this is not always the case. There are times when the patient load overwhelms available medical resources and patients have to be triaged not only in terms of their individual medical conditions but also with the goal of saving as many lives as possible as a governing principle. In battlefield medicine, when there are no overriding military interests, this principle of saving as many lives as possible should also govern decisions regarding when to treat patients. As discussed, though, military medicine is obligated to take into account the needs of the military. To aid in this decision-making process, the U.S. Army has established six Medical Battlefield Rules in *Field Manual 8-55: Planning for Health Service Support*: (1) maintain medical presence with the soldier, (2) maintain the health of the command, (3) save lives, (4) clear the battlefield, (5) provide state-of-the-art care, and (6) return soldiers to duty as early as possible. These are listed in order of priority and indicate which should govern the decision-making process if all six cannot be accomplished. Note that saving lives is third on the list of priorities. So there are two considerations that have to be taken into account that can supersede this principle of saving as many lives as possible. The first is the need to maintain medical readiness in the combat zone. The military has an obligation not only to those soldiers who are currently injured but also to those soldiers who will become injured. This means that medical units cannot exhaust their medical resources on their current patients at the expense of providing care to future patients. The second overriding interest is in protecting the medical readiness of the command. Unlike most soldiers on the battlefield, those in key command positions are not easily replaceable. Their continued role in combat operations is seen as being critical to the military's ability to achieve its overall objectives. The number of individuals who constitute this group is relatively small, and their treatment does not often affect military medicine practices. It is worth noting that returning soldiers to duty to sustain the fighting force is sixth on the list and does not override the governing principle of saving as many lives as possible.

The Decision How to Treat

Similarly, as with civilian medicine, in a nonbattlefield medicine/nonemergency setting there are often multiple forms of treatment that can accomplish the same medical objective. In such circumstances, the patient should be provided the opportunity to select his or her treatment. In battlefield medicine, when the medical injuries are more severe and medical resources, including the availability of medical personnel, are limited, such options are often impractical.

Once the decisions of whether and when to treat have been made, the decision of how to treat should be based on the standards of medical practice, the medical condition of the patient, and the availability of medical resources. The only potentially

relevant military interest that could affect the decision of how to treat a patient is the need to return soldiers to duty as soon as possible. But as previously indicated, this is considered the lowest priority in making battlefield medical decisions.

Jason Gatliff

See also Bioethics; Decisions Faced by Hospital Ethics Committees

Further Readings

Beam, T. E., & Sparacino, L. R. (Eds.). (2003). *Military medical ethics*. Bethesda, MD: Office of the Surgeon General, Department of the Army.

Gatliff, J. (2007). *Terrorism and just war tradition: Issues of compatibility*. Germany: VDM Verlag Dr. Müller.

Gross, M. L. (2004). Bioethics and armed conflict: Mapping the moral dimensions of medicine and war. *Hastings Center Report, 34*(6), 22–30.

Howe, E. G. (2003). Dilemmas in military medical ethics since 9/11. *Kennedy Institute of Ethics Journal, 13*(2), 175–188.

United States Code, Title 10. Armed Forces, Subtitle B. Army, Part II. Personnel (chap. 355). Hospitalization, Section 3723. Approved November 13, 1998.

U.S. Department of the Army. (1994). *Army field manual 8-55: Planning for health service support*. Washington, DC: Author.

Medical Errors and Errors in Healthcare Delivery

Medical error can be defined as any mistake in the delivery of care, by any healthcare professional, regardless of outcome. The specific reference to outcome is important because errors can result in actual adverse outcomes for patients or near misses. Adverse events are injuries that are caused by medical management rather than the underlying disease. They prolong hospitalization, produce a disability at discharge, or both. While the media tend to highlight the catastrophic injuries that result from medical error, there is often little understanding of the context of the clinical decision making in practice. Healthcare is an inherently uncertain and dynamic environment. Individual patients vary in their responses to treatment, and their health status can change rapidly with little warning. Clinical knowledge is frequently distributed among clinical team members, requiring both proactive and reactive decisions to be made, often under difficult circumstances such as limited resources and staff shortages. Medical error is omnipresent in healthcare, and the costs to the community are considerable. In this entry, the types of error that can occur in healthcare are defined, followed by an outline of the incidence of medical error and the common errors that may result from faulty decision making. The etiology of errors and the changes being implemented globally to address the problem are the focus of the final discussion.

Defining Medical Error

There is much debate in the medical literature surrounding the definitions of error, adverse events, mistakes, and near misses. However, the most commonly used definition of error is from the seminal report by the Institute of Medicine (IOM), *To Err Is Human: Building a Safer Health System*. As an exploration of medical error, *error* was defined in this publication as occurring when persons fail to complete an action as planned or intended (an act of omission), or they use an incorrect plan to achieve an aim (an act of commission). In doing something wrong or not doing something right, an undesirable outcome may or may not result.

If the resultant injury is caused by the medical care received by the patient rather than the patient's underlying illness, it is considered a preventable adverse event. It can result from a single error or an accumulation of errors. If the error results in serious harm or death to the patient, it is referred to as a *sentinel event*. Sentinel events usually require further investigation and may often reveal significant deficits in policies or current practice.

Some adverse events may be defined in legal terms as negligent adverse events. In these cases, a legal ruling is made as to whether an injury resulted because the care did not meet a standard of care reasonably expected to be delivered by an average clinician.

Sometimes an error does not result in a patient injury; it is then considered a *near miss* or *close call*. Near misses are potentially harmful incidents or errors that do not cause harm to patients either

because effective recovery action was taken or because no harm resulted from the error. For example, a nurse identifies an incorrect drug prescription prior to drug administration.

James Reason referred to errors as either active or latent. In healthcare, active errors are usually more apparent and result from a contact between a frontline clinician and some aspect of the system, such as the interface with machinery, for example, when a nurse programs a mechanical ventilator incorrectly. Active errors are mostly noticed by the person involved in committing the error—thus they are also referred to as *errors at the sharp end*.

In contrast, latent errors are more indistinct and result from system failures in the general operations of an organization that cause harm to patients. These blunt end errors, as described by Reason, often affect the person at the sharp end and are often referred to in retrospect as *accidents waiting to happen*. For example, an active error or failure would be the incorrect programming of an infusion pump, whereas a latent error or failure would be caused by an organization that has multiple types of infusion pumps, making a programming error by a clinician more likely because he or she may be confused.

Incidence of Medical and Healthcare Error

The incidence of adverse events is a significant avoidable cause of human suffering, with a high toll in financial loss and opportunity cost to health services. Since the early 1990s, research studies conducted in the United States of America, the United Kingdom, and Australia have found that between 4% and 17.7% of patients suffer from some kind of harm (including permanent disability and death) as a result of adverse events while in hospital. Andrews and colleagues found that more than one in six hospitalized patients suffered medical injuries that prolonged their hospital stays. Studies that have used a review of medical histories to identify medical injuries, such as the Harvard Medical Study in New York, determined that a significant proportion (up to 69%) of the medical injuries identified were due to error and, therefore, could have been prevented. The incidence of near misses is difficult to quantify because they are not always reported or recognized; however, they are thought to occur up to 100 times more frequently than do adverse events. Both adverse events and near misses are likely to share the same causal circumstances and hence be reflective of weaknesses in care delivery systems that cause risk for patients. Although estimates of the frequency of medical errors and injuries vary considerably, even the most conservative estimates indicate that the problem is widespread and requires serious attention. The magnitude of the problem has generally been taken to underscore the need for robust safety processes to be implemented to ensure optimal patient safety outcomes.

Governments, healthcare services, and health professional groups around the world are attending to the development and implementation of processes aimed at reducing the incidence and impact of preventable adverse events in healthcare and to generally improving the safety and quality of their healthcare services. Quality and safety initiatives take place within a complex health environment that includes hospital care, outpatient and ambulatory services, residential aged care services, and primary care. The provision of healthcare is affected by an increasing emphasis on accountability and public involvement in health. It is further challenged by an aging population, many of whom experience chronic disease and disability, as well as other burdens of disease that affect the broader population, such as obesity and associated cardiovascular disease and diabetes.

Types of Medical and Healthcare Error

There are many types of medical error that can occur within the multiple contexts in which healthcare is delivered and the increasing complexity of care. The most common types of error include medication errors, preventable nosocomial (hospital-acquired) infections, diagnostic and treatment errors, equipment errors, prevention errors, and unnecessary treatment. Interestingly, an Australian study by Weingart and colleagues showed that preventable cognitive errors, such as making incorrect diagnoses or choosing the wrong medication, were more likely to result in permanent disability than were technical errors such as choice of the wrong surgical technique.

Medication error is defined by the IOM as any preventable event that may cause or lead to

inappropriate medication use or patient harm while the medication is in the control of the healthcare professional, patient, or consumer. It may be related to professional practice, healthcare products, procedures, and systems, including prescribing; communication of orders; product labeling, packaging, and dispensing; distribution; administration; education; monitoring; and use. The IOM report cited one study finding that about 2% of hospital admissions experienced a preventable adverse drug event, although the majority were not fatal. The most common error involving medications is related to administration of an improper dose of medicine. This is followed by giving the wrong drug and using the wrong route of administration. Almost half of fatal medication errors occur in people over the age of 60. Older people may be at greatest risk for medication errors because they often take multiple prescription medications.

Infections acquired during a hospital stay are called nosocomial infections. They are defined as infections arising after 48 hours of hospital admission and can present as urinary tract infections, surgical site infections, respiratory infections (especially nosocomial pneumonia), blood infections/bacteremia, gastrointestinal tract infections, and central nervous system infections. Although these are not necessarily due to an identifiable error by healthcare professionals, greater emphasis on preventive measures such as hand washing and sterilization will prevent many nosocomial infections.

Diagnostic errors include wrong and missed diagnoses and failure to diagnose complications, underlying disease, or associated diseases and extend to failure to diagnose others in cases where family members or others exposed to disease are not investigated. Diagnostic errors can be due to pathology laboratory errors such as wrong biopsy results, mixing samples, or known test errors and risks associated with false positives and false negatives.

There are numerous ways that errors can occur in medical treatment. These errors can occur when patients use the wrong treatment, the wrong condition is treated or the wrong choice of treatment plan is made, or the wrong treatment is applied, such as can occur with medication errors or wrong blood transfusion. Treatment errors involving surgical mistakes include the following: wrong patient, wrong site, wrong organ, and equipment left inside. Anesthesia errors can include too much or too little anesthesia. Treatment errors include delays in applying appropriate treatment interventions and unnecessary medical treatment.

Prevention errors refer to failure to prevent conditions and include failure to prevent known complications of a diagnosed disease, failure to treat family members or others in cases of genetic diseases or infectious diseases, and failure to address risk factors for various conditions.

Adverse events can occur as a consequence of equipment failure or errors in the application of equipment such as occur with dislodgment of intravenous infusion devices or not renewing batteries in equipment.

Etiology of Medical and Healthcare Error

Traditionally, when an error occurred in medicine, the clinician delivering the care to the patient was blamed for the error. This view is called the person approach. However, a more recent approach has shifted toward systems thinking. Systems thinking acknowledges that human error in healthcare is a considerable threat to patient safety that can be traumatic to those involved and a challenge to learn from. To protect patients, organizations concentrate on the conditions that individuals work under and position internal safeguards to prevent errors occurring. Although this approach has been highly successful in numerous high-risk organizations, including the aviation, oil, and nuclear power industries, only recently has it been considered applicable in healthcare. The analysis of medical error has generally been underdeveloped and oversimplified.

In a systems approach, error is the result of a chain of events and many contributory factors. Psychologist James Reason's "Swiss cheese model" was drawn from a major review of organizational accidents. Reason found that single errors occurring at the sharp end of care were rarely enough to cause major harm to patients. In contrast, errors that resulted from multiple layers of structural weakness in organizations or through the alignment of holes in the Swiss cheese model were more likely to result in a disastrous effect. Thus, the aim of systems thinking is to strengthen the structure within an organization by developing layers of protection or safeguards to reduce latent errors.

As a means of reducing latent errors in medicine, healthcare organizations typically review the antecedent events of the error, using a root cause analysis approach. All errors are comprehensively dissected, usually by an interdisciplinary committee, who search for the triggers or root causes. Key facts and timelines are established after ascertaining the context of the situation the error occurred within from those involved. The analysis is used to identify system gaps and inadequacies. The facts are then presented to the root cause analysis (RCA) committee, who focus on what can be learned from the error to safeguard the system.

In adapting Reason's model for healthcare, Charles Vincent developed a framework of contributory factors that occur in clinical practice and produce conditions that result in error or latent failures. The framework categorizes the major influences on clinicians' decision making and associated underlying causes of error (see Table 1). It has been suggested that healthcare agencies use the framework to systematically analyze adverse outcomes and the antecedent events to prevent further adverse events.

The person approach to error focuses on the sharp end, where the interface between the frontline clinician and system occurs. This view regards the errors and violations as atypical cognitive processes, such as inattention, carelessness, and negligence, whereby people have the capacity to choose between safe and unsafe modes of behavior. Essentially humans have two behavioral control modes—conscious and automatic. The conscious mode is slow, prone to error, and sequential. It also has limited capacity and, if focused on one thing, cannot focus on another. Conversely, the automatic mode is mostly unconscious, fast, and able to multitask simultaneously. It is the mode that we use for the routine tasks that do not require high-level cognition.

Clinical tasks may range from routine everyday tasks that require little thought to novel problems that require deep thought and problem solving. Reason categorized human error as being either slips (skill-based errors) or mistakes (knowledge- or rule-based errors). Slips are defined as errors that occur as the result of inadvertent and unconscious behavior, which may occur when performing some automatic task. Slips tend to occur when individuals are on autopilot, with multiple sensory inputs occurring simultaneously. In contrast, mistakes are conscious errors that result from incorrect choices being made. Mistakes may be a consequence of inadequate knowledge or information, lack of experience or training, or the application of the wrong rule in the given situation.

It is believed that conscious behaviors are more likely to result in a mistake than automatic behaviors, which are likely to result in a slip. However, because a significant proportion of the activities that clinicians do are automatic in nature, it is thought that slips present a far greater risk to patient safety.

The person approach is focused on reducing variability in human behavior, creating a sense of fear, blame, and litigation. It is criticized for decontextualizing the incident, which ignores recurrent patterns of error in similar circumstances with different individuals. In contrast, the systems approach focuses on changing the work conditions to build organizational defenses. If an error occurs, the critical question is why it happened, not who did it. Hence, worldwide initiatives to prevent errors are focused toward changing the way healthcare is delivered to improve safety.

Future Directions

Internationally, within multiple governance and regulatory environments, there are a considerable number of approaches to improving the quality and safety of healthcare. The major themes in the various approaches include a strong focus on the issues related to governance and leadership frameworks and educational strategies, attention to organizational climate or culture, and workforce issues such as work hours and conditions. An important development is the emerging body of opinion that, in addition to healthcare providers, patients and their families have an important role to play in monitoring and improving patient safety. Underpinning this view is a growing appreciation of the unique relationship that patients and families have with each other and of the benefit of their continuous vigilance over both the patient's health condition and the care that is given. It is also being increasingly recognized that patients often become well-versed in their own illnesses and care during the trajectory of their healthcare experience and can play a significant

Table 1 Framework of factors influencing clinical practice and contributing to adverse events

Framework	*Contributory Factors*	*Examples of Problems That Contribute to Errors*
Institutional	Regulatory context Medicolegal environment	Insufficient priority given by regulators to safety issues; legal pressures against open discussion, preventing the opportunity to learn from adverse events
Organization and management	Financial resources and constraints Policy standards and goals Safety culture and priorities	Lack of awareness of safety issues on the part of senior management; policies leading to inadequate staffing levels
Work environment	Staffing levels and mix of skills Patterns in workload and shift Design, availability, and maintenance of equipment Administrative and managerial support	Heavy workloads, leading to fatigue; limited access to essential equipment; inadequate administrative support, leading to reduced time with patients
Team	Verbal communication Written communication Supervision and willingness to seek help Team leadership	Poor supervision of junior staff; poor communication among different professions; unwillingness of junior staff to seek assistance
Individual staff member	Knowledge and skills Motivation and attitude Physical and mental health	Lack of knowledge or experience; long-term fatigue and stress
Task	Availability and use of protocols Availability and accuracy of test results	Unavailability of test results or delay in obtaining them; lack of clear protocols and guidelines
Patient	Complexity and seriousness of condition Language and communication Personality and social factors	Distress; language barriers between patients and caregivers

Source: Reproduced with permission from Vincent, C., Taylor-Adams, S., Stanhope, N. (1998). Framework for analysing risk and safety in clinical medicine. *British Medical Journal, 316*, 1154–1157. Reproduced with permission from BMJ Publishing Group. License Number: 1953450752344 License Date: May 21, 2008; and Vincent, C. (2003). Understanding and responding to adverse events. *New England Journal of Medicine, 348*, 1051–1056. Copyright © 2003 Massachusetts Medical Society. All rights reserved. Reference Number: PS-2009-0025 Date June 3, 2008.

role in recognizing and rescuing errors and adverse events not otherwise detected by healthcare systems. This has led to changes in systems of care delivery that explicitly recognize the potential contribution that many patients and families can make to safety processes and how patients and families could inform clinical care and guide further improvement initiatives.

Tracey Bucknall and Mari Botti

See also Contextual Error; Error and Human Factors Analyses; Errors in Clinical Reasoning; Health Risk Management

Further Readings

Entwistle, V., Mello, M., & Brennan, T. (2005). Advising patients about patient safety: Current initiatives risk shifting responsibility. *Joint Commission Journal on Quality and Patient Safety, 31*(9), 483–494.

Kohn, L., Corrigan, J., & Donaldson, M. (Eds.). (2000). *To err is human: Building a safer health system*. Washington, DC: National Academy Press.

Leape, L. L., Brennan, T. A., Laird, N. M., Lawthers, A. G., Localio, A. R., Barnes, B. A., et al., (1991). The nature of adverse events and negligence in hospitalized patients: Results of the Harvard Medical Practice Study II. *New England Journal of Medicine, 324,* 377–384.

Reason, J. T. (1990). *Human error*. Cambridge, UK: Cambridge University Press.

Rosenthal, M. M., & Sutcliffe, K. M. (Eds.). (2002). *Medical error: What do we know? What do we do?* San Francisco: Jossey-Bass.

Vincent, C. (2003). Understanding and responding to adverse events. *New England Journal of Medicine, 348,* 1051–1056.

Vincent, C. (2005). *Patient safety*. London: Elsevier.

Weingart, S. N., Wilson, R. M., Gibberd, R. W., & Harrison, B. (2000). Epidemiology of medical error. *British Medical Journal, 320,* 774–777.

World Health Organization. (2002). *Quality of care: Patient safety*. Report by the Secretariat. Fifty-Fifth World Health Assembly, A55/13 23 March, Provisional agenda item 13.9. Geneva: Author.

World Health Organization. (2004). *World alliance for patient safety: Forward programme 2005*. Geneva: Author.

Medicare

Medicare is the health insurance program provided by the U.S. federal government for older Americans once they reach the age of 65. Individuals with certain conditions, such as those who are permanently disabled or have end-stage renal disease, may also be eligible for Medicare, regardless of their age. The Medicare program is administered by the Centers for Medicare and Medicaid Services (CMS), a federal agency headquartered in Baltimore, Maryland.

Medicare is a complex program with multiple options for receiving insurance coverage and varying features and rules that accompany each option. The program is continually evolving as new legislation is introduced and plan types, features, or payment policies are changed. While older adults often require more healthcare services due to chronic medical conditions or declining health, they may also have limitations in available income and health insurance options due to retirement from work, resulting in a greater need to understand the insurance benefits provided to them through Medicare. This entry provides a brief history of the Medicare program, including the introduction of prescription drug coverage, a description of the types of decisions Medicare beneficiaries must make, and an overview of special considerations when conducting research on Medicare decision making.

History of Medicare Program

The Medicare program was established in 1965 through Title XVIII of the Social Security Act. The program was designed to provide health insurance coverage for adults aged 65 and older. The eligibility criteria for the program were modified in 1972 to waive the age requirements for individuals with long-term disabilities or end-stage renal disease so that they could begin receiving Medicare coverage. In 1977, a government agency, the Health Care Financing Administration (HCFA; now the Centers for Medicare and Medicaid Services [CMS]) was created to oversee the Medicare program. The Medicare Catastrophic Coverage Act of 1988 expanded benefits and added coverage for outpatient prescription drugs; however, this act was repealed in 1989 due to complaints regarding high premiums charged for those with higher incomes.

The Balanced Budget Act of 1997 (BBA) was designed to control spending on Medicare and resulted in substantial changes to the Medicare program. The act introduced additional health plan choices for people with Medicare, including Medicare-managed care plans and private fee-for-service plans, referred to as Medicare + Choice plans. The act also mandated that HCFA/CMS educate and inform beneficiaries, including providing comparative plan information, with the goal of helping beneficiaries make more informed health plan decisions.

The most recent changes to the program resulted from the passing of the Medicare Drug Improvement and Medicare Modernization Act (MMA) in 2003. The act continued the evolution of Medicare with the Medicare Advantage (MA) program (replacing the Medicare + Choice program of the BBA), a prescription drug discount card program to assist beneficiaries with the cost of prescription drugs, until the prescription drug insurance program was

implemented in 2006, and additional coverage for preventive care services.

Decision Making in Medicare

Beneficiaries must go through several steps when selecting their Medicare coverage, including deciding whether to purchase certain components (parts) of Medicare, understanding the various plan types and the rules and cost implications for each, selecting a specific plan within the type they have chosen after evaluating the individual plan's premiums and co-payments, and considering other possible types of coverage available to them.

Parts of Medicare

Medicare insurance coverage currently consists of four parts, Parts A to D. Medicare Part A is the basic Medicare coverage, which is generally provided free of charge to beneficiaries, provided they or their spouses have paid Medicare taxes. Part A primarily covers inpatient care (i.e., hospital stays, hospice care, skilled nursing facilities) and home healthcare. Beneficiaries may pay a monthly premium (generally deducted from their social security checks) to receive Medicare Part B, which covers outpatient care (e.g., doctor's visits), preventive services (e.g., mammograms, colorectal cancer screening), certain medical supplies (e.g., diabetes supplies), and some home healthcare services not covered by Part A.

Rather than enrolling in Medicare Parts A and B, beneficiaries may choose to receive their Medicare coverage for these services through a MA plan by enrolling in Part C. MA plans are administered by private insurance companies rather than the Medicare program. The final Medicare part, Part D, provides coverage for prescription medicines and was added as result of the MMA of 2003. Those who are in a MA plan via Part C may receive prescription drug coverage as a part of their plan or may have to enroll in Part D separately to receive that coverage. As with Part C, the Medicare Prescription Drug Plans for Part D are run by private insurance companies.

Types of Medicare Plans

There are two primary avenues for receiving Medicare coverage. The first, Original Medicare, encompasses Parts A and B, described above. The majority of beneficiaries are enrolled in Original Medicare, which is a fee-for-service plan operated by the Medicare program. After meeting a deductible, participants pay a coinsurance amount for covered services. A benefit of the Original Medicare plan is the flexibility to go to any provider or hospital that accepts Medicare. However, beneficiaries may need to locate a provider that specifically agrees to accept assignment from Medicare to ensure that they pay only the amount Medicare has determined for a particular service without any additional costs.

The other primary avenue for getting insurance coverage through Medicare is enrolling in a MA plan operated by a private insurance company. There are several variations of plans that fall under the MA umbrella. Probably the most familiar are health maintenance organizations (HMOs) and preferred provider organizations (PPOs). HMOs require participants to use healthcare providers within the plan's specified provider network to receive coverage. PPOs are similar to HMOs but provide some coverage for services that are received outside the network, although at a greater cost to participants than if they used providers within the network. Private fee-for-service (PFFS) plans are similar to Original Medicare but are operated by private insurance companies rather than Medicare; they allow beneficiaries to go to any provider that accepts the plan. In contrast to HMOs and PPOs, PFFS plans have fewer reporting requirements concerning the quality of care provided to their participants. Additional MA plan types include the following, which are not described here and generally serve a very small proportion of Medicare beneficiaries: Medicare Medical Savings Account plans, Medicare Special Needs Plans, Medicare Cost Plans, Demonstration/Pilot Programs, and Programs of All-Inclusive Care for the Elderly.

Potential benefits of an MA plan may include coverage of some items not covered by Original Medicare, such as prescription drugs, and possibly lower out-of-pocket costs. However, potential drawbacks are that beneficiaries often must use providers within the plan's network and may need referrals from the primary care physician to see a specialist. Being restricted to in-network providers can be problematic for beneficiaries who have a long-standing, established relationship with a

provider who is not in the plan's network or leaves the network. Plan restrictions may be particularly difficult for beneficiaries who live in rural areas and may be unable to find a local provider within the network. When selecting a plan type, beneficiaries must balance possible cost savings with plan rules and limitations.

Incorporating Other Types of Coverage

Original Medicare alone generally covers only about half of beneficiaries' healthcare costs. Along with evaluating and selecting among the Medicare coverage options, beneficiaries may also need to choose among other types of insurance to pay for costs not covered by Medicare, such as Medigap insurance, employer-sponsored coverage, and Medicaid.

Beneficiaries who sign up for Original Medicare (Parts A and B) commonly also purchase a supplemental insurance plan (also referred to as Medigap coverage) through a private insurance company. Although sold by private insurance companies, the options for Medigap coverage are standardized (currently labeled with the letters A to L), so that all Medigap plans of the same type offer the same benefits, thereby allowing beneficiaries to more easily compare plans across insurance companies. By purchasing Medigap insurance, beneficiaries can lower some of the deductibles or co-payments they would have to pay with Medicare alone. Some of the plan types also offer additional benefits, such as coverage for prescription drugs. Certain requirements and regulations may affect whether or when an individual can purchase Medigap coverage. For example, beneficiaries who are enrolled in a Medicare Advantage plan (Part C) may not also purchase a Medigap plan.

Older adults who continue to work after age 65 may rely on employer-sponsored insurance as their primary source of coverage and may need to understand how to integrate this coverage with Medicare. Employers generally pay a portion of the insurance premiums, making this coverage less costly for their employees. As a part of their retirement plans, some individuals who are no longer working may also be able to obtain insurance coverage through their former employers. However, given rising costs of insurance, fewer employers are now offering this benefit.

Depending on their incomes, Medicare beneficiaries may also qualify for the Medicaid program, which is administered together by CMS and state Medicaid programs. These individuals are often referred to as "dually eligible" beneficiaries. Medicaid provides health insurance for those with low incomes; the specific income requirements and the benefits of the program vary from state to state. Individuals with some disabilities may also be eligible for Medicaid. While dually eligible beneficiaries have a greater proportion of their healthcare costs paid through their combined coverage, navigating two insurance programs may be challenging, particularly for a population that has traditionally experienced lower education and health literacy levels. In addition, special Medicare rules or options may apply to those who are dually eligible.

Prescription Drug Coverage in Medicare

The feature of the Medicare program that has probably garnered the most public attention and debate is the introduction of prescription drug coverage through the Medicare Modernization Act of 2003. Many older adults use one or more prescription medications on a regular basis, which can become quite costly, particularly for those with limited incomes. However, prior to the MMA, Medicare did not routinely cover outpatient prescription medications, leaving a significant gap in its coverage of healthcare expenses. After the passing of the MMA, Medicare beneficiaries were able to purchase prescription drug discount cards, which assisted them with the cost of prescription drugs in 2004 and 2005 until the Medicare prescription drug coverage went into effect in 2006.

Currently, beneficiaries may receive Medicare drug coverage either as a part of a Medicare Advantage plan or by purchasing a separate prescription drug plan. Both types of plans are run by private insurance companies and may have different premiums, co-payments, or formularies. When selecting a prescription drug plan, beneficiaries may need to carefully compare plan formularies (i.e., list of medications, generic and brand-name, that are covered by the plan) to ensure that their medications are included, especially if they are taking multiple medications. They may also need to stay up-to-date on possible changes to the plan's formularies. Individuals should also be aware of

certain rules or conditions regarding plan enrollment. For example, beneficiaries who do not enroll in a Medicare prescription drug plan when they first become eligible may have to pay a penalty if they decide to enroll at a later time. Furthermore, Medicare beneficiaries with low incomes may also need to explore whether they qualify for a subsidy to pay for prescription drugs available through the Social Security Administration.

A common source of confusion among beneficiaries and a primary point of contention among supporters and opponents of the current drug coverage is the gap in Medicare prescription drug coverage when expenditures reach a certain amount, often called the "donut hole" in coverage. Once beneficiaries' drug expenditures reach this threshold (approximately $2,400 in 2007), they no longer receive coverage until their expenditures exceed another threshold (approximately $3,800 in 2007), at which time their coverage will resume. In other words, those who fall into the donut hole will pay the full price for their medications, as if they did not have coverage.

Educating Beneficiaries About Medicare

Given the complexities of the Medicare program, it is not surprising that past research has generally found beneficiary knowledge about the Medicare program and the available coverage options to be quite low, although beneficiaries tend to be most knowledgeable about the particular type of coverage they currently have. Those who are in the process of choosing an insurance plan may require educational information and materials to help them compare the cost, benefits, and performance of the various plans to make an informed decision. However, beneficiary education must be an ongoing effort that goes beyond just plan comparison information. Even after selecting a plan, beneficiaries need to stay informed about the changes in the Medicare program (e.g., available plan options) and their individual plans (e.g., changes in drug formularies). They must also understand the rules (e.g., when they may change plans) and their rights (e.g., right to appeal a coverage decision).

A provision of the BBA was to educate and inform Medicare beneficiaries. In 1998, CMS implemented the National Medicare Education Program to provide beneficiaries with information they need to make informed decisions about their Medicare coverage. As a part of this program, beneficiaries have access to Medicare information via print copies of the *Medicare & You* handbook, which is updated yearly, the Medicare help line (800-MEDICARE), and the Medicare Web site (www.medicare.gov). Many states also have State Health Insurance Counseling and Assistance Programs (SHIPs) that can assist beneficiaries with Medicare coverage questions.

Conducting Research on Medicare Decision Making

Complementary to efforts to educate beneficiaries are research studies to understand how beneficiaries make decisions regarding their Medicare coverage, including how much they know about the Medicare program, what information they use in selecting their insurance plans, and which approaches are most effective for assisting beneficiaries in making informed decisions. There are several considerations when conducting decision-making research with the Medicare population.

Study Sample

As a part of the study design process, researchers must determine whether to include a sample that is representative of the entire Medicare population or to use specific inclusion and exclusion criteria to restrict the sample. One segment of the Medicare population includes elderly beneficiaries in nursing homes who may be physically or cognitively unable to fully participate in a research study. In addition to older adults, individuals with certain disabilities or conditions, such as end-stage renal disease, are eligible for Medicare. These beneficiaries could have different experiences with the Medicare program than those who became eligible simply due to their age.

Mode of Data Collection

Decisions concerning the study sample should also inform the selection of an appropriate mode of data collection, while balancing practical considerations, such as budget and time. Studies that include ill or institutionalized respondents may be most effective using in-person interviews; however,

this approach may be expensive. Telephone surveys may need modifications for those who are hard of hearing, while mail surveys may need to use larger print for those with eyesight problems. Web surveys may not be effective, given the unfamiliarity with and sometimes reluctance among some older adults to use the Internet.

Survey Questions

Care should be taken when designing survey questions to ensure the validity of the data collected. It is particularly important to ensure that beneficiaries understand any terminology used in the survey questions (unless the goal of the questions is to test beneficiary understanding). There may sometimes be a disconnect between the terminology used by insurers and providers and the way that beneficiaries themselves refer to certain aspects of their Medicare coverage. Cognitive interviewing is one useful approach for tapping into respondents' perceptions and identifying potential problems with the survey questions.

Another consideration when selecting the number and type of survey questions is respondent burden. Some beneficiaries, particularly those in institutions, may be quite ill and unable to answer a lengthy questionnaire. In addition, some older adults experience declines in cognitive abilities over time. For example, declines in working memory, which are common with increasing age, would make it difficult for respondents to retain lengthy response options in their memory to effectively answer questions administered verbally. Questions that will be administered verbally should use short response options.

Study Materials

As with survey questions, study materials should be designed to be easy to comprehend and use. The health literacy of the target population should be considered when designing materials and delivering messages. Possible demographic differences in Medicare knowledge and decision-making approaches may be used to target messages and materials. It may also be helpful to consider preferences among segments of the population. For example, some older beneficiaries may be less likely to use the Internet as a source of insurance information. Other beneficiaries may tend to assume a less active role in medical decision making, possibly relying on doctors, pharmacists, or other professionals to provide them the information they need.

Including Nonbeneficiaries as Respondents

While research studies generally include a sample of participants from the population of interest, studies on Medicare decision making may need to include individuals who are not Medicare beneficiaries themselves. One such situation is when respondents are too ill or cognitively impaired to answer the study question by themselves. In these cases, it may be possible to include a proxy respondent who can provide assistance or answer the questions for the respondent. However, the validity of proxy responses may be a concern, depending on the familiarity of the proxy with the sample member and his or her Medicare coverage and the types of questions asked (i.e., those asking for objective information vs. subjective perceptions or preferences).

Finally, given the complexity of the Medicare program and the available options, even those beneficiaries who are able to respond to a survey themselves may receive assistance from others, such as a spouse, children, or friends, when making health insurance decisions. Some of these individuals may not be enrolled in the Medicare program themselves and may be unaware of the informational resources available to beneficiaries. Researchers may need to investigate the role of these informal caregivers in the Medicare decision-making process.

Carla Bann

See also Decision-Making Competence, Aging and Mental Status; Government Perspective, Informed Policy Choice; Informed Decision Making; Medicaid

Further Readings

Bann, C. M., Berkman, N., & Kuo, T. (2004). Insurance knowledge and decision-making practices among Medicare beneficiaries and their caregivers. *Medical Care, 42*(11), 1091–1099.

Centers for Medicare and Medicaid Services. (2008). *Medicare & you 2008*. Retrieved February 9, 2009,

from http://www.medicare.gov/Publications/Pubs/pdf/10050.pdf
Greenwald, L. M., McCormack, L. A., Uhrig, J. D., & West, N. (2006). Measures and predictors of Medicare knowledge: A review of the literature. *Health Care Financing Review, 27*(4), 1–12.
Oliver, T. R., Lee, P. R., & Lipton, H. L. (2004). A political history of Medicare and prescription drug coverage. *Milbank Quarterly, 82*(2), 283–354.
Regan, J. F., & Petroski, C. A. (2008). Prescription drug coverage among Medicare beneficiaries. *Health Care Financing Review, 29*(1), 119–125.

Memory Reconstruction

All aspects of episodic (information about events) and semantic (knowledge about the world) memory can be influenced by reconstructions based on inferences due to general knowledge or current beliefs, including stereotypical ones. This aspect of memory may serve the once adaptive purpose to keep a consistent and updated model of the world, but it may also lead to memory illusions under certain conditions. Erroneous beliefs about past events may be held with strong confidence, and currently, there is no reliable method that allows one to distinguish true from false memories. In medical decision making, the hindsight bias or "knew it all along" effect may be most relevant. Whenever possible, one should strive for external documentations of current beliefs and decisions to have an external validation of memories later. Cuing techniques borrowed from the cognitive interview technique may enhance the chances for retrieving true memories if an external validation source does not exist.

Background

Since the beginnings of empirical research on memory, theorizing has been dominated by the "storehouse metaphor." Theories of memory tend to describe structures and substructures in which information about events (episodic memory) or knowledge about the world (semantic memory) is stored and can later possibly be retrieved. Forgetting has been attributed either to the decay of information or to the lack of effective retrieval cues for finding and reactivating the trace in the store with more empirical support for the latter conception. Hence, according to this view, information is first encoded into a representation which in turn is stored somewhere in a memory trace which resides there until it is retrieved (or not). It can be refreshed by rehearsal or repeated retrieval, but otherwise it remains essentially unchanged.

In his classic book *Remembering*, published in 1932, Frederic C. Bartlett challenged this static view and demonstrated the shortcomings of the metaphor. He presented an old Indian myth to his British university students and had them reproduce the contents repeatedly after various time intervals. His observation was that the reproductions of the exotic story were increasingly and systematically distorted toward common knowledge about typical stories in the Western culture. Hence, remembering is an active process which uses generic world knowledge (schemata) to reconstruct information encountered in the past. Although controlled replications of Bartlett's original study have been somewhat elusive, there is now a wealth of empirical evidence with other materials demonstrating the influence of knowledge-based reconstruction and judgment processes in episodic memory. One early experimental example by Carmichael and colleagues is presented in Figure 1, which shows the systematic influence on the reproductions of simple line drawings that had been labeled with "memory cues" for reproduction.

Reconstruction and Distortion Phenomena

Memory reconstructions have been demonstrated experimentally in a variety of domains: Research on the so-called hindsight bias, eyewitness suggestibility, and source memory has shown that reconstructions of past events can be manipulated in predictable ways. Furthermore, research on schemata, false memories and so-called implanted childhood memories has demonstrated that memory illusions can even be created for events that never happened or for items that were not presented. The phenomena are sketched below. The most prominent memory and judgment distortion is the hindsight bias or "knew it all along" effect. According to this phenomenon, events seem less surprising after the fact. For example, participants in a typical experimental study may be asked to judge the likelihoods of future events (e.g., soccer

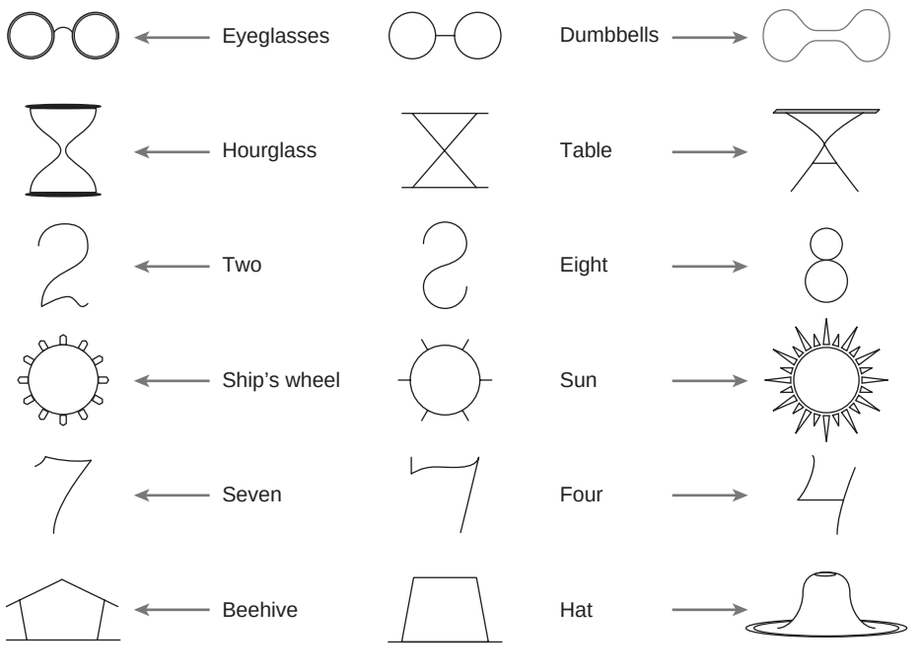


Figure 1 Examples of stimuli (middle column) and typical reproductions in two conditions with different verbal labels

Source: Adapted from Carmichael, L., Hogan, H. P., & Walter, A. (1932). *Journal of Experimental Psychology, 15,* 73–86.

results). If they are instructed to recall their initial estimates after the outcomes are known, the mean judgments tend to move toward the actual outcomes. This procedure also works with numerical estimates that are not probabilities, for example, almanac questions ("How high is the Eiffel tower?"). The phenomenon is extremely robust, and neither financial incentives for accuracy nor the reduction of self-presentation motivations reduces it by a substantial amount. Hence, it is seen as a genuinely cognitive phenomenon by most researchers. It is obvious that the tendency to overestimate the predictability of events after the fact may lead to substantial misjudgments that can also affect medical decisions and their post hoc evaluation.

Closely related phenomena are the misinformation effects or suggestibilty effects in memory that have been studied extensively by Elizabeth K. Loftus and colleagues. In the basic paradigm, participants are presented with slide shows or videos of events (e.g., car accidents or staged crimes), followed by questions about the events that either contain misleading information (e.g., "Did another car pass the red car while it stopped at the intersection with the stop sign?" when there was actually a yield sign) or some suggestive wordings of the questions (e.g., "How fast were the cars going when they smashed into each other?"). Later memory tests regularly show an increased tendency to report suggested details (e.g., the stop sign) or details in accordance with the suggestion (e.g., higher velocity estimates and the illusion of having seen broken glass when the word *smashed* rather than *hit* was used in the leading question). Similar results have been reported repeatedly, and they are extremely relevant for eyewitness testimony and possible influences of suggestive interrogations on eyewitness accuracy.

Recent research on source monitoring (remembering the context of information acquisition) has shown that the reconstruction of the episodic context is particularly susceptible to judgment processes that rely on generic knowledge about schemata, stereotypes, or correlations between aspects of the information and information source. Reconstructing the source of information therefore

resembles a problem-solving task more than a simple retrieval of contextual details.

More dramatic instances of illusory memory have shown that people may "remember" items that never appeared or events that never happened. Research on schemata in the 70s of the past century revealed that after the presentation of information that fits a certain schema (e.g., a visit to the restaurant), participants tend to accept schema-typical events as having been presented (e.g., the waiter presents the bill) although these were not stated explicitly. Later research with simple semantically associated word lists revealed that the belief that certain nonpresented items were presented can be held with high confidence, and participants sometimes claim that they vividly "remember" the presentation of the item. For example, the presentation of items such as a table, stool, wood, back, sofa, or rocker will enhance the probability that participants later recognize or recall the word *chair* as presented before although it was not. The false memory rate may even exceed that for some actually presented items, and their "forgetting" rate seems to be slower. This illusion is also quite robust and can easily be provoked with appropriate lists in classroom demonstrations. It is, however, not restricted to artificial word lists; even autobiographical pseudo-memories can be "implanted" by plausible suggestions, and the confidence with which participants report details of these memories increases with repeated testing.

Theoretical Accounts and Debates

Although there is no single theory of the phenomena listed above, their common denominator is that the apparent retrieval of memory traces is rather a process of reconstruction that is influenced by general world knowledge, associations with the context and between items, and response biases. Memory distortions are probably the price we pay for an effective memory system that tries to maintain a consistent representation of the world and fills gaps in the memory representations by betting on the most likely information, given the continuously updated knowledge in the system. The demonstration of malleable memories is commonly interpreted as a genuine maladaptive bias. However, the experimental settings are created deliberately in a biased way to make the reconstruction processes visible to the researcher in the way optical illusions operate at the boundaries of our perceptual system and reveal the principles according to which it works. In all the examples cited above, including the false memory paradigm with word lists (except perhaps implanted childhood memories), the reconstructions produced by the memory system are "intelligent" ones that fill in information that is likely, given the state of the world and knowledge about it. One can speculate that these guesses are in most cases accurate and hence contribute to an effective knowledge system that has evolved.

Whereas early accounts of the hindsight bias and the misinformation effect claimed that the memory trace is damaged by the outcome knowledge and the misleading information, respectively, later studies suggested that the original memory trace (if encoded in the first place) is still available but that its accessibility may be reduced. Given a lack of successful retrieval, knowledge-based reconstruction fills in the gap. Theoretical approaches emphasize associative, similarity-based, or judgmental mechanisms, but probably all mechanisms contribute to the reconstructions.

Research on false memories has focused on individual and situational factors that determine the amount of suggestibility or allow distinguishing true from false memories. For example, there are individual differences in the susceptibility for memory distortions, but they affect only the amount of the illusion, not its occurrence or nonoccurrence. The same is true for cognitive processes such as visual imagination that can enhance the effect sizes.

Concerning the possibility to distinguish between true and false memories, the results are mixed: Some studies found differences in experienced vividness and emotional intensity between true and implanted memories; others did not. Anyway, there is no clear dividing line, and the expert commission of the American Psychological Association investigating the relation between potential suggestion effects and the increasing reports about recovered memories of childhood sexual abuse concluded that currently there is no reliable method for distinguishing true memories from false memories.

Countermeasures

Given the pessimistic conclusion on the distinguishability of true and false (or reconstructed)

memories, the most obvious countermeasure to avoid potentially fallacious reconstructions after the fact is a documentation in external representations (written, recorded, or videotaped). Especially to avoid hindsight biases in medical decisions, evidence and conclusions drawn from them should be documented. If new facts appear later (e.g., a tumor is detected), it will otherwise sometimes be hard to achieve a fair evaluation of previous judgments. In the same manner, it is recommended to write down memory protocols immediately after an important event has been witnessed for potential cases of later questioning.

If, however, external records do not exist, a method has been developed for eyewitness interrogations called the *cognitive interview*. This method draws on the fact that the chances of retrieving memories are improved by reinstating (at least in imagination) as many aspects of the episodic context as possible. These may serve as effective cues that aid retrieval of original information in memory. Of course, this method tries to avoid all kinds of leading questions or suggestions of information. Similar principles can be used for retrieving one's own previous judgments.

Arndt Bröder

See also Bias; Cognitive Psychology and Processes; Fuzzy-Trace Theory

Further Readings

Alpert, J. L., Brown, L. S., Ceci, S. J., Courtois, C. A., Loftus, E. F., & Ornstein, P. A. (1998). Final report of the American Psychological Association working group on investigation of memories of childhood abuse. *Psychology, Public Policy, and Law, 4*(4), 931–1068.

Memon, A., & Higham, P. A. (1999). A review of the cognitive interview. *Psychology, Crime & Law, 5*(1–2), 177–196.

Pohl, R. (Ed.). (2004). *Cognitive illusions: A handbook on fallacies and biases in thinking, judgement, and memory*. Hove, UK: Psychology Press.

Mental Accounting

The terms *mental account* and *mental accounting* were coined in the 1980s by Richard Thaler to refer to certain mental representations (*accounts*) and cognitive processes (*accounting*) related to decision outcomes and events, particularly transactions involving money. It is useful to distinguish core mental accounts, which are relatively stable structures, from specific mental accounts constructed to represent a new economic decision.

Core Mental Accounts

Behavioral life cycle theory aims to explain how people deal with the economic transactions they encounter in everyday life across the lifespan. It is a formal economic model incorporating assumptions of bounded rationality theory, notably that people construct simplified mental representations of their economic world. One assumed simplification is that people mentally partition their income and expenditure transactions over time into discrete budget periods, often weekly or monthly, to coincide with significant recurring events such as payday, or utility or housing payments. Another is that economic resources are allocated to one of three core mental accounts: *current income*, *current assets,* or *future income*. The way that these accounts relate to expenditure decisions varies. Each has a different budget constraint, with the current income account having the lowest resistance to spending and the future income account the highest. In addition, different categories of expenditure may relate to these broad accounts differently: For example, low-cost, frequently occurring purchases such as a newspaper are more likely to be allocated to the current income account. Finally, subaccounts for specific purposes may also be constructed, such as "holiday money." All this implies that money is mentally categorized or labeled.

Behavioral life cycle theory predicts that people will violate a principle of rational economic behavior known as *fungibility*—in essence, the principle that money from all sources should be interchangeable. It can explain certain anomalies of economic behavior (violations of fungibility) such as holding savings attracting a lower rate of interest while borrowing money at a higher rate (more common than some readers might think). The notion of core mental accounts as simplified representations of long-term resources has mainly been applied to consumer behavior and to basic financial decisions

such as saving versus spending unexpected (windfall) income. However, it is also clearly relevant to personal healthcare decisions, including health insurance, which involves uncertain future benefits but ongoing costs, that is, monthly premium payments that would normally be allocated to a current income account.

Specific Mental Accounts

A second use of the term *mental account* derives from prospect theory and is related to the framing effect. In contrast to the ongoing representation discussed above, Amos Tversky and Daniel Kahneman defined a mental account as an outcome frame set up for a specific consumer choice or transaction (initially these authors used the term *psychological account*). They distinguished three levels of account, differing in the extent to which contextual information might be included, that were investigated in the Jacket and Calculator decision problem. In one version of the problem, participants were asked to imagine they were about to purchase a jacket for $125 and a calculator for $15 and had been informed that the calculator was on sale for $10 at another branch, a 20 minutes' drive away. Only 29% of participants responded that they would drive to the other store, but when the problem was rephrased so that the prices of jacket and calculator were reversed, with the calculator being $125 in one store and $120 in the other, 68% were prepared to do so. Although the difference in the price of the calculator and the cost of the whole shopping trip are the same in both forms of the problem, respondents tended to represent it differently, which can be understood in terms of the specific mental account primed for the transaction. For example, the calculator price difference could be framed in terms of a minimal, topical, or more comprehensive mental account. For a *minimal account*, it would be evaluated relative to the status quo (a saving of $5). For a *topical account*, a reference point from which to evaluate the price difference would be derived from the "topic" of the decision: in this case the actual price of one of the calculators (e.g., $5 less than $15). For a more *comprehensive account*, the reference point would be based on a wider context, for example, the whole shopping bill ($5 less than $140). The change in majority preference across versions of the problem is explained by a tendency to construct topical accounts that incorporate the most relevant, but not all, aspects of the transaction.

Mental Accounting Processes

Mental accounting theory assumes that three basic cognitive processes are used to allocate financial outcomes to core or specific mental accounts: coding, integration, and segregation. Coding is one of the basic processes of original prospect theory by which decision outcomes are coded as gains or losses relative to a reference point and evaluated according to a value function that is concave for gains and convex for losses. The above analysis in terms of topical accounts implies that relative, rather than absolute, differences are coded and evaluated as gains or losses.

Turning to integration and segregation processes, these refer to the construction of representations of two or more decision outcomes: specifically, whether outcomes are allocated to the same or different mental accounts, respectively. These processes have been used to explain a range of decision-related phenomena, including conditions under which prior outcomes may influence current decisions. Suppose a player has just gambled on a roulette wheel and won $10 from the house. This newly won money could be integrated into a core, current income account, or alternatively, it could be segregated from other resources and allocated to a new subaccount, the "house money." The player may subsequently integrate the possible outcomes of the next spin of the roulette wheel with this house money account and be more prepared to gamble this than he or she would money from the general current income account. In fact, consistent with a process of integrating winnings in this way, some studies have found that players were more likely to make riskier gambling choices after a prior win.

The important consequence of integrating versus segregating sequences of decision outcomes is that the overall evaluation of the events is different, assuming that gains and losses are evaluated according to the value function of prospect theory. For example, in the case of two successive roulette wins of $10, the overall evaluation if they remained segregated would be greater than the evaluation of the corresponding integrated gain of $20. There is

some evidence that preferences for clearly segregated gains and losses compared with the equivalent integrated ones are consistent with this prediction. However, contrary to it, one study found that two segregated losses were evaluated less negatively than was the integrated larger loss. Other effects of prior outcomes on current decisions that can be explained in terms of these mental accounting processes include sunk costs and escalation of commitment, where current decisions are integrated with prior outcomes, and de-escalation of commitment, where they are segregated.

Another issue that has received attention is whether the gains and losses *within* a transaction are integrated or segregated: for example, whether the pleasure of acquiring a new car is segregated from, or integrated with, the pain of paying for it, or whether the experience of work and being paid are integrated or segregated. Drazen Prelec and George Loewenstein's double-entry mental accounting theory argues that gains (the black—income, consumption) and losses (the red—paying, working) are mentally represented in separate accounts that interact in different ways, particularly depending on how such outcomes are distributed over time. Support was found for the prediction that paying for something before acquiring or consuming it is generally less painful than the other way round because of the mechanism of prospective accounting (the pain of paying for a holiday in advance is buffered by the prospect of the positive experience to come). As well as time, contextual factors such as the form of payment have been found to moderate the integrating, or coupling, as opposed to the segregation, or decoupling, of outcomes on the red and the black sides of the account.

Psychological Functions of Mental Accounts

Several adaptive functions of mental accounts and accounting processes have been proposed. First, an important function of core mental accounts is self-control. Resources in current asset and future income accounts are resistant to the temptations of immediate consumption, thereby supporting longer-term economic well-being. Second, an important cognitive function of both specific and core accounts is to mentally represent the important features of the personal economic world effectively and efficiently. Since the personal economy is too complex for the limited information processing capacities of the human mind to represent completely, simplifications are necessary. Despite these simplifications, however, core mental accounts are sufficient to facilitate effective budgeting. Finally, a function of mental accounting processes that has been proposed is hedonic editing. This is the notion that people actively choose to segregate or integrate decision outcomes to maximize their subjective value. Alternatively, some evidence suggests that integration of prior outcomes may be motivated by a desire to avoid losses. However, as discussed earlier, preferences for segregated or integrated outcomes may reflect the consequences of these processes rather than prior motivation to derive greater satisfaction from decision outcomes.

Extensions and Issues

The range of decision phenomena that have been elucidated by mental accounting theory since the 1980s is rather extensive, particularly in consumer psychology. For example, as well as those already discussed, important insights have been gained on consumers' responses to discounts and surcharges, their reactions to unexpected price changes, consumer credit decisions, and evaluations of the bundling or separation of product features. Also, our understanding of important aspects of personal finance, such as saving and tax-related behavior, has been advanced. On the other hand, boundary conditions and other limitations have been identified, and several domains, including health decisions, have been underresearched.

One criticism has been that some of the decision anomalies explained by mental accounting theory are rather context-specific. In addition, the validity of core mental accounts has been questioned: Consistent effects on economic behavior and expectations have not always been found, and alternative core account structures have been proposed. Furthermore, it has been argued that core mental accounts are in practice rather malleable, with money mentally transferred from one to the other rather easily. Finally, the overextensive use of the term *mental accounting* to refer to basic coding and editing processes has also been criticized as adding little to what is understood via prospect theory and cognitive process models of decision making. Nevertheless, the fundamental

insight of mental accounting theory, that financial resources are not fungible, has been consistently validated and shown to be important across many domains of decision making. In conclusion, to misquote George Orwell's *Animal Farm*: All money is equal, but some monies are more equal than others.

Rob Ranyard

See also Bounded Rationality and Emotions; Editing, Segregation of Prospects; Prospect Theory; Sunk Costs

Further Readings

Gärling, T., Karlsson, N., & Selart, M. (1999). The role of mental accounting in everyday economic decision making. In P. Juslin & H. Montgomery (Eds.), *Through a fruitful lens: Forty years of Swedish research on judgment and decision making*. Mahwah, NJ: Lawrence Erlbaum.

Prelec, D., & Loewenstein, G. (1998). The red and the black: Mental accounting of savings and debt. *Marketing Science, 17,* 4–27.

Ranyard, R., & Abdel-Nabi, A. (1993). Mental accounting and the process of multiattribute choice. *Acta Psychologica, 84,* 161–177.

Ranyard, R., Hinkley, L., Williamson, J., & McHugh, S. (2006). The role of mental accounting in consumer credit decision processes. *Journal of Economic Psychology, 27,* 571–588.

Shefrin, H., & Thaler, R. H. (1988). The behavioral life-cycle hypothesis. *Economic Inquiry, 26,* 609–643.

Thaler, R. H. (1985). Mental accounting and consumer choice. *Marketing Science, 4,* 199–214.

Thaler, R. H. (1999). Mental accounting matters. *Journal of Behavioral Decision Making, 12,* 183–206.

Thaler, R. H., & Johnson, E. (1990). Gambling with the house money and trying to break even: The effects of prior outcomes on risky choice. *Management Science, 36,* 643–660.

Tversky, A., & Kahneman, D. (1981). The framing of decisions and the psychology of choice. *Science, 211,* 453–458.

Meta-Analysis and Literature Review

Literature reviews designed to summarize large volumes of information are frequently published. When a review is done systematically, following certain criteria, and the results are pooled and analyzed quantitatively, it is called a meta-analysis. A well-designed and -interpreted meta-analysis can provide valuable information for decision making. However, there are several critical caveats in performing and interpreting them.

The information generated in medical research has tremendously increased in recent years. New studies are constantly being published, and clinicians, researchers, and policy makers may find it nearly impossible to stay current. More and more review articles that pool the results of multiple studies are seen.

Combining available information seems reasonable and can save considerable time, effort, and money. Nowadays, meta-analyses are used to design future research, to provide evidence in the regulatory process, and especially to modify clinical decision making. A meta-analysis is powerful but also controversial because several conditions are critical, and small violations of those conditions can lead to misleading results. Under scrutiny, some meta-analyses have been inappropriate and their conclusions not fully warranted. This entry covers basic concepts of meta-analysis and discusses its caveats.

Main Aims of a Meta-Analysis

The main aims of a meta-analysis are as follows:

1. To summarize results from several individual studies
2. To evaluate differences in the results among studies
3. To overcome small sample sizes of individual studies
4. To increase precision in estimating effects
5. To evaluate effects in subsets of patients
6. To determine if new studies are needed to further investigate a topic

Critical Issues in Performing a Meta-Analysis

Identification and Selection of Studies

Two phases need to be followed when selecting studies for a meta-analysis: (1) the literature

search, in which potential studies are identified; and (2) the clear definition of inclusion criteria. Three problems affect these phases: publication bias and search bias in the former, and selection bias in the latter.

Publication Bias

Searches of databases can yield many studies. However, these databases include only published studies. Such searches are unlikely to yield a representative sample because studies that show a positive result (usually in favor of a new or standard treatment) are more likely to be published. This selective publication is called publication bias. Consider the case of the publication status of studies on antidepressants. Based on studies registered with the FDA, 97% of the positive studies are published versus only 12% of the negative ones. Furthermore, when the nonpublished studies are not included, the positive effects of individual drugs increase between 11% and 69%.

One reason for publication bias is that drug manufacturers are not generally interested in publishing negative studies. Also, journal editors favor positive studies because these are the ones that make the headlines. To ameliorate the effect of publication bias, a serious effort should be made to identify unpublished studies. This is much easier now due to improved communication between researchers and by registries in which all the studies of a certain disease or treatment are reported. In some medical areas, the exclusion of studies conducted in non-English-speaking countries can increase publication bias.

The National Institutes of Health maintains a registry of all the studies it supports, and the U.S. Food and Drug Administration keeps a registry and database in which drug companies must register all trials they sponsor. Registries of published and unpublished trials supported by pharmaceutical companies are also available (e.g., GlaxoSmithKline's Clinical Study Register). The Cochrane collaboration keeps records of systematic reviews and meta-analyses of many diseases and interventions.

Search Bias

Even in the ideal case of no publication bias, a faulty search can miss some publications. In searching databases, care should be given to using a set of key words that is as complete as possible. This step is so critical that most recent meta-analyses include the list of key words used. The search engine (e.g., PubMed, Embase, Web of Science, or Octopus) is also critical, affecting the type and number of studies that are found. Small differences in search strategies can produce large differences in the set of studies found.

Selection Bias

The identification phase usually yields a long list of potential studies, many of which are not directly relevant to the topic of the meta-analysis. This list is then subject to predefined inclusion criteria. This critical step is also designed to reduce differences among studies, to eliminate duplication of data or studies, and to improve data quality.

To reduce selection bias, it is crucial that inclusion criteria for the studies be clearly defined and that these studies be evaluated by at least two researchers, with the final list chosen by consensus. The objective is to select studies that are as similar as possible with respect to these criteria. Even with careful selection, differences among studies will remain, and it becomes hard to justify pooling the results to obtain an overall conclusion.

In some cases, it is particularly difficult to find similar studies, and sometimes the discrepancies and low quality of the studies can prevent a reasonable integration of results. Authors may decide not to pool the results, due to a systematic qualitative inadequacy of almost all trials and a lack of consistency in the studies and their methods.

Stratification is an effective way to deal with inherent differences among studies and to improve the quality and usefulness of the conclusions. An added advantage to stratification is that insight can be gained by investigating discrepancies among strata. There are many ways to create coherent subgroups of studies. For example, clinical trials can be stratified according to their quality scores. Commonly used scores are based on the use of allocation concealment, the use of blinding, the drop-out rate, the outcome measurement, and the use of intention-to-treat analysis.

Funnel Plot

The funnel plot is a technique used to investigate the possibility of biases in the identification

and selection phases. In a funnel plot, the size of the effect (defined as a measure of the difference between treatment and control) in each study is plotted on the horizontal axis against precision or sample size on the vertical axis. If there are no biases, the graph will tend to have a symmetrical funnel shape centered on the average effect of the studies. When negative studies are missing, the graph shows asymmetry. Funnel plots are simple, but their objective is to detect a complex effect, and they may be misleading. For example, lack of symmetry in a funnel plot can also be caused by heterogeneity in the studies. Another problem with funnel plots is that they are difficult to evaluate when the number of studies is small.

For example, consider a meta-analysis that evaluates the effect of anticoagulant treatment to prevent venous thromboembolism (PE) in hospitalized patients. The treatment was useful to prevent PE, with no significant increase in major bleeding. Figure 1 shows the funnel plots for these two outcomes. The asymmetry in the top plot suggests bias due to a lack of inclusion of small studies showing an increase in the risk of PE. The bottom plot shows the symmetry of the funnel plot for major bleeding, suggesting absence of bias.

Evaluation of Heterogeneity of Results

Heterogeneity refers to the degree of dissimilarity in the results of individual studies. The dissimilarities can be related to their inclusion criteria but sometimes might not be easy to elucidate. As the level of heterogeneity increases, the justification for an integrated result becomes more difficult. A tool

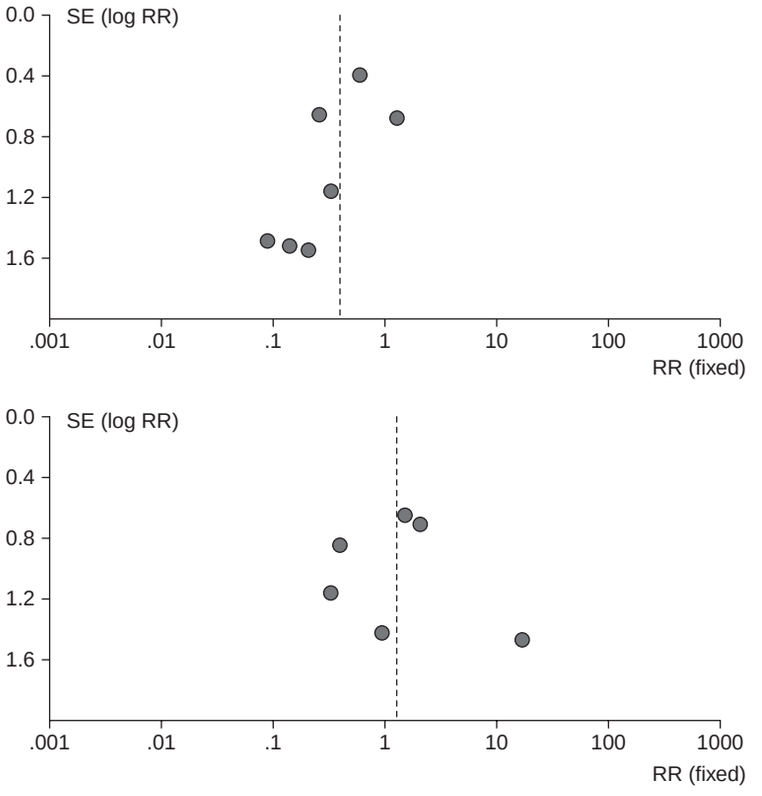


Figure 1 The funnel plot

Notes: Top—An asymmetrical funnel plot of studies that measured the presence of pulmonary embolism in patients with anticoagulant prophylaxis, suggesting that small studies where the association was positive are missing. *Bottom*—A symmetrical funnel plot of studies that measured major bleeding, suggesting absence of selection bias. The standard error of the log relative risk (*SE* log([RR]) is used as a measure of precision.

that is very effective to display the level of heterogeneity is the forest plot. Here, the estimated effect of each study, along with a line representing a confidence interval, is drawn. When the confidence intervals overlap, the heterogeneity is low. The forest plot includes a reference line at the point of no effect (e.g., 1 for relative risks, odds ratios, and hazard ratios; 0 for risk difference). When some effects lie on opposite sides of the reference line, it means that heterogeneity is high, and the conclusions of a meta-analysis are compromised.

Consider a meta-analysis that evaluated the effect of high-dose versus standard-dose statin therapy on the risk of coronary death or myocardial infarction (MI) in patients with stable coronary disease or acute coronary syndromes. The forest plot showed a homogeneous benefit of high-dose statin therapy across trials (Figure 2). In contrast, consider another meta-analysis that studied the association between statin use and the risk of breast cancer. The forest plot shows a heterogeneous association across case-control studies (Figure 3). Cochran's Q test and the I^2 test are frequently used to determine the significance of heterogeneity.

A meta-analysis of clinical trials compared the survival of patients with esophageal carcinoma who received neo-adjuvant chemotherapy versus those who underwent surgery alone. Only one of the eight studies showed that neo-adjuvant chemotherapy was significantly beneficial. Three of the studies suggested that it was harmful although not statistically significant. The pooled result was marginally significant in favor of the treatment (p = .05). This positive result was largely due to the fact that the only study with a significantly positive result was also the largest (with 400 patients in each arm vs. an average of 68 per arm for the other studies). Even though the test for heterogeneity was not significant, the marginal p value and the differences in study size make the results of this meta-analysis suspect.

Availability of Information

Most reports of individual studies include only summary results, such as means, standard deviations, proportions, odds ratios, relative risks, and/or hazard ratios. Other than the possibility of errors in reporting, the lack of information can severely limit the types of analyses and conclusions. The lack of information from individual studies can preclude the comparison of effects in predetermined subgroups of patients. The best scenario is when data at the patient level are available. In such cases, the researcher has great flexibility in the analysis. Consider a meta-analysis of the value of microvessel density in predicting survival in non-small-cell lung cancer. Information on individual patients was obtained by contacting research centers directly. The data allowed varying the cutoff point to classify the microvessel density as high or low and to use statistical methods to ameliorate heterogeneity.

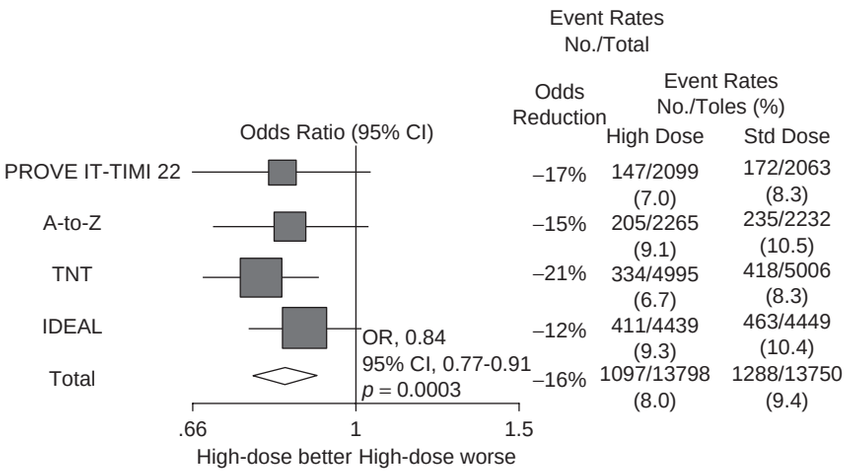


Figure 2 Low level of heterogeneity

Note: All trials show better outcome with high-dose statin therapy, indicating a low level of heterogeneity.

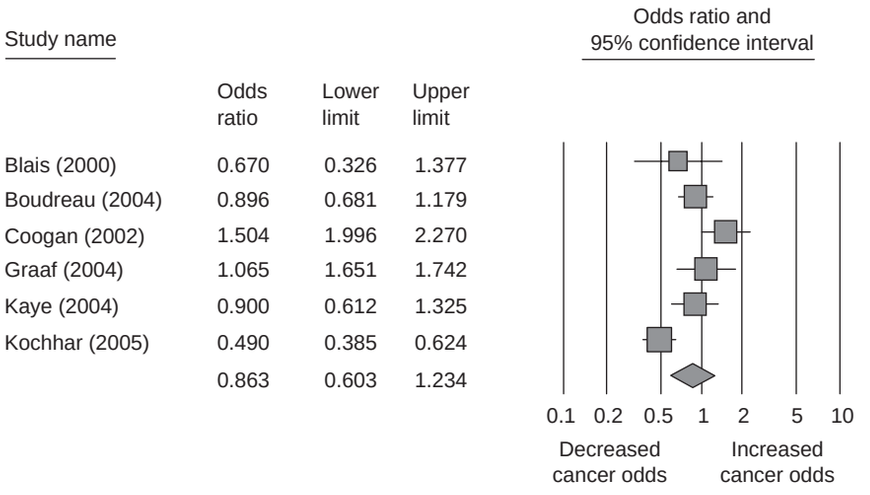


Figure 3 High level of heterogeneity

Note: Some studies show increased cancer odds with the use of statins, and others show decreased cancer odds, representing a high level of heterogeneity. The overall effect is shown as a diamond at the bottom of the figure.

Appropriate Analysis of Data

There are specific statistical techniques used in meta-analysis to analyze and integrate the information. The data from individual studies can be analyzed using either of two models: (1) The fixed-effects model assumes that the treatment effect is the same across studies. This common effect is unknown, and the purpose of the analysis is to estimate it with more precision than in the individual studies. (2) The random-effects model, on the other hand, assumes that the treatment effect is not the same across studies. The goal is to estimate the average effect in the studies.

In the fixed-effects model, the results of individual studies are pooled using weights that depend on the sample size of the study, whereas in the random-effects model each study is weighted equally. Due to the heterogeneity among studies, the random-effects model yields wider confidence intervals. Both models have pros and cons. In many cases, the assumption that the treatment effect is the same in all the studies is not tenable, and the random-effects model is preferable. When the effect of interest is large, the results of both models tend to agree, particularly when the studies are balanced (i.e., they have similar numbers of patients in the treatment arm and in the control arm) and the study sizes are similar. But when the effect is small or when the level of heterogeneity of the studies is high, the result of the meta-analysis is likely to depend on the model used. In such cases, the analysis should be done and presented using both models.

It is highly desirable for a meta-analysis to include a sensitivity analysis to determine the robustness of the results. The most common way to do this is by analyzing the data using various methods and to present the results when some studies are added (or removed) from the analysis. If these actions cause serious changes in the overall results, the credibility of the results is compromised.

Testing of effects suggested by the data and not planned a priori increases considerably the risk of false-positive results. One common problem is the practice of performing multiple subgroup analyses according to baseline characteristics. The best way to prevent the possibility of false-positive results is to determine the effects to be tested before the data are collected and analyzed. Another method is to adjust the p value according to the number of analyses. In general, post hoc analyses should be deemed exploratory, and the reader should be aware to judge the validity of the conclusions.

Meta-Analysis of Rare Events

Lately, meta-analysis has been used to analyze outcomes that are rare and that individual studies

were not designed to test. The sample size of individual studies provides inadequate power to test rare outcomes. Scarcity of events causes serious problems in any statistical analysis. The reason is that, with rare events, small changes in the data can cause big changes in the results. This problem can persist even after pooling data from many studies. Instability of results is also exacerbated by the use of relative measures (e.g., relative risk and odds ratio) instead of absolute measures of risk (e.g., risk difference). Adverse or harmful events are prime examples of important rare outcomes that are not always formally analyzed statistically.

Consider a recent meta-analysis that combined 42 studies to examine the effect of rosiglitazone on the risk of MI and death from cardiovascular causes. The overall estimated incidence of MI in the treatment groups was low: 0.006 (86/14,376), or 6 in 1,000. Furthermore, four studies did not have any events in either group, and 2 of the 42 studies accounted for 28% of the patients in the study. Using a fixed-effects model, the odds ratio was 1.42, that is, the odds of MI were 42% higher in patients using rosiglitazone, and the difference was statistically significant (95% confidence interval 1.03–1.98). Given the low frequency of MI, this translates into an increase of only 1.78 MIs per 1,000 patients (from 4.22 to 6 per 1,000). Furthermore, when the data were analyzed using other methods or if the two large studies were removed, the effect became nonsignificant.

Future Directions

Like many other statistical techniques, meta-analysis is a powerful tool when used judiciously; however, there are many caveats in its application. Clearly, meta-analysis has an important role in medical research, public policy, and medical decision making. Its use and value will likely increase, given the amount and the speed at which new knowledge is being created and the availability of specialized software for performing it.

A meta-analysis needs to fulfill several key requirements to ensure the validity of its results: well-defined objectives, including precise definitions of clinical variables and outcomes; appropriate and well-documented study search and selection strategy; evaluation of bias in the identification and selection of studies; description and evaluation of heterogeneity and quality of studies; justification of data analytic techniques; and use and description of sensitivity analysis. It is imperative that clinicians and researchers be able to assess critically the value and reliability of the conclusions of meta-analyses to apply them to the decision-making process.

Adrian V. Hernandez

See also Evidence-Based Medicine

Further Readings

Bailar, J. C. (1997). The promise and problems of meta-analysis. *New England Journal of Medicine, 337,* 559–561.

Cochrane Collaboration: http://www.cochrane.org

Egger, M., Davey Smith, G., & Altman, D. (Eds.). (2001). *Systematic reviews in health care: Meta-analysis in context*. London: BMJ Books.

GlaxoSmithKline, Clinical Study Register: http://ctr.gsk.co.uk/welcome.asp

Moher, D., Cook, D. J., Eastwood, S., Olkin, I., Rennie, D., & Stroup, D. F. (1999). Improving the quality of reports of meta-analyses of randomised controlled trials: The QUOROM statement. Quality of reporting of meta-analyses. *Lancet, 354,* 1896–1900.

Whitehead, A. (2002). *Meta-analysis of controlled clinical trials*. New York: Wiley.

Minerva-DM

Minerva-DM (DM = decision making) is a memory-based model of choice, probability judgment, and frequency judgment. Minerva-DM and its predecessor, Minerva 2, are similar to neural network models and can be used to simulate human behavior. Minerva-DM has been used to account for many of the common heuristics and biases discovered in the judgment and decision-making literature, including the availability and representativeness heuristics, base-rate neglect, mere-exposure effect, overconfidence effect, Bayesian conservatism, and frequency judgment.

Minerva-DM was developed on the premise that memory processes serve as input into the higher-order processes of probability and frequency

judgment. Thus, errors and biases that arise as part of the memory encoding or retrieval process are assumed to cascade into errors and biases in judgment. A number of studies support this contention. For example, overconfidence has been shown to covary with two main factors: the structure of the environment (the ecology) and how well information has been encoded in long-term memory. The idea that overconfidence is affected by how well information has been encoded suggests that the overconfidence effect is, in large part, a memory phenomenon rather than a judgment phenomenon. Moreover, it suggests that remediation of the overconfidence effect should focus on memory variables, not judgment variables.

Model Description

Minerva-DM and Minerva 2 are akin to a single-layer neural network model, where the input corresponds to a pattern of features extracted from the environment and the output is a function of the contents of memory that are activated by the input. Both Minerva 2 and Minerva-DM assume an exemplar-based memory representation where each individual experience (i.e., episode) is represented by a distinct memory trace. Multiple experiences of similar events are therefore assumed to result in multiple, albeit similar, memory traces stored in memory. Because Minerva-DM preserved the representational and computational details inherent in Minerva 2, it can be used to simulate a variety of effects in the recognition memory literature, as well as the aforementioned phenomena in the frequency judgment and probability judgment literatures. Through the use of simulation methodology, Minerva-DM is able to make a priori predictions regarding the relationship between memory and judgment without the need to evoke specialized heuristic mechanisms.

Assumptions

Minerva-DM is based on the recognition memory model called Minerva 2. Minerva 2 makes two fundamental assumptions: (1) Memory consists of a database of instances that represent an individual's past experiences, and (2) recognition memory judgments are based on a global familiarity signal derived by matching a memory cue against all traces in memory simultaneously. Minerva 2 has been used successfully to model the influence of different types of experience on people's recognition and frequency judgments.

Minerva-DM extended Minerva 2's capability by adding two additional assumptions. First, Minerva-DM assumes that memory traces can be partitioned into components that represent "hypotheses," "data," and "context." Hypotheses correspond to events about which the participant is making a judgment, such as a disease hypothesis or a treatment hypothesis. The data component corresponds to the information on which the participant is making his or her assessment, such as the symptoms associated with diseases in past patients. The context component corresponds to environmental or task information available to the decision maker. The second additional assumption is that Minerva-DM assumes a *conditional* memory search process rather than a global memory search. The conditional memory search process involves first activating those memory traces in long-term memory that are consistent with the observable data (the presenting symptoms of the patient being diagnosed). The participant is then assumed to estimate the relative frequency of various disease hypotheses within the set of traces activated by the initial observable data. The relative frequencies are then normalized through a comparison process to derive a conditional probability judgment corresponding to the probability of the hypothesis in light of the data, that is, $p(H|D)$.

Both Minerva 2 and Minerva-DM assume that memory is accessed by *probing* memory with information provided to the decision maker. In Minerva-DM, decision makers are assumed to access memory by probing with observable "data," such as a patient's presenting symptoms. This initial probing of memory allows the decision maker to partition episodic memory into those past patients who had similar presenting symptoms (the relevant set) and those who had different presenting symptoms. Once partitioned, the decision maker is then assumed to access the relevant set to determine how many past cases correspond to a particular hypothesis (e.g., pneumonia). For example, if asked for a judgment of p(pneumonia|outcome of test X), the decision maker would be assumed to first activate all traces in episodic memory of patients who had a similar score on test X and then probe this

activated subset with the pneumonia hypothesis to determine how many within the activated set correspond to pneumonia.

Connection to Bayesian Probabilities

The conditional probability judgment rendered by Minerva-DM is analogous to Bayesian conditional probability, and Minerva-DM mimics Bayesian probabilities under appropriate parameterization. However, Minerva-DM also anticipates the psychological and task variables that lead to systematic deviation from Bayes's theorem.

Extension of Minerva-DM to Hypothesis Generation

More recent research has extended Minerva-DM to deal with hypothesis generation and information search processes. The extended model, HyGene, accounts for a variety of new judgment phenomena, including the subadditivity effect and the relationship between judgment, hypothesis generation, and individual differences in working memory capacity. Like Minerva-DM, HyGene places a premium on understanding the relationship between memory and judgment by explicitly modeling the interrelationship between long-term memory and judgment. However, unlike Minerva-DM, HyGene is able to model how people generate the to-be-evaluated hypotheses from long-term memory and posits a limited-capacity working memory system that constrains the number of hypotheses one can include in the comparison process. Thus, HyGene describes how people generate hypotheses from long-term memory, and the psychological processes that govern how these hypotheses are fed into the processes of probability judgment, information search, and hypothesis testing.

Michael R. Dougherty and Rick P. Thomas

See also Cognitive Psychology and Processes; Errors in Clinical Reasoning; Heuristics; Learning and Memory in Medical Training; Support Theory; Unreliability of Memory

Further Readings

Dougherty, M. R. P. (2001). Integration of the ecological and error models of overconfidence using a multiple-trace memory model. *Journal of Experimental Psychology: General, 130,* 579–599.

Dougherty, M. R. P., Gettys, C. F., & Ogden, E. E. (1999). MINERVA-DM: A memory processes model for judgments of likelihood. *Psychological Review, 106,* 180–209.

Dougherty, M. R. P., & Hunter, J. E. (2003). Hypothesis generation, probability judgment and individual differences in working memory capacity. *Acta Psychologica, 113,* 263–282.

Hintzman, D. L. (1988). Judgments of frequency and recognition memory in a multiple-trace memory model. *Psychological Review, 95,* 528–551.

Moore, D., & Healy, P. J. (2008). The trouble with overconfidence. *Psychological Review, 115*(2), 502–517.

Thomas, R. P., Dougherty, M. R., Sprenger, A., & Harbison, J. I. (2008). Diagnostic hypothesis generation and human judgment. *Psychological Review, 115,* 155–185.

Mixed and Indirect Comparisons

Mixed and indirect comparisons are specialized methods of performing meta-analysis. These methods can be useful in replacing or augmenting a meta-analysis, but one must be aware of potential pitfalls when using them. This entry presents definitions, methods, assumptions, and examples of each of the terms.

Indirect Comparisons

Definition

A direct comparison of two interventions occurs when they are compared within the same randomized controlled trial (RCT). An indirect comparison is any method of comparing two interventions without the use of direct comparisons between the two. It can be used in meta-analysis of RCTs when a reviewer wishes to compare two interventions and no direct comparisons exist. It can also be used in conjunction with direct evidence to strengthen results.

Methods

There are two possible methods of performing indirect comparisons. One method is to take all

evidence regarding the effects of two interventions from various sources and compare the two interventions as if they came from the same trial. This method has been referred to as unadjusted indirect comparison or the naive method. This method should be avoided since it violates the inherent randomization that occurs within the trials and is known to produce misleading results.

All further references to indirect comparisons in this entry refer to adjusted indirect comparisons. In this method, two interventions are compared indirectly by using their direct comparisons with a third common intervention.

This method is most easily demonstrated with an example. Suppose a reviewer wishes to assess the difference in efficacy between two drugs, A and B. While there are no RCTs directly comparing the two interventions, there are trials comparing each drug with a placebo (i.e., A vs. P and B vs. P). If efficacy is measured in terms of a mean difference, then two separate meta-analyses can be performed: one comparing Drug A with the placebo, resulting in a mean difference of d_{AP}, and one comparing Drug B with the placebo, resulting in a mean difference of d_{BP}. The mean difference between A and B (d_{AB}) can be expressed as a "difference of differences"—that is,

$$d_{AB} = d_{AP} - d_{BP}.$$

Using the standard formula for the variance of a difference of independent variables, one can compute the variance of this difference as

$$\mathrm{Var}(d_{AB}) = \mathrm{Var}(d_{AP}) + \mathrm{Var}(d_{BP}).$$

The variance of an estimated quantity can be defined generally as the uncertainty one has as to the estimate. The uncertainty increases with an indirect estimate as the variance will be higher than either of the direct estimates.

With this information, one can compute a standard estimate with a confidence interval for the mean difference between the two drugs. If the efficacy is measured by a risk ratio or an odds ratio, the procedure is the same, but the ratios must be converted to the log scale first and then exponentiated to obtain the final results. This will result in a "ratio of ratios."

This estimate will be unbiased as long as there is no interaction between the magnitude of the treatment effect and the covariates that define the subgroups in the corresponding studies—that is to say that the effects are transitive and the populations exchangeable. While this assumption is difficult to verify, it should be noted that it is the same assumption that is made in a standard meta-analysis of direct comparisons.

Combining Direct and Indirect Evidence

If both direct and indirect evidence for a comparison exist, one may wish to combine the evidence to strengthen the result. This can be done using standard meta-analytic techniques. If one has a point estimate with corresponding standard error (defined as the square root of the variance) for both direct and indirect evidence, then, provided there are no studies included in both estimates (i.e., three-arm trials—A vs. B vs. P—should be included only in direct estimation), one can combine the evidence using the inverse variance meta-analytic method. Either fixed or random effects can be employed. A fixed effects meta-analysis assumes that each study is measuring the same underlying effect, while a random effects meta-analysis allows for the possibility that each study is estimating a different effect and the researcher is attempting to find the "average study effect."

Example

In a study on drug treatments for chronic insomnia, the main drugs were classified into two groups: benzodiazepines and nonbenzodiazepines. Two separate meta-analyses for the primary outcome, sleep onset latency (SOL), were conducted comparing each drug class with a placebo using direct evidence. Both drug classes were found to be superior to the placebo, reducing SOL by an average of 10.0 minutes (95% CI: 3.4, 16.6) for benzodiazepines and by an average of 12.8 minutes (95% CI: 8.8, 16.9) for nonbenzodiazepines. The two drug interventions can be compared indirectly from these estimates. The point estimate of the difference between the two drug classes can now be estimated indirectly as 2.8 minutes (12.8 minus 10.0) with 95% confidence interval (–4.9, 10.5).

This confidence interval can be computed using the aforementioned formula by converting the confidence intervals to variances and then converting the corresponding computation of variance back to a confidence interval. From this estimate, one would say that while nonbenzodiazepines appeared superior to benzodiazepines, the difference was not statistically significant because the confidence interval includes zero.

If in addition to this indirect estimate, there was direct evidence comparing the two interventions, they could be combined. For example, if direct evidence showed a difference of 1.6 minutes (95% CI: –3.4, 6.5), then the two estimates (direct and indirect) can be combined in a standard inverse variance meta-analysis (Figure 1). Thus a combined estimate of 1.9 minutes (95% CI: –2.2, 6.1) is obtained that is a combination of all available evidence, both direct and indirect.

Mixed Comparisons

Definition

Mixed comparison is a meta-analytic technique that combines direct and indirect evidence to formulate comparisons among interventions. This could be as simple as the method described in the previous example but is more generally used to describe network meta-analyses and mixed treatment comparisons. This latter method combines all available evidence to compare all interventions simultaneously. The method is flexible in that it can properly use multiarm trials and yield estimates among interventions for any number of different interventions that have been examined, as long as there is a connected network of studies comparing all interventions.

Network Meta-Analysis

Network meta-analysis is the process of estimating the difference between two treatments indirectly through a connected network of studies where the two interventions are compared through other interventions directly. Say the researchers wish to compare treatments A and B, and they have studies involving other treatments C, D, E, and F. This can be done indirectly if a connected network can be established through these other treatments. For example, if there are studies comparing AC, CD, BC, and BD, one could compare A and B both through their direct comparisons with C and through A's indirect comparison with D and B's direct comparison with D. On the other hand, if the only existing studies compared AC, AD, CD, BE, BF, and EF, there would be no way to compare A and B even indirectly since there is one network of ACD and another of BEF and there is no connection between them.

This idea of a connected network of comparisons is inherent to conducting mixed treatment comparisons.

Mixed Treatment Comparisons

With mixed treatment comparisons, the focus shifts from comparing two interventions to attempting to simultaneously compare three or more interventions. By using the networks discussed in the previous section, this can be achieved as long as all interventions can be connected with a network of available studies.

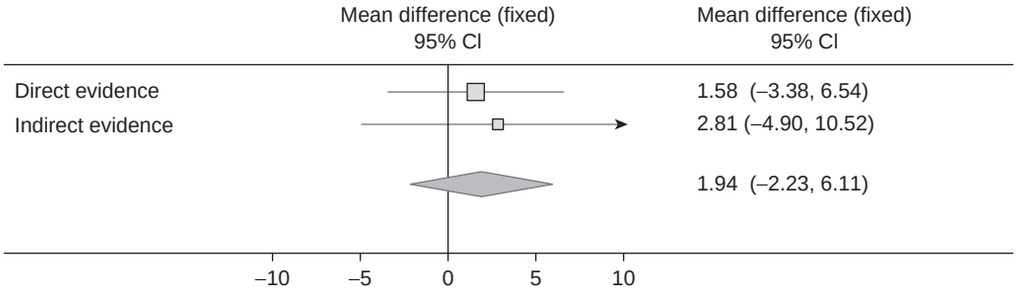


Figure 1 Combining direct and indirect evidence

Mixed-treatment comparisons are best used when multiple treatments (three or more) have been compared across numerous studies in various combinations and the objective is to determine which of the treatments is either best overall or best in a specific situation or population.

For example, consider a situation where six different interventions have been examined in several two- and three-arm trials in various combinations. Fifteen separate pairwise comparisons can be made, and thus a standard meta-analysis will be of little help in determining the best intervention. However, by simultaneously comparing all interventions through a mixed treatment comparison, it becomes more straightforward to determine the best one or to rank them in order of decreasing benefit.

Methods

To conduct a mixed treatment comparison, one must first select one treatment as a reference treatment. This is usually the placebo or standard treatment if one exists. All other treatments are compared with this reference to define the basic parameters, and all other comparisons can be defined as functional parameters of these basic parameters. These estimates can be calculated with the fixed or random effects assumptions.

For example, if there are four treatments A, B, C, and D, and A is chosen as the reference, direct and indirect evidence could be used to establish d_{AB}, d_{AC}, and d_{AD}—the differences between each of the other treatments and the reference treatment. All other contrasts can be defined as functions of these basic contrasts:

$$d_{BC} = d_{AC} - d_{AB},$$

$$d_{BD} = d_{AD} - d_{AB}, \text{ and}$$

$$d_{CD} = d_{AD} - d_{AC}.$$

Using Bayesian statistical methods via Gibbs sampling, it is straightforward to attach probabilities to each of the interventions as to their likelihood of being the best intervention. In Gibbs sampling, thousands (or even tens or hundreds of thousands) of results are simulated. The interventions can be ranked at each of the sampled iterations, and the number of times each intervention was ranked the best can be counted.

Example

In a review on short-acting agents for emergency room procedural sedation, there were a total of six RCTs that examined four different interventions: midazolam (M), etomidate (E), propofol (P), and ketofol (K). The trials had the following forms: M versus E (two trials), M versus E versus P, M versus P, E versus P, and P versus K. Since there is a connected network (i.e., M, E, and P are all compared directly with each other and K can be connected indirectly to M and E through its comparison with P), a mixed treatment comparison can be performed. Midazolam, being the standard analgesic used in practice, is chosen as the reference drug. Using a Bayesian formulation and using Gibbs sampling to combine the data for the primary outcome (procedure time), the results were computed as shown in Figure 2.

All three "active interventions" significantly reduced procedure time compared with midazolam, with ketofol having the greatest effect. While there was no direct evidence linking ketofol to midazolam, a comparison using the indirect evidence could still be made. The wider confidence interval of the ketofol comparison (compared with those of propofol and etomidate) is partially a result of the lack of direct comparisons between the two treatments. Although not shown here, the pairwise estimates and confidence (or credible) intervals for all the active interventions can also be computed.

The Gibbs sampling also provides estimates of the probability of each intervention being the best intervention based on the evidence. In this example, in 73% of the iterations, ketofol had the greatest difference in procedure time from midazolam; thus this represents an approximation of its probability of being the best intervention of the four drugs being compared. In a similar way, probabilities for any rank or the mean rank of each intervention can be estimated.

Assumptions

As with simple indirect comparisons, mixed treatment comparisons require similar assumptions

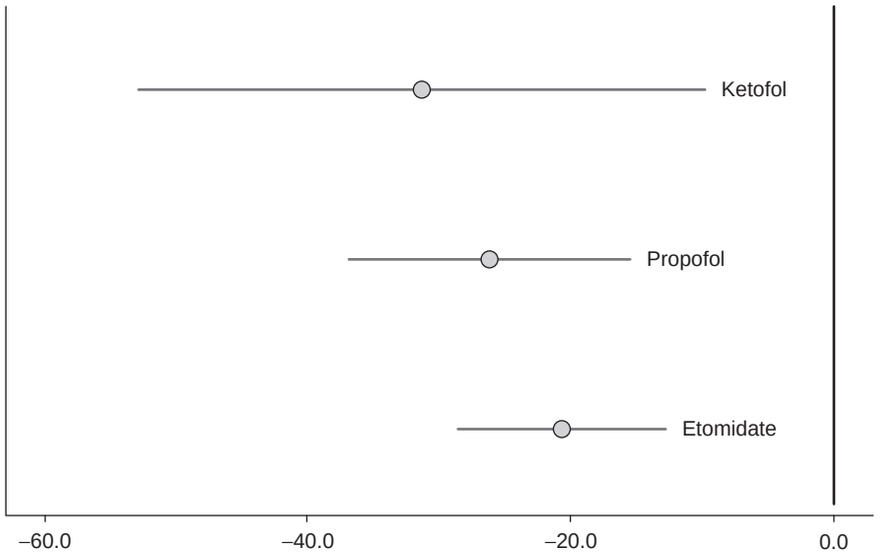


Figure 2 Mixed treatment comparison results

to those of a standard meta-analysis. In a fixed-effects analysis, the assumption is made that all true differences between interventions are the same across trials, even if a particular trial did not examine one or more of the interventions. For example, the true difference between A and B would be identical in trials that compared A versus B, A versus D, B versus C, or even C versus D if A or B or both had been examined in those trials that omit them. In a random effects analysis, this assumption is relaxed and one allows effects to vary between studies and assumes only that the variance between studies is constant. One must also be cautious of incoherence, which occurs when an indirect effect can have opposite conclusions, depending on which indirect comparator is used. As with any meta-analysis, these assumptions are difficult to verify, but it is important to note that as long as unadjusted comparisons are avoided (as they are in the mixed treatment comparisons described above), the within-trial randomization is still being preserved, and if standard meta-analytic assumptions hold, there will be no bias.

Ben Vandermeer

See also Bayesian Evidence Synthesis; Confidence Intervals; Evidence Synthesis; Fixed Versus Random Effects; Meta-Analysis and Literature Review; Randomized Clinical Trials; Variance and Covariance

Further Readings

Bucher, H. C., Guyatt, G. H., Griffith, L. E., & Walter, S. D. (1997). The results of direct and indirect treatment comparisons in meta-analysis of randomized controlled trials. *Journal of Clinical Epidemiology, 50,* 683–691.

Caldwell, D. M., Ades, A. E., & Higgins, J. P. T. (2005). Simultaneous comparison of multiple treatments: Combining direct and indirect evidence. *British Medical Journal, 331,* 897–900.

Glenny, A. M., Altman, D. G., Song, F., Sakarovitch, C., Deeks, J. J., D'Amico, R., et al. (2005). Indirect comparisons of competing interventions. *Health Technology Assessment, 9*(26), 1–134.

Lu, G., & Ades, A. E. (2004). Combination of direct and indirect evidence in mixed treatment comparisons. *Statistics in Medicine, 23,* 3105–3124.

Lumley, T. (2002). Network meta-analysis for indirect treatment comparisons. *Statistics in Medicine, 21,* 2313–2324.

Song, F., Altman, D. G., Glenny, A. M., & Deeks, J. J. (2003). Validity of indirect comparison for estimating efficacy of competing interventions: Empirical evidence from published meta-analyses. *British Medical Journal, 326,* 472–476.

Models of Physician–Patient Relationship

Clinical decision making is the use of diverse strategies to generate and test potential solutions to problems that are presented by patients. It involves using, acquiring, and interpreting the indicators and then generating and evaluating hypotheses. Clinical decision making takes place within the context of the physician–patient relationship and thus is embedded within the consultation process. Indeed, even if the duration is brief, individuals involved in dyadic interactions can, and often do, influence each other's cognitions, emotions, and behaviors. Theories about decision making suggest that people do not have stable and preexisting beliefs about self-interest but construct them in the process of eliciting information. Therefore, the way information is provided by the health provider is crucial in assisting patients to construct preferences and then deciding on a course of action. Thus, ignoring the nature of the physician–patient relationship that occurs during clinical encounters could undermine our understanding of how current clinical decision-making processes can be improved.

The first section of the entry presents the main characteristics of the two basic models of physician–patient relationships: the doctor-centered model and the patient-centered model. In response to the growing expectations of patients as well as the burden of managing uncertainty in routine clinical decision making, the second section briefly summarizes how a third model, the shared decision-making model, has evolved in the past decade. The last section highlights the gaps in knowledge and areas needing further research.

Basic Models

Patient-Centered Model

The patient-centered model of care is grounded in the client-centered psychotherapy model. It refers to a philosophy of care that aims at the best integration possible of the patient's perspective. This philosophy of care was proposed as an answer to a medical model that was focused mainly on the disease and that was felt to be unsatisfying. In line with this philosophy, the patient-centered model has been further developed to fit routine clinical practice. The main characteristics of the patient-centered model are (a) exploring the disease and the experience of the disease (the illness experience), (b) understanding the person and his or her situation, (c) finding common ground, (d) integrating health prevention and promotion, (e) promoting the physician–patient relationship, and (f) fostering realistic expectations. Finding common ground is defined as an agreement between the physician and the patient on three elements: the nature of the problem, the goals of treatment, and the roles each wants to play in decision making. Congruence between patients and their physician on the nature of the problem, the options, and their roles in decision making is expected to foster favorable patient outcomes, especially in the area of mental health (e.g., depression outcomes, well-being). The ultimate goal of the patient-centered model is the appropriate level of involvement of individuals in decisions affecting their health.

In recent years, in response to the difficulties associated with its definition and evaluation in its current form, the patient-centered model has been reconceptualized with a focus on five main characteristics: (1) the biopsychosocial perspective (the understanding of the person and his or her situation), (2) the patient as a person (exploration of the disease and the experience of the disease), (3) the sharing of power and responsibilities (finding common ground), (4) the therapeutic relationship (the promotion of the physician–patient relationship), and (5) the physician as a person (recognition of the influence of the personal qualities and subjectivity of the physician in the practice of medicine). This reconceptualization puts the focus on the sharing of responsibilities in decision making and the necessity of considering both perspectives, those of the patient and those of the physician.

Informed Decision-Making Model

In the decision-making community, the patient-centered model is closely related to the informed decision-making model, also known as the consumerist model or the informative model. The informed decision-making model refers to a model in which the information transfer is one-way (the physician transfers to the patient all the

medical information that is needed for making a decision, the decision deliberation is made by the patient or sometimes by the patient with family members or some other individuals, and the decision about implementing the treatment is solely under the responsibility of the patient. In other words, the physician provides information, the patient applies values, and then the patient decides. This model is in contrast to the paternalistic decision-making model.

Doctor-Centered Model

The term *doctor-centered model* refers to a model of care in which the physician interprets the problems that are presented by the patient in terms of his or her own explanatory framework. It is understood as a more conventional model with a low sensibility to the unique context of the patient. In this model of relationship, the physician attempts to assign the patient's problems to one of the preestablished disease categories. The physician-centered model has its origins in the 19th century and is still better known as the structured approach to patients' consultation, namely, the subjective/objective/assessment/plan model (S.O.A.P. model):

1. Subjective complaints of the patients are sought.
2. Objective signs are found during the examination.
3. An assessment is made.
4. A plan is proposed.

Paternalistic Decision-Making Model

In the decision-making community, the physician-centered model is closely related to the paternalistic decision-making model. The term *paternalistic decision-making model* refers to a model in which the information transfer is one-way (the physician transfers to the patient a minimum of medical information that is needed for informed consent), the decision deliberation is made by the physician alone (or the medical team) and sometimes with other physicians, and the decision about implementing the treatment is solely under the responsibility of the physician. In other words, the physician makes decisions for the patient's benefit independent of the patient's values or desires. This model is in contrast to the informed decision-making model.

Shared Decision-Making Model

Shared decision making is defined as a process by which a healthcare choice is made by practitioners together with the patient and is said to be the crux of patient-centered care. It locates itself in the middle of a continuum between the paternalistic decision-making model and the informed decision-making model. Shared decision making rests on the best evidence as to the risks and benefits of all available options, including doing nothing. It includes the following components: establishing a context in which patients' views about treatment options are valued and seen as necessary; transferring technical information; making sure patients understand this information; helping patients base their preference on the best evidence; eliciting patients' preferences; sharing treatment recommendations; and making explicit the component of uncertainty in the clinical decision-making process. It relies on the best evidence about risks and benefits associated with all available options (including doing nothing) and on the values and preferences of patients, without excluding those of health professionals.

Shared decision making stresses the negotiation process between two divergent explanatory models of illness: the medical model and the patient model. In other words, it is a transactional model that makes explicit the tensions and ineffectiveness in clinical encounters on the basis of communication problems between models that contain the patient's and doctor's understanding of the cause of the patient's illness. Therefore, the clinical encounter is the local setting in which patient–doctor interactions are transactions between their respective explanatory models. This implies that the clinical encounter should foster the following: (1) the development of a therapeutic or working alliance through the establishment of an empathic milieu; (2) the eliciting of the patient's explanatory model and illness problems; (3) the presentation, by the doctor, of his explanatory model in layman's terms; (4) the shifting by the patient toward the doctor's model to make a working alliance possible; and (5) the open acknowledgment by the doctor of discrepancies between exposed models. The final result of this negotiation process is

(6) the change in position by one or both of the doctor and patient to set up a mutually agreed position. Overall, this transactional model is congruent with the revised version of the patient-centered model and the current shared decision-making model. Collaborative decision making is deemed even more crucial in situations of clinical uncertainty where a "good" decision might lead to an undesirable outcome. Consequently, fostering shared decision making in clinical settings has the potential to help both health providers and patients recognize the uncertainty that is present in the decision-making process, a first step for managing uncertainty in routine clinical decisions.

Interpretive Decision-Making Model

In this model of decision making, the patient is uncertain about his or her own values. The physician role consists in assisting the patient in elucidating his or her own values. The physician is a counselor to the patient but remains neutral regarding which values should be favored.

Deliberative Decision-Making Model

In this model of decision making, the patient is uncertain about his or her values and is open to considering suggestions. The physician role consists in teaching the desirable values. In contrast with the physician in the interpretative model, the physician is a coach who identifies which values should be favored.

Further Research

Exemplary medical practice is based on a greater consideration being given to factors related to the patient as well as to factors related to the doctor, with the goal of developing the greatest possible mutual understanding between these two individuals. Poor practice in medicine can result from insufficient clinical knowledge, the absence of a relationship with patients, a lack of understanding of the behavior, perceptions, and problems of patients, and insensitivity to the context. Thus, exemplary practice of medicine takes into consideration the doctor, the patient, and the context. An understanding of the conduct of patient-doctor interactions thus becomes a prerequisite to a better understanding of the medical discipline. Although many models of the physician–patient relationship are available, it is not clear which model best fits the diverse types of patients or clinical situations. Therefore, it remains essential that future studies investigate the fit between the models and the types of patients or clinical situations. This has the potential to streamline the process of clinical decision making in overburdened healthcare clinical settings and ensure quality of care.

France Légaré

See also Decisions Faced by Patients: Primary Care; Managing Variability and Uncertainty; Patient Decision Aids; Risk Communication; Shared Decision Making

Further Readings

Balint, M. (1972). *Patient-centred medicine*. London: Regional Doctor Publications. (Based on the first international conference of the Balint Society in Great Britain on "The doctor, his patient, and the illness," held on March 23–25, 1972, at the Royal College of Physicians, London)

Charles, C., Gafni, A., & Whelan, T. (1997). Shared decision-making in the medical encounter: What does it mean? (or It takes at least two to tango). *Social Science & Medicine, 44*(5), 681–692.

Emanuel, E. J., & Emanuel, L. L. (1992). Four models of the physician-patient relationship. *Journal of the American Medical Association, 267*(16), 2221–2226.

Kleinman, A., Eisenberg, L., & Good, B. (1978). Culture, illness and care: Clinical lessons from anthropological and cross-cultural research. *Annals of Internal Medicine, 88,* 251–258.

McWhinney, I. R. (1985). Patient-centred and doctor-centred models of clinical decision-making. In M. Sheldon, J. Brooke, & A. Rector (Eds.), *Decision-making in general practice* (pp. 31–46). New York: Stockton Press.

Mead, N., & Bower, P. (2000). Patient-centredness: A conceptual framework and review of the empirical literature. *Social Science & Medicine, 51*(7), 1087–110.

O'Connor, A. M., Tugwell, P., Wells, G. A., Elmslie, T., Jolly, E., Hollingworth, G., et al. (1998). A decision aid for women considering hormone therapy after menopause: Decision support framework and evaluation. *Patient Education and Counseling, 33*(3), 267–279.

Rogers, C. R., & Sanford, R. C. (1984). Client-centered psychotherapy. In H. I. Kaplan & B. J. Sadock (Eds.), *Comprehensive textbook of psychiatry* (4th ed., pp. 1374–1388). Baltimore: Williams & Wilkins.

Stewart, M., Belle Brown, J., Weston, W., McWhinney, I. R., McMillan, C. L., & Freeman, T. R. (1995). *Patient-centered medicine: Transforming the clinical method*. London: Sage.

Tversky, A., & Kahneman, D. (1981). The framing of decisions and the psychology of choice. *Science, 211,* 453–458.

Monetary Value

By avoiding the difficult and uncomfortable issue of assigning monetary values to intangible health benefits, cost-effectiveness analysis (CEA) sacrifices the easy decision rule of cost-benefit analysis (CBA): An intervention or program represents an economically appropriate use of scarce resources if and only if the value of its health benefits exceeds its costs. Unfortunately, for the analyst preparing a CEA and the decision-maker faced with the findings from a CEA, the monetization of intangible health benefits by CEA can only be avoided in appearance. The decision maker must still wrestle with the value versus cost trade-off to interpret the results of the analysis. Therefore, the analyst also faces the dilemma of whether or not to provide the decision maker with guidance as to whether or not the intervention or program is worth its costs. That guidance often takes the form of a comparison of the incremental cost per health outcome achieved with a threshold value. This entry provides an overview of the key issues that analysts and decision makers should consider when using a threshold value (measured as cost per quality-adjusted life year [QALY]) to determine whether or not an intervention or program is cost-effective.

Theoretically Appropriate Threshold

The decision problem can be conceptualized as an attempt to maximize health benefits for a given cost. The implicit $/QALY is the shadow price on a QALY from this maximization, which represents the opportunity cost of the resources devoted to producing an additional QALY (technically speaking, the shadow price is the value of the Lagrange multiplier on the budget constraint). Hence, the threshold depends on any parameter of the decision context that affects the objective function or the constraints. Examples of factors that would affect the optimal threshold include the nature of the decision maker's utility function (e.g., extent of risk aversion), the position on utility function (how many other programs already funded), the available budget, and the discount rate. In addition to these factors, some research implies that the willingness to pay may be higher when the condition being treated is life threatening versus a condition that impairs only quality of life or that social willingness to pay for health benefits may differ by characteristics of the beneficiary such as age. This suggests that the basic premise that "a QALY is a QALY" may not be valid in some contexts, which would further require different thresholds based on the type of condition being treated.

The many determinants of the optimal threshold value should give the analyst and decision maker pause in adopting a "generic" threshold that is not attuned to their specific context. However, without appealing to a generic threshold, the decision maker would be left with little of a framework on which to make necessary adjustments. Similarly, the analyst is often writing for an unknown client with an unknown context (e.g., when publishing a CEA in the peer-reviewed literature). Therefore, the analyst's desire to conclude with a generic recommendation is also understandable.

Common Generic Thresholds

Rules of Thumb

Given the psychological predilection for round numbers, it should come as no surprise that rules of thumb (ROTs) such as $50,000 or $100,000 per QALY are among the most common thresholds proposed in CEA studies. The most often used threshold appears to be $50,000 per QALY. The origin of the $50,000 threshold is, in part, that it approximated U.S. spending per renal dialysis patient year more than 20 years ago. Given the legislative decision to cover care for kidney failure patients by providing them with an entitlement to Medicare, it has been argued that other care with comparable or better cost-effectiveness should also

be covered. However, note that this "dialysis standard" is based on an approximation of $50,000 per life year; the implied $/QALY threshold would actually be considerably higher, given that dialysis patients' quality-of-life scores are well below 1.0. Furthermore, under the reasoning that care for dialysis patients is worth its cost, that would imply the $50,000 threshold is a floor for a reasonable threshold, not a ceiling.

In the United Kingdom, the role of CEA is institutionalized via the National Institute for Clinical Excellence (NICE). NICE has adopted a round number "range of acceptable cost-effectiveness" (£20,000 to £30,000). Notably, the midpoint of that range is very close to $50,000 at exchange rates prevailing in early 2008, though it has generally been less than $50,000 at historical exchange rates. However, a recent analysis concluded that while the probability of rejection of an intervention by NICE did rise with cost/QALY, other factors also influenced decisions, and the threshold appears to be closer to £45,000 in practice.

One particular difficulty with round number ROTs is that they do not change with inflation, income growth, healthcare spending, or other factors unless they take a discrete (and large) jump to another round number. The issue of adjusting threshold values over time is taken up in more detail below.

Another set of ROT thresholds was developed as part of the World Health Organization's Choosing Interventions That Are Cost Effective (WHO-CHOICE) project. That project deemed interventions to be highly cost-effective within a country if the cost per QALY was less than that country's gross domestic product (GDP) per capita, cost-effective if it was between 1 and 3 times GDP per capita, and not cost-effective if it was greater than 3 times GDP per capita. WHO reported thresholds in year 2000 dollars for 14 regions of the world. These thresholds ranged from $1,381 (highly cost-effective care in the WHO's lowest income region) to $94,431 (cost-effective care in WHO's highest income region).

League Tables

The league table approach examines the cost-effectiveness of various interventions relative to each other. A given intervention's cost-effectiveness is then judged based on whether or not other interventions with similar or higher cost/QALY values are deemed to be generally accepted medical practices. A significant concern with the league table approach is that it can be self-referential and self-fulfilling. To the extent that current practice is influenced by factors such as the existence of third-party payment, the decision maker should not be comfortable with the implicit presumption that current practice reflects an optimal resource allocation. In addition, comparisons via league tables require a strong presumption that the studies are directly comparable. Differences such as quality of life assessments by the general population versus those who have experienced and adapted to the target health state, or differences based on assessment methodology (e.g., standard gamble vs. time trade-off) would lead to questions regarding the validity of some cross-study comparisons.

Using Empirical Data to Infer a $/QALY Threshold

A large body of literature has estimated the value of a statistical life by studying actual behavior (e.g., wage variation for jobs according to occupational mortality, willingness to pay for safety improvements) or contingent valuation methods (surveys on hypothetical money vs. risk trade-offs). Conversion of the findings from this literature into an implied value of a QALY in 1997 U.S. dollars implied values that varied widely across studies. However, the median study implied a value per QALY of $265,000, and 80% of the reviewed studies implied values over $100,000. This provides empirical evidence suggesting that common round number ROTs of $50,000 and even $100,000 may be too low. All the reviewed studies were from developed countries (the United States, Canada, the United Kingdom, Denmark, and France), but non-U.S. studies implied values similar to those of U.S. studies.

Additional empirical evidence is consistent with threshold values far above conventional ROTs. Data regarding the costs and benefits of medical advances and willingness to purchase unsubsidized health insurance coverage have been used to infer thresholds. Maintaining the assumption that society has, on average, been willing to pay for advances in medical technology since

1950 leads to an inference that the lower bound for the value of a QALY is $183,000 in year 2003 dollars. Similarly, maintaining the assumption that those without access to subsidized health insurance are similar to the entire population leads to an upper bound for the value of a QALY of $264,000.

Valuations of Health Benefits Based on Calibration of Microeconomic Models

Several threshold estimates are based on the specification of plausible utility function parameters for values and risk aversion. Basically, these estimates ask how much a utility maximizing person with "reasonable" preferences should be willing to spend for an expected health benefit. Research using this approach suggested a threshold of about double a person's annual income or approximately $72,500 on average in the United States in 2006.

However, recent research based on utility maximization models, though not aimed specifically at estimating a $/QALY threshold, is consistent with very high values of health benefits. These estimates and their implications for health spending are likely to generate substantial controversy. One study argued that the monetary value of 20th-century health improvements in the United States averaged $1.2 million per person, with improvements in longevity between 1970 and 2000 adding about $3.2 trillion annually to the national wealth. These findings arise from a strong complementarity between consumption and health. Essentially, spending more on healthcare lowers current consumers' current utility because some nonhealth consumption must be forgone, but it raises future utility even more by extending the number of time periods in which consumption can take place. An extension of this argument contended that even small health gains from end-of-life care can have value far in excess of the typically assumed value of a life-year. Provided the bequest motive is weak, a utility-maximizing individual near the end of life has no better use for his or her assets than to spend them on healthcare even if healthcare offers only a modest chance of a modest extension of life. The key dynamic driving these conclusions is the relatively low opportunity cost of extra health spending.

The $/QALY Threshold and Healthcare Cost Growth

To this point, thresholds have been discussed primarily as a decision rule for CEA. In addition, the chosen threshold (if enforced) has substantial implications for healthcare cost growth. The higher the chosen threshold, the higher the healthcare cost growth rate will be as a greater variety of current and future interventions will be deemed cost-effective.

Politically, cost control may even be viewed as the primary objective of the application of CEA to healthcare. However, the objective of CEA, as rooted in welfare economics, is not to contain cost growth. CEA's basis in welfare economics implies that its purpose is solely to help assure that whatever is purchased is indeed worth its cost; total costs may even rise rapidly due to the development and adoption of cost-effective interventions. Nonetheless, the higher the chosen threshold, the higher the healthcare cost growth rate will be as a greater variety of current and future interventions will be deemed cost-effective. For example, suppose that a $200,000 per QALY threshold is chosen instead of a $100,000 per QALY threshold. Any existing interventions whose cost-effectiveness lies between $100,000 and $200,000 will be deemed acceptable under the higher threshold, and manufacturers will find it attractive to invest in interventions likely to yield relatively high cost-effectiveness ratios. Therefore, if the threshold is reset optimally, it may actually decline over time in response to the higher spending encouraged by a higher threshold. As more cost-effective but cost-increasing interventions are developed and adopted under a high current threshold, the opportunity cost of further spending will rise. Therefore, the marginal QALY that seems affordable today may no longer seem affordable in the future, and it should not be assumed that the QALY threshold should necessarily rise in lockstep with inflation and incomes.

Given the possibility of monopoly pricing power for some healthcare services, a high threshold can also contribute to cost growth in a more subtle way. For example, if a new drug will be approved for payment if it meets the cost-effectiveness threshold, the patent-holder for that drug will have little reason not to price it in a manner that will place it very near the threshold. Therefore, setting

and enforcing a lower threshold can serve as a mechanism to force pricing restraint.

Future Directions

Most empirical evidence indicates that the arbitrary ROT thresholds in use today are too low. Continuing to use low thresholds may have beneficial cost control features but may also threaten the credibility of recommendations based on CEA. Given the desire for a decision rule for CEA, the appeal of a threshold is unlikely to diminish. Therefore, a consensus process such as that employed by the U.S. Public Health Services to standardize CEA methods could be used to develop and periodically update a threshold value and could also provide guidance to decision makers about how such a threshold could be adjusted to their particular contexts. Similarly, it has recently been suggested that the U.K. National Health Service establish an independent committee to develop a threshold. By making the basis for a threshold more transparent, such efforts should be encouraged.

Richard A. Hirth

See also Cost-Benefit Analysis; Cost-Effectiveness Analysis; Costs, Opportunity; Cost-Utility Analysis; Decision Rules; Health Status Measurement Standards; Quality-Adjusted Life Years (QALYs); Willingness to Pay

Further Readings

Appleby, J., Devlin, N., & Parkin, D. (2007). NICE's cost effectiveness threshold. *British Medical Journal, 335,* 358–359.

Byrne, M. M., O'Malley, K., & Suarez-Almazor, M. E. (2005). Willingness to pay per quality-adjusted life year in a study of knee osteoarthritis. *Medical Decision Making, 25,* 655–666.

Devlin, N., & Parkin, D. (2004). Does NICE have a cost-effectiveness threshold and what other factors influence its decisions? A binary choice analysis. *Health Economics, 13,* 437–452.

Garber, A. M., & Phelps, C. E. (1997). Economic foundations of cost-effectiveness analysis. *Journal of Health Economics, 16,* 1–31.

Hirth, R. A., Chernew, M. E., Miller, E., Fendrick, A. M., & Weissert, W. G. (2000). Willingness to pay for a quality-adjusted life year: In search of a standard. *Medical Decision Making, 20,* 332–342.

Murphy, K. M., & Topel, R. H. (2006). The value of health and longevity. *Journal of Political Economy, 114,* 871–904.

Neumann, P. J., Sandberg, E. A., Bell, C. M., Stone, P. W., & Chapman, R. H. (2000). Are pharmaceuticals cost-effective? A review of the evidence. *Health Affairs, 19,* 92–109.

Owens, D. K. (1998). Interpretation of cost-effectiveness analyses. *Journal of General Internal Medicine, 13,* 716–717.

Tengs, T. O., Adams, M. E., Pliskin, J. S., Safran, D. G., Siegel, J. E., Weinstein, M. C., et al. (1995). Five hundred life-saving interventions and their cost-effectiveness. *Risk Analysis, 15,* 369–390.

Ubel, P. A., Hirth, R. A., Chernew, M. E., & Fendrick, A. M. (2003). What is the price of life and why doesn't it increase with inflation? *Archives of Internal Medicine, 163,* 1637–1641.

Mood Effects

Among emotional influences in decision making, the concept of "mood" has always been of specific interest. Nevertheless, one of the most critical aspects in dealing with mood is its definition. Although contents are overlapping, *mood* only refers to the valence dimension of emotion and appears usually to be less intense. In contrast to affects, states of mood do not change rapidly and tend to last for longer periods of time—some authors even refer to depression as a state of negative mood. Taken together, studies examining this concept tend to lack a clear disambiguation to what states exactly they refer when talking about mood; a commonly accepted usage of the term is still to be achieved.

Applied to a medical environment, moods might play a distinctive role in several ways. Health specialists often have to make fast decisions under uncertainty—especially when the time frame is tight or previous knowledge is scarce. In these situations, contextual information is taken into account, including emotional states like mood. Findings show that there is an influence of specific moods enhancing or impairing cognitive processes involved in decision making. The following provides several examples of this.

Even highly experienced medical staff have to examine carefully the symptoms to give a

diagnosis—during this examination process, moods might play an important role. A direct impact of mood in decision making can be derived from findings focusing on the *abstraction level of information processing*. There exists strong evidence that individuals in a happy mood perceive incoming information in a more generalized way (focusing on more general aspects or characteristics) than those experiencing a sad mood who are normally concentrating on more specific aspects. Imagining a routine checkup with a health professional, mood might be an influential factor regarding the diagnosis. For example, it could be responsible for an underestimation of the patient's symptoms. On the one hand, patients in a good mood may not report specifically enough about their physical or psychological state—which could lead to difficulties for the physician to find the right diagnosis. On the other hand, a physician in a bad mood may focus on the specificity of the visitor's health aspects too hard and therefore lack the ability to grasp the bigger picture.

Another risky aspect comes from the fact that happy moods, in contrast to sad moods, are found to be related to a more *heuristic strategy of processing* incoming information. Despite the elaborated previous knowledge of a physician, a patient might be better off when having an appointment with a sad doctor. Sad moods are not only found to support the systematic or analytic elaboration of the actualities but also to avoid the (sometimes) inappropriate use of stereotypical thinking. Stereotypes are derived from the application of broader categories based on a general knowledge basis—and are therefore a result of heuristic processing. Moreover, applied to the context of medical decision making, the influence of a happy mood could be problematic for the examination of stigmatized patients: They might be judged by stereotypical expectations rather than objective criteria. In contrast, a sad mood leads to a more systematic analysis as the provided information undergoes an individualized elaboration. In line with these findings, research shows that prior general knowledge is more influential when individuals are experiencing a happy rather than sad mood. Hence, due to their prior beliefs, happy professionals run the risk of deciding in favor of a *confirmation bias* (i.e., accepting only findings confirming their assumptions) and disregarding any deviant information.

An explanation for the results about reliance on heuristics and stereotypes during happy moods is assumed by a lack of *motivation*. Based on the assumption that individuals in general try to maintain positive states and avoid negative ones, a negative mood might fulfill the need of changing the current situation. In contrast, a positive mood may not invite investment of any additional effort—however, motivation may enhance the effort, with the prospect of a positive outcome and enjoyment of the task.

The influence of mood in decision making has various aspects that are widely confirmed; naturally, it contains advantages and disadvantages regarding the outcome. Especially in the medical context, where decisions are affecting people's lives, any performance has to take the mood of the participating agents into account. To what extent medical decision making is influenced by mood has to be further examined. A matter of particular interest could lie in the impact mood has on previous knowledge when decisions and judgments of health professionals are required.

Stephanie Müller and Rocio Garcia-Retamero

See also Confirmation Bias; Decision Making and Affect; Diagnostic Process, Making a Diagnosis; Emotion and Choice

Further Readings

Allan, L. G., Siegel, S., & Hannah, S. (2007). The sad truth about depressive realism. *Quarterly Journal of Experimental Psychology, 60,* 482–495.

Bless, H., Mackie, D. M., & Schwarz, N. (1992). Mood effects on encoding and judgmental processes in persuasion. *Journal of Personality and Social Psychology, 63,* 585–595.

Bless, H., Schwarz, N., & Wieland, R. (1996). Mood and stereotyping: The impact of category and individuating information. *European Journal of Social Psychology, 26,* 935–959.

Bodenhausen, G. V. (1993). Emotions, arousal, and stereotype-based discrimination: A heuristic model of affect and stereotyping. In D. M. Mackie & D. L. Hamilton (Eds.), *Affect, cognition, and stereotyping: Interactive processes in group perception* (pp. 13–35). San Diego, CA: Academic Press.

Caruso, E. M., & Shafir, E. (2007). Now that I think about it, I'm in the mood for laughs: Decisions

focused on mood. *Journal of Behavioral Decision Making, 19,* 155–169.
Erber, R., & Erber, W. E. (2000). The self-regulation of moods: Second thoughts on the importance of happiness in everyday life. *Psychological Inquiry, 11,* 142–148.
Gasper, K., & Clore, G. L. (2002). Attending to the big picture: Mood and global versus local processing of visual information. *Psychological Science, 13,* 34–40.
Martin, L. L., & Core, G. L. (2001). *Theories of mood and cognition: A user's handbook*. London: Lawrence Erlbaum.
Morris, W. N. (1989). *Mood: The frame of mind*. New York: Springer.
Morris, W. N. (1992). A functional analysis of the role of mood in affective systems. In M. S. Clark (Ed.), *Emotion* (pp. 256–293). Newbury Park, CA: Sage.
Schwarz, N., & Clore, G. L. (1983). Moods, misattribution and judgements of well-being: Informative and directive functions of affective states. *Journal of Personality and Social Psychology, 45,* 513–523.

Moral Choice and Public Policy

This entry on moral choice and public policy in medical decision making focuses on the irreducible components of the physician–patient relationship as it is the foundational nature of that relationship that determines the moral character of decisions in healthcare. The focus on the physician–patient relationship is historically and fundamentally appropriate, but it is also dated. Medical decision making for the individual, as well as in public policy, is currently determined by additional professionals and entities. Professionals, such as nurses, have a larger role than ever before in the management of patients, both individually and collectively. Entities such as the government, insurance companies, hospitals, and health maintenance organizations also bear responsibility for decisions made for the individual and for the public. However, a discussion that is focused on the moral aspects of the physician–patient relationship can be a guide for judging the moral correctness of decisions made in healthcare for the entire spectrum of decision makers.

Moral Choice

The physician–patient relationship is at the heart of what it means to generate ethical medical decision making. The physician has many roles, including those of a technician, wage earner or entrepreneur, agent of public well-being, and advocate for public policy. These roles, however, do not constitute what it means to be a professional and a physician. What a physician professes, and that which is at the root of this relationship, is that the physician is obligated to place the welfare of the patient above all other considerations. This role as a healer remains as the irreducible character trait of the physician and is the moral foundation for ethical medical decision making. Note that the role of healer goes beyond that of technician to the body human because it encompasses consideration of the person's spiritual well-being also.

A second aspect of moral choice in medical decision making has been guided by the paradigm of patient autonomy. This view holds that the final authority for determining the treatment and direction of a patient's medical care lies with the patient or the patient's surrogate. The exercise of that right creates a responsibility for the physician to provide that the patient, or the patient's surrogate, is sufficiently informed and the resulting decisions are not coerced and are free of undue stress and self-interest. Furthermore, decision-making capacity is a developmentally regulated process, and among the determinants of that capacity are the age and developmental status of the individual, his or her relative health, the nature of the choice to be made, and the stress under which that person finds himself or herself. As this applies to ethical decision making in public policy, this would entail that decision making in the public arena requires an informed populace that is able to have an input into decisions made on their behalf. This input should be solicited in a noncoerced way that is free of undue stressors. Finally, not all decisions in healthcare carry a moral imperative. Certain questions, such as the appropriate antibiotics, are a technical matter, and other moral questions, such as priority listing for patients requiring solid organ transplantation in a country without the resources to perform such procedures, are not at issue on solely circumstantial grounds.

Public Policy

The physician has a responsibility to the public that can exceed his or her responsibility to the individual patient. When the patient is a risk to others, the physician's obligation supersedes the patient's wishes for privacy. Reporting of certain infectious diseases is one example. Hence, the right to confidentiality is not universal, and patients when they seek a physician's assistance are generally aware that certain conditions must be reported, even within the bounds of the physician–patient relationship. This is a part of the social contract we share. By accepting this social contract, individuals place their trust in the policy initiatives that society mandates to guard their best interests. The creation of healthcare policy is a result of this social contract. The morality of the decisions of that contract can be adjudicated by reference to the precepts of the physician–patient relationship.

Moral choice in medical decision making is determined by the character of the physician–patient relationship. The primary requirement for a decision to be considered ethical is that it places the well-being of the patient (or public) foremost and the welfare of the government or other entity after. Such decisions recognize the stake that the public has in health policy decisions and most adequately inform the patient or public, solicit input, and recognize and aim to reduce any undue stresses or coercion that might influence such input. Conflicts of interest will occur, and physicians occupy multiple roles at any given time. This is readily apparent in a society where a physician's action may be constrained by hospitals, managed care organizations, and the government. When such a conflict arises, it is the duty of the physician, or other decision maker, to recognize his or her primary role as a healer, with all other interests being secondary. A decision that is not in line with this fundamental role cannot be considered moral.

Criteria and Course of Action

Moral choice in public policy must recognize the selfsame criteria as moral choice at the bedside. In a question with a moral dimension, the criteria to recognize are as follows:

1. Moral choice in medical decision making is bound by having, as its primary objective, the well-being of the patient (or public).
2. The patient (or public) is an autonomous agent and should be
   a. informed,
   b. uncoerced, and
   c. free of undue stressors.
3. Conflict exists, and where it cannot be eliminated, it needs to be recognized and accounted for.

Public policy shifts and societal mores evolve with time, and therefore grounding moral choice for both public and private medical decision making is necessary to avoid discussions of morality that are subject to accusations of caprice. Medical decision making therefore has to foremost consider the needs and wishes of the patient or community in helping to define an appropriate course of action. This course of action must not violate certain foundational aspects of ethical medical decision making such as doing no harm and acting in a way that serves the interest of the patient, not the caregiver.

Robert K. Pretzlaff

See also Decision Making and Affect; Models of Physician–Patient Relationship; Shared Decision Making; Government Perspective, Informed Policy Choice

Further Readings

Beam, T. E., & Sparacino, L. R. (2003). *Military medical ethics*. Bethesda, MD: Office of the Surgeon General Uniformed Services University of the Health Sciences.

Loewy, E. H., & Loewy, R. S. (2004). *Textbook of healthcare ethics* (2nd ed.). Dordrecht, the Netherlands: Kluwer Academic.

U.S. Department of Health, Education, and Welfare. (1979). *The Belmont report: Ethical principles and guidelines for the protection of human subjects of research* (DHEW Publication No. (OS) 78-0012). Washington, DC: Government Printing Office.

Moral Factors

Moral factors are elements in decision making that relate to our relationships, as individuals and communities, with other people and the values used to

structure ways of living well. Moral factors pervade every medical decision but often go unnoticed because in standard practice there is consensus on how to balance competing moral factors best. These factors become most apparent in particularly troubling circumstances, such as a diminished mental capacity, limited resources, or vulnerability of subjects in research. Western medicine has a long tradition of attempting to enumerate and codify these factors to provide a practical guide for both ethical conduct and decision making. Attempts to do so include the Hippocratic oath, the American Medical Association's Code of Ethics, the Nuremberg Code, the Declaration of Helsinki, and the Belmont Report. Code-based approaches to understanding moral factors essentially rely on a snapshot of circumstances. Focusing on one set of values or a standard set of circumstances to the exclusion of others can create a blindness to other relevant factors. The consequences of losing sight of the diversity of moral factors can lead to a failure to identify a lack of consensus and the need for an evaluative moral calculus. Moral philosophers such as Aristotle, Kant, Mill, and Dewey each proffer theories on moral factors based on different understandings of the good life. In the end, moving from theory to practice requires that training, experience, and careful analysis be used for identification of relevant moral factors.

The sheer scope of moral factors necessitates that only a broad outline of kinds can be delineated in the current entry. The variety of relevant moral factors in any set of situations in a clinical environment can be analyzed at policy, care provider, and patient levels. These factors become more complex when treatment occurs in a research context. The sections below address three of the most interesting and important moral factors that manifest during considerations of complex medical choices. These include issues of autonomy and bodily integrity, clinical research, and nonmedical elements. Within each topic, the moral factors for various stakeholders are explored.

Autonomy and Bodily Integrity

A fundamental moral factor in any clinical medical decision is respect for a patient's choices. In modern medicine, there is a long-standing tradition of respecting a patient's autonomy—his or her right to self-determination. A competent adult patient has the right to accept or refuse available medical options based on his or her own valuing of relevant moral factors. Generally, performing medical interventions on a competent adult without his or her consent violates the patient's right to bodily integrity. Legally and morally, there is no substantive difference between performing an unwanted medical procedure on a competent patient and any other kind of assault. This right to self-determination entails certain moral and legal obligations on the part of healthcare professionals. Because a patient must have a sufficient understanding of the medical situation to exercise his or her autonomy, healthcare professionals have an obligation to disclose all relevant information necessary for the patient to provide informed consent (see *Canterbury v. Spence,* 1972), and they are generally obligated not to subject patients to medical treatments or procedures against their wishes. In other words, patients have the right to refuse treatment, even life-preserving treatment. The classic example of this is the Jehovah's Witness refusing blood products (see *Stamford Hospital v. Vega,* 1996). Competent adults have an equal right to withdraw an unwanted treatment as they do to refuse its initiation (see *In re Quinlan,* 1976; *Cruzan v. Director, Missouri Department of Health,* 1990).

A patient's right to self-determination does not necessarily entail a right to demand all and any intervention. While the right of a competent well-informed adult patient to refuse treatment is virtually limitless, there are significant limits on the positive demands a patient can make for medical interventions. A patient cannot demand a healthcare professional to engage in any activities that violate standards of medical practice. Respecting patients' autonomy does not include healthcare professionals abdicating their obligation to act in accordance with their professional judgment. Healthcare professionals have the right to refuse to participate in any procedure that violates their moral conscience. However, healthcare professionals may not abandon their patients. If they are unwilling to perform a particular procedure, then generally they must assist the patient in finding adequate care. In this way, a physician's professional and moral stance may be a significant moral factor. Finally, patients do not have a right to demand scarce resources, such as solid organs, to

which they are not otherwise entitled. The moral factor related to needs of others and the community must be considered.

Not every patient, though, is a competent adult. In the case of adult patients who lack the ability to make their own decisions, an attempt is still made to respect their wishes. In such instances, a surrogate decision maker, chosen by the patient, the courts, or state statute, makes decisions based on what the patient would have done if the patient had not lacked the capacity. In the case of children or adults who have never been in a position to express their own set of values, medical decisions are made based on the best medical interests of the patient. It is important to understand that the right to determine the course of one's medical treatment and to refuse or withdraw unwanted treatments is not restricted to competent adults. Since moral factors of control of bodily integrity and autonomy are valued highly, every person regardless of his or her cognitive and developmental abilities should be allotted control commensurate with his or her goals, values, and articulation of wishes.

Clinical Research

The research endeavor interjects a variety of new moral factors into the decision-making process. In clinical research, the physician–patient dyad expands to include at least the researcher, the funding source, and the population that the research aims to help. Physicians may find themselves playing a dual role. As physicians, they are primarily concerned with the patient immediately before them. As researchers, they value elements such as uniformity, repeatability, quality of data, and statistical significance directed toward understanding something about an aggregate of patients. This duality creates a morally relevant fact for both the patient and the physician in attempting best decision making. A natural conflict of interest may arise in interjecting various research-related moral factors since best clinical practice may not be best for scientific advancement. The relationship adds the researcher-subject dimension to the existing physician–patient interaction. This conflict is exacerbated when the relationship puts undue influence on the patient to participate in research.

Beyond the inherent difference of values in research, moral factors include the types of subjects being recruited. Elements such as a subject's medical condition, cognitive ability, age, and membership in an identifiable vulnerable population all become important moral factors because they affect the subject's ability to consent voluntarily to be a subject, or they affect the burdens, risks of harm, and likelihood of benefits for that subject. The burdens, harms, and benefits faced by the subject must be balanced against the clinical options available to the subject and the overall value of the research for the patient population.

A plethora of further important moral factors should be considered in clinical research, including things such as appropriate levels of oversight, peer review, limiting patient liberty, dissemination of information, degree to which research may be useful, and ownership of findings. These are among many additional factors that must enter into good decision making when patient care includes some aspect of clinical research. With every new activity added onto the patient–physician relationship, there are new moral factors to be considered.

Nonmedical Considerations

The above discussion of moral factors focuses on medical considerations as they affect particular patients. While in theory the only relevant interests should be those of the particular patient, in practice the interests of a multitude of persons connected to the patient may be affected. These include family members, friends, healthcare providers, and in some cases the community at large. One of the most prevalent effects on persons other than the patient is the emotional affects that treatment or withdrawal of treatment can have on those close to a patient. In difficult end-of-life choices, family concerns are often included more fully as relevant moral factors. For instance, the decision to withdraw life-sustaining treatment may be delayed to allow family members time to "adjust" to a situation. Furthermore, healthcare professionals are not immune to the emotional costs of caring for patients. In particular, nurses who are responsible for the daily care of patients can suffer from moral distress and become overwhelmed by the suffering of their patients when little meaning is attached to that suffering. These providers are moral agents, and their concerns should be weighed in some degree as appropriate moral factors.

Treatment decisions can affect third parties in many ways beyond the emotional. For instance, there are financial considerations, fiduciary obligations (children, work, other patients), and community resources. The decision to treat or refrain from treating can affect future burdens of care. The collateral cost to others is exemplified by the controversial case reported in 2006 in the *Archives of Pediatrics & Adolescent Medicine* regarding a severely disabled girl, known as Ashley X, who underwent surgery to attenuate her growth and sexual maturation. Part of the discussion regarding the parents' treatment decisions revolved around their desire to avoid burdens.

Finally, as a community, explicit and implicit rationing occur in healthcare. Since there are finite resources, instances arise where the proper allocation becomes a morally relevant factor. This might be the use of blood in transfusion, an ICU bed, a specialist's time, or entry into a research protocol.

Final Thoughts

Moral factors become particularly prevalent and critical in medical decisions because their consequences often have far-reaching effects on quality of life. Medical decisions force individuals and communities to confront core values in unaccustomed ways. The inherent uncertainty in medical practice exacerbates the decision making in such a way that accounting and weighing of moral factors is always probabilistic as well as subjective. Even when a particular moral factor is given priority in decision making, the patient, physicians, or community may still not preserve what was intended because the medical condition may follow an unexpected path.

Jason Gatliff and Paul J. Ford

See also Beneficence; Bioethics; Cost-Benefit Analysis; Cultural Issues; Decisions Faced by Surrogates or Proxies for the Patient, Durable Power of Attorney; Disability-Adjusted Life Years (DALYs); Distributive Justice; Disutility; Equity; Government Perspective, Public Health Issues; Informed Consent; International Differences in Healthcare Systems; Medical Decisions and Ethics in the Military Context; Motivation; Patient Rights; Quality-Adjusted Life Years (QALYs); Rationing; Risk Aversion; Social Factors

Further Readings

Canterbury v. Spence, 464 F.2d 772, 775 (D.C. Cir. 1972).
Copp, D. (Ed.). (2006). *Oxford handbook of ethical theory*. New York: Oxford University Press.
Council on Ethical and Judicial Affairs, American Medical Association. (2006). Capital punishment. In *Code of medical ethics: Current opinions with annotations 2006–2007*. Chicago: American Medical Association.
Cruzan v. Director MDH, 497 U.S. 261 (1990).
Graham, D. (2000). Revisiting Hippocrates: Does an oath really matter? *Journal of the American Medical Association, 284*(22), 2841–2842.
Gunther, D. F., & Diekema, D. S. (2006). Attenuating growth in children with profound developmental disability: A new approach to an old dilemma. *Archives of Pediatric Adolescent Medicine, 160,* 1013–1017.
In re Quinlan, 70 N.J. 10, 355 A.2d 647 (1976).
McGee, G. (2003). *Pragmatic bioethics*. Cambridge: MIT Press.
Stamford Hospital v. Vega, 236 Conn. 646, 663, 674 A.2d 821 (1996).
U.S. Department of Health, Education, and Welfare. (1978). *The Belmont Report: Ethical principles and guidelines for the protection of human subjects of research* (DHEW Publication No. (OS) 78-0012). Washington, DC: Government Printing Office.
U.S. Government Printing Office. (1949). *Trials of war criminals before the Nuremberg military tribunals under control council law* (No. 10, vol. 2, pp. 181–182). Washington, DC: Author.
World Medical Organization. (1996). Declaration of Helsinki. *British Medical Journal*, *313*(7070), 1448–1449.

Morbidity

Morbidity refers to the absence of health, including physical and psychological well-being. In medical decision making, one cares about morbidity in the following three constructs: (1) the baseline incidence or prevalence of morbidity in the absence of an intervention; (2) the decrease in morbidity in the presence of an efficacious, effective, and cost-effective prevention or treatment intervention; and, in some cases, (3) the increase in morbidity because of an intervention, in terms of adverse events.

Morbidity is commonly measured by the *incidence* or *prevalence* of disease or injury. Incidence is the number of new cases of disease or injury in a population over a period of time. Incidence is a rate and describes the probability that healthy or disease-free people will develop the disease or injury over a specific time period, oftentimes a year. Thus, incidence is frequently used to measure diseases with a rapid onset, such as infectious diseases. If a city with a population of 100,000 persons has 500 new cases of influenza over a period of 3 months, then the incidence rate for this city would be 500 cases per 100,000 per 3 months. By comparing incidence rates of disease or injury among population groups varying in one or more identified factors (such as race/ethnicity or gender), health practitioners can make better informed decisions regarding the use of scarce health resources. For example, if a school finds that children living in a particular zip code are five times more likely to have dental caries than do other students, then public health officials can strategically intervene to prevent further morbidity.

Prevalence is the number of persons in a population who have a disease or injury at a given time, regardless of when the disease first occurred. Prevalence is a proportion and measures the probability of having the disease at a point in time. In decision making, prevalence is important in determining human resource needs, workload constraints, and the utilization of healthcare resources associated with particular diseases and injuries. It is also useful for expressing the burden of some attribute, disease, or condition in a population. An awareness that 25% of Hispanic adults aged 50 years or older have diabetes in a geographic area next to a particular hospital, for example, indicates how pervasive this disease is and can guide decision makers at that hospital to effectively allocate resources. Unlike incidence, prevalence is not a rate and therefore should not be used to measure diseases with rapid onsets or infectious diseases.

The relationship between prevalence and incidence is another important aspect of how these measures describe morbidity in a population: Prevalence = Incidence × Duration. Thus, treatments that prolong life and increase duration of disease, such as insulin for diabetics, have a profound effect on disease prevalence. Furthermore, an increased incidence of diabetes contributes to a rapidly increasing prevalence of diabetes in the United States.

In medical decision making, prevalence can also be used to describe the baseline population in the absence of an intervention and to serve as an outcome variable. Treatment and prevention strategies can affect the prevalence of disease by reducing the length of time an individual might suffer from disease or by affecting the development of new cases or both. The use of Pap smears to identify preinvasive and invasive cervical cancer is an example of an intervention that can reduce cervical cancer prevalence by reducing the length of time a woman would suffer from cervical cancer as well as by preventing the development of new cases.

In addition to incidence and prevalence, morbidity in a population can also be measured by impacts on disability or functional capacity, quality of life, life expectancy, or costs. *Disability* describes the temporary or long-term reduction in an individual's functional capacity that is associated with morbidity. For example, impairments to activities of daily living (or ADLs) may be associated with a morbidity such as a broken hip or Alzheimer's disease.

Quality of life (QOL) is another measure used in medical decision making to address morbidity. QOL is a multidimensional measure of the physical, emotional, cognitive, and social impacts of disease, treatment, or sequelae of an injury or disability. QOL is often measured using standardized, validated instruments completed by patients or the general public to determine preferences for specific morbidity or health states under consideration. QOL instruments can either be generic ones that can be used by all individuals, such as the EuroQol 5-Dimension (EQ-5D) or Short Form-36 (SF-36), or disease-specific ones that provide additional details on the disease of interest.

The impact that morbidity may have on *life expectancy* is also important in medical decision making. For example, in the early years of the HIV epidemic, the average life expectancy for HIV-positive persons was less than 10 years. With the development of effective HIV treatment, however, life expectancy is now less affected by HIV morbidity. When combined with QOL, morbidity's impact on life expectancy can be determined by quality-adjusted life expectancy (QALE), which is another key outcome measure for use in medical

decision making. In particular, the quality-adjusted life year, or QALY, is the main effectiveness measure for use in cost-effectiveness analyses of health interventions.

In some cases, treatments aimed at decreasing morbidity in the long run may have adverse consequences and cause morbidity in the short term. For example, chemotherapy treatment for cancer patients can result in a temporary decrease in QOL and an impairment in ADLs in the short term. In this instance, medical decision makers must balance the trade-off between increased short-term morbidity with the expected decrease in long-term morbidity.

Other measures of morbidity used in medical decision making include the impact of morbidity on *costs,* such as increased healthcare utilization, including emergency department visits, hospitalizations, and prescription drugs, and losses in productivity. These measures are typically summarized in cost of illness (COI) analyses. For example, a COI analysis associated with child maltreatment would include the marginal increased utilization of health and mental health services, increased special education costs, costs to the criminal justice and child welfare systems, and reductions in educational and employment outcomes associated with the morbidity resulting from maltreatment. COI analyses may incorporate the economic impact of either incidence-based morbidity or prevalence-based morbidity, depending on whether the decision maker is interested in knowing the lifetime costs associated with the morbidity (incidence-based) or the annual costs associated with the morbidity (prevalence-based).

Regardless of the approach, the perspective of the decision maker affects how morbidity is addressed in medical decision making. For example, in a COI analysis where the decision maker has a societal perspective, productivity losses and other economic losses, or opportunity costs, to society would be included. That is, all morbidity costs, regardless of to whom they accrue, would be included in the analysis. From a healthcare system perspective, however, costs of morbidity would likely include only the value of healthcare resources required to treat or prevent the disease or injury and not the societal losses associated with decreased productivity. In this case, economic costs, or costs of activities that were forgone due to morbidity, would not factor into the decision model.

Finally, *comorbidity* describes the effect of all other diseases or injuries a patient has other than the primary disease or injury of interest. Thus, comorbidity refers to the coexistence of two or more disease processes, which may have a substantial impact on medical decision making. For example, a diabetic who also has heart disease is considered to have a comorbid condition (heart disease) that must be accounted for in the treatment and prevention of the primary condition (diabetes). This comorbidity may affect the effectiveness and costs of treatments and the overall morbidity of this diabetic and is therefore a critical component of medical decisions. For example, the presence of substantive comorbidities can make it difficult for clinicians to diagnose diseases and can contribute to complications in treating disease, both of which potentially result in increased costs and decreased effectiveness of treatment.

Phaedra Corso and Heather Edelblute

See also Mortality; Quality-Adjusted Life Years (QALYs)

Further Readings

Haddix, A., Teutsch, S., & Corso, P. S. (Eds.). (2003). *Prevention effectiveness: A guide to decision analysis and economic evaluation.* New York: Oxford University Press.

Kleinbaum, D. G., Sullivan, K., & Barker, N. (2007). *A pocket guide to epidemiology.* New York: Springer.

Mortality

Mortality data are generally collected by vital statistics agencies. There is a long tradition of using mortality data to evaluate the burden of diseases on a population basis. From these data, it is possible to build key mortality indicators that may help to detect important public health problems and inequalities in health among regions, countries, or subgroups of populations.

It is also possible to represent mortality data using maps showing the geographical patterns of mortality that might be useful for epidemiologists and public health researchers to formulate etiologic hypotheses or for public health policy

makers to make decisions concerning allocation of health funds.

Simple Mortality Indicators

The simplest mortality indicator is the crude or total mortality rate (CMR), which is defined as the number of deaths per 100,000 population.

Suppose one wants to calculate the CMR among females in Spain in the period from 1994 to 1998, where one has observed a total of Y deaths and N is the number of person-years in the period (calculated by multiplying the number of females in the Spanish population by the observation period of 5 years). The CMR per 100,000 population per year is $Y/N \times 100{,}000$.

To calculate the CMR in Spain in 1 year per 100,000 population one need only divide the number of deaths in that year by the midyear population and then multiply the result by 100,000, that is,

$$\text{CMR} = (\text{Number of deaths in one year}/\text{Midyear population}) \times 100{,}000.$$

When the cause of death is unspecified, it is referred to as global mortality or mortality by all possible causes. However, although global mortality has an interest per se as an overall indicator of mortality, epidemiologists and public health researchers are also interested in classifying mortality data by specific cause of death. The *International Classification of Diseases*, 10th revision (ICD-10), is used to code and classify mortality data from death certificates. The use of a standard and well-designed coding system facilitates the comparison of results from different sources. The ICD-10 is copyrighted by the World Health Organization (WHO), which owns and publishes the classification. The ICD-10 was adopted in 1990 and came into effect in 1993. Unfortunately, many countries have been using the previous ICD-9 classification until very recently.

Mortality rates are often age standardized. The underlying reason is that two populations that have the same age-specific mortality rates for a specific cause of death will have different crude mortality rates if the age distributions of the two populations are different. To standardize rates, one may use a direct or an indirect standardization method. However, direct standardization is becoming very popular due to the increasing availability of age-specific mortality rates (AMRs). The AMRs are calculated as is the CMR but by restricting the numerator and denominator of the CMR to a particular age-group. For example, the AMR for age-group 35 to 54 years is the number of deaths that occur in that age-group per 100,000 persons of age 35 to 54. Suppose that in a certain country in the year 2000 there are 2,000,000 people in that age-group and 3,200 people die. Then, the AMR for that age-group is $(3200/2000000) \times 100000 = 160$ per 100,000.

It is of course possible to standardize (adjust) the rates for the effects of more than one variable. For example, one may want to adjust by gender and age.

Direct Standardization

The *age-standardized mortality rate* is the rate that would have been observed in a population with the same age structure of some reference population, called the standard population. This direct standardization method provides an age-adjusted mortality rate that is a weighted average of the age-specific rates. The weights are taken from the standard population. The direct standardized mortality rate can be calculated with the formula

$$\frac{\sum_j p_{js} d_j}{\sum_j p_{js}}, j = 1, \ldots, J,$$

where J is the number of age-groups, p_{js} is the standard population in the jth age-group (percentages or counts), and d_j is the AMR in the jth age-group.

Example

One could calculate from Table 1 the direct standardized mortality rate for Country A:

$$\frac{26.2 \times 190 + 32.2 \times 180 + 25.2 \times 630 + 13.4 \times 3260 + 3 \times 20390}{26.2 + 32.2 + 25.2 + 13.4 + 3} = 1315.4 \textit{ per } 100{,}000 \textit{ population}.$$

Table 1 Population (in percentages) and age mortality rates in Country A, and standard population (artificial)

Age-Group (Years)	*Population (%)*	*AMR per 100,000*	*Population (%)*	*AMR per 100,000*
0–14	21.0	190	26.2	220
15–34	27.9	180	32.2	310
35–54	29.5	630	25.2	600
55–74	15.5	3,260	13.4	3,000
75+	6.1	20,390	3	19,900

Age-standardized mortality rates are comparable only if they are calculated using the same standard population. WHO has proposed a standard population based on the mean world population age structure projected for the period 2000 to 2025. The use of this standard world population by different countries will permit comparison of mortality rates among them. Sometimes vital statistics agencies are interested in comparing age-specific rates within a country, and then the national population is used as the standard.

Indirect Standardization

The *standardized mortality ratio* (SMR) is calculated as the total number of mortality cases in the study population divided by the expected number of cases. The expected number of cases is obtained by applying the age rates of the standard population to the age structure of the study population. In other words, one is computing the expected cases considering that the study population has the same rates as the standard population.

The main advantage of the SMR is that it involves only the total number of mortality cases, so one does not need to know in which age categories the mortality cases occur in the study population. Another practical advantage is that the SMR does not tend to be sensitive to numerical instabilities in one or two of the age-specific rates.

An important note of caution for practitioners is that generally it is not possible to compare SMRs in different populations. A single SMR compares the observed number of mortality cases in the study population with the expected number of cases computed using a standard population. What is interesting is to test if the SMR is significantly different from 1 or, equivalently, to calculate confidence intervals for the SMR. A value of the SMR greater than 1 means that an excess of mortality in the study population is observed when compared with the standard population. Often, SMRs are presented as percentages by multiplying them by 100.

However, there is one case where different SMRs are comparable. Suppose there are several exposure categories. If the stratum-specific mortality rates for each exposure class are proportional to the external standard rates, then it is possible to compare the SMRs of the different categories.

Example

Suppose one has observed 1,000 mortality cases in metal workers in a certain region (for data, see

Table 2 Data needed to calculate the standardized mortality ratio (SMR) in metal workers (invented)

Age-Groups (Years)	*Number of Metal Workers*	*Expected Cases*	*AMR in the National Population per 100,000*
25–34	40,000	80	200
35–44	30,000	120	400
45–54	20,000	140	700
55–64	10,000	300	3,000
	100,000	640	

Table 2). Since the national mortality rate for the age-group 25 to 34 is 200 per 100,000 population, the number of expected cases for that particular age-group is $40000 \times (200/100000) = 80$. One can calculate the rest of the expected cases in a similar way. The total number of expected cases is 640. Then, the SMR = 1000/640 = 1.5625. This means that the mortality level of metal workers in that region is 56% higher than the mortality experienced by the national population.

Possible Drawbacks of Standardized Rates

It is not appropriate to calculate the age-standardized mortality rate as the single measure of mortality for a given region in a certain period. We would then be concentrating on a single measure all the information concerning mortality, and, as a consequence, a lot of information could be lost. For example, with certain diseases, one may expect that pollution effects in a region will have different consequences for different age-groups. Then, the age-specific rates should always be the starting point of any thorough analysis of mortality data.

Mapping Mortality Data

Maps displaying the geographic distribution of mortality or disease incidence have several important functions: For example, they are used by epidemiologists to identify factors that may be linked to various causes of mortality, or they may be used by policy makers for the purposes of allocation of health funding. In the latter case, it is usually of interest not only to obtain a smoothed picture of mortality risks across the region being studied but also to pinpoint areas that appear extreme.

Prior to the publication of the first atlases of mortality rates, many epidemiologists questioned the utility of mapping rates. However, it is now clear that geographical patterns of mortality could not be discovered from lists of mortality rates. The first U.S. cancer atlas identified a strong clustering of high oral cancer rates in the southern part of the United States that a posterior epidemiological study found to be due to snuff dipping. This first atlas showed that mapping small-area mortality rates is a valuable public health tool for generating etiologic hypotheses and identifying high-rate regions where intervention efforts might be warranted.

Some atlases have displayed measures of relative risk, usually SMRs, while others have represented the statistical significance of local deviations of risks from the overall rates. However, mapping raw or direct measures, such as SMRs, is not very reliable and has been criticized as they are usually highly variable in thinly populated areas or when the number of observed counts is very small. So, imprecise estimates of the SMRs might be dominating the geographical pattern. This is particularly evident when the focus of the investigation is on mortality caused by a rare disease. To address this problem, statistical models discussed in the recent literature have explored ideas surrounding smoothing methods, which consist in pooling information across the regions under study to provide mortality ratio estimators that are more stable.

M. Dolores Ugarte

See also Confidence Intervals; Hypothesis Testing; Life Expectancy; Morbidity; Odds and Odds Ratio, Risk Ratio

Further Readings

Ahmad, O. B., Boschi-Pinto, C., Lopez, A. D., Murray, C. J. L., Lozano, R., & Inoue, M. (2000). *Age standardization of rates: A new WHO standard* (GPE Discussion Paper Series: No. 31, EIP/GPE/EBD). Geneva: World Health Organization.

Besag, J., York, J., & Mollié, A. (1991). Bayesian image restoration with two applications in spatial statistics. *Annals of the Institute of Statistical Mathematics*, *43*, 1–59.

Breslow, N. E., & Day, N. E. (1987). *Statistical methods in cancer research: Vol. 2. The design and analysis of cohort studies* (Scientific Publication No. 82). Lyon, France: IARC.

Clayton, D., & Kaldor, J. (1987). Empirical Bayes estimates of age-standardized relative risks for use in disease mapping. *Biometrics*, *43*, 671–681.

Inskip, H. (1998). Standardization methods. In P. Armitage & T. Corton (Eds.), *Encyclopedia of biostatistics* (Vol. 6, pp. 4237–4250). New York: Wiley.

Mason, T. J., McKay, F. W., Hoover, R., et al. (1975). *Atlas of cancer mortality for U S. counties: 1950–1969* (DHEW Pub. No. [NIH] 75-780). Washington, DC: Government Printing Office.

Mollié, A. (1996). Bayesian mapping of disease. In W. R. Gilks, S. Richardson, & D. J. Spiegelhalter

(Eds.), *Markov chain Monte Carlo in practice*. London: Chapman & Hall.

Ugarte, M. D., Ibáñez, B., & Militino, A. F. (2005). Detection of spatial variation in risk when using CAR models for smoothing relative risks. *Stochastic Environmental Research and Risk Assessment, 19*(1), 33–40.

Ugarte, M. D., Ibáñez, B., & Militino, A. F. (2006). Modelling risks in disease mapping. *Statistical Methods in Medical Research, 15,* 21–35.

van Belle, G., Fisher, L. D., Heagerty, P. J., & Lumley, T. (2004). *Biostatistics: A methodology for the health sciences*. New York: Wiley.

Winn, D. M., Blot, W. J., Shy, C. M., Pickle, L. W., Toledo, A., & Fraumeni, J. F. (1981). Snuff dipping and oral cancer among women in the southern United States. *New England Journal of Medicine, 304,* 745–749.

Motivation

The health service industry is undergoing massive transformations due, in part, to advances in technology and the metamorphosis occurring in the demographics and diversity of the workforce. Changes in healthcare will likely continue at an accelerated pace, and with these changes the need for training will become even more important. The use of technology in training has led to heightened access, faster distribution, innovation, and increased collaboration. However, this increase of technology implies challenges in keeping up with the latest developments in technology, an increased pace of training, depersonalization, and fear of the unknown. With these challenges, understanding *what motivates health workers,* as well as *how leaders motivate the entire organization,* becomes essential to promote a proper work environment. This entry examines the conceptual issues and empirical research concerning motivation at the workplace in health organizations. First, motivation as a concept is explained. Second, a brief overview of one of the main theories of motivation is provided and applied to the challenge of increasing motivation in health workers. Finally, several relevant results are described reviewing the literature on the topic of effective leadership styles to increase motivation.

Maslow's Theory

There is a general consensus that motivation is an internal state or condition—sometimes described as a need or desire—that serves to activate or energize behavior and give it direction. One of the most influential authors in the area of motivation is Abraham Maslow, who attempted to synthesize a large body of research related to the topic. Prior to Maslow, researchers generally focused separately on factors such as biology, achievement, or power to explain what energizes, directs, and sustains human behavior. Maslow posited a hierarchy or pyramid of human needs divided into five levels. At the bottom of the pyramid is the physiological level, which includes food, water, and shelter—the most basic needs for human survival. The premise is that unless an individual's basic needs have been met, higher levels in the pyramid are of no relevance as survival is the most basic human component. When the basic survival needs have been met, the individual aspires to the next level seeking safety, including freedom from anxiety and stress. Stress, unfortunately, appears to be a constant in our culture. Although anxiety and stress may be a constant, one needs to look at these elements on an individual basis and on a continuum from moderate to extreme. Once the stress level has been moderated, or is acceptable, for the individual and basic safety conditions have been met, the individual would then look to the third stage, which is identified as the social level. This level includes the need for belongingness, friendship, and love. Having obtained relationship/belongingness security, it is possible to look to the fourth level, self-esteem. In this arena, the individual seeks to feel competent, confident, and self-assured. Finally, having accomplished the needs on all four levels, the individual is able to pursue self-actualization—or to "be all that one can be."

Maslow's hierarchy of needs model provides a means for motivating employees in a rapidly changing healthcare industry. In a work setting, the first basic level that must be satisfied for an employee to be motivated is that of wages. Maslow posited that basic survival needs can be fulfilled with wages because money is an equivalent to shelter, food, water, heat, clothing, and so forth. Once the survival needs are covered in the form of adequate wages, the individual seeks safety on the job. This

includes not only physical, but also mental, safety, implying a decrease in anxiety. Training plays an important role at this level because the worker consciously and subconsciously relates training to safety. When the terms for safety on the job have been met, individuals aspire to satisfy the third level: social belongingness in the workplace. We seek pleasant working relationships with coworkers, peers, and others; it is important to us to find our place in formal and informal work groups. The social needs wax and wane on the strength of our personal relationships and our participation with others in the organization. Training provides the individual with additional opportunities to meet people in the work environment, to discover others with the same interests or job responsibilities, and to establish new lines of communication. The fourth level, self-esteem in the organization, is generally based on the individual's successful performance appraisals, incentives, rewards received, and recognitions obtained—which all are related to self-confidence. To enhance self-esteem, training provides a possible source to feel and actually be more productive and confident in the work environment. In turn, as the individual's confidence level grows, there is greater opportunity to obtain rewards, recognition, and positive performance appraisals. Finally, training allows the individual to move toward self-actualization; to develop one's potential, to learn new things, to take risks, and to feel even more confident in what one does.

Health employees could be motivated in the face of increased demands by making them feel secure, needed, and appreciated. This is not easy at all, but leaders would be able to enhance employees' motivation and commitment through training that covers two main aspects: (1) the needs of the individuals and (2) the demands of new technologies that provide challenges and opportunities for meeting those needs. The question that remains open is which leadership styles will have a stronger impact on employees' motivation.

Effective Leadership Styles

In today's world, the role of a leader is changing from being that of one who controls workers to achieve results to the role of one who visualizes the future and empowers workers with the necessary skills and behaviors to be successful. To control others, a leader must be perceived as having authority, power, and influence. In contrast, today's effective leadership focuses more on influence and less on power and authority. Researchers have compared the effectiveness of several leadership styles in health organizations. Transformational and transaction leadership styles are two representative examples.

Transactional leadership involves creating and clearly communicating employee expectations along with identifying rewards and punishments. *Transactional leadership* thus implies trading rewards for achieving specific goals. The transactional leader knows about the needs and wants of the employees and offers these needs and wants as rewards for performance. While some researchers indicate favorable organizational outcomes with the use of contingent rewards, employees are extrinsically motivated to perform only at the minimum levels that are required for achievement. There is no encouragement for employees to go above and beyond expectations or to try innovative solutions (e.g., creative thinking), which would be crucial in the healthcare system. There are other difficulties that stem from a reward system based on achievements of goals. One difficulty arises from the tendency of employees to focus on the actual reward rather than on the quality of the work and eventually become immune to rewards (i.e., much larger rewards are needed to fulfill the same expectations). This leadership style can produce positive short-term outcomes; however, it rarely produces long-term results because it can force competition between coworkers, decrease teamwork, and even destroy any intrinsic motivation, which leads to a decrease of the quality of work performed by employees. In healthcare organizations, transactional leaders would motivate their employees by offering rewards or incentives for successful completion of tasks and would clearly communicate errors incurred by the staff.

In contrast to transactional leadership, *transformational leadership* theory is founded on the ability of the leader to raise awareness of the organization's vision and mission in terms such that employees believe in the organization's needs and put them ahead of their own needs. A transformational leader achieves this alignment through the skillful art of articulating visions. That way, employees do not only accept but also take ownership of the

visions. Because employees are empowered to take ownership of the visions, they perceive the organization's success as their own. Transformational leadership theory focuses on transforming employees' own values and beliefs so that they can expand and elevate their goals to perform at a higher level. The transformational leader works at instilling trust, admiration, loyalty, and respect with each employee. It is important that employees believe that the leader has also taken ownership of the organizational visions and is putting all efforts into attaining these goals. Leaders who are intrinsically motivated exhibit a much deeper sense of enjoyment and purpose while they are performing their jobs. Enjoyment, along with other positive emotions and arousing attitudes displayed by the leader, positively influences the motivation of employees. In healthcare organizations, transformational leaders would motivate their staff by clearly articulating the department's mission and vision and explain how those factors give meaning and satisfaction to what the staff does. The importance of ownership and delivery of quality patient care by each employee would be highly stressed.

Empirical studies contrasting these leadership styles suggest that transformational leadership is very effective. For instance, in a study conducted by Dong Jung, two groups, one with a transactional leader and the other with a transformational leader, were compared according to their creativity levels. The group with the transactional leader formed significantly fewer creative and unique ideas compared with the group with the transformational leader. Jung also indicated that transformational leadership behaviors cultivate a greater level of creativity and flexibility in followers' thought processes. As a conclusion, Jung stated that this is achieved from the followers' focus on the intrinsic rewards and feeling of satisfaction and increased self-esteem brought about by the achievement of their goals.

In a similar vein, Megan Joffe and Sean Glynn described transformational structural changes in a global pharmaceutical company. Changes were needed to increase employee morale and job satisfaction as well as decrease the high staff turnover rate. The leadership in the organization wished to empower the employees and allow them to participate in the changes, giving them ownership of the organization's direction and values. In their study, a new mission and vision were developed, and the role of the scientists was incorporated clearly into the new vision and goals of the organization. Some key components of the changes focused on leadership, clear communication, identification with the organization, and feelings of value and recognition. Employees were encouraged to take risks and to be creative in their ideas, activities, and actions. After a year of introducing those changes, 90% of employees had taken the initiative in identifying areas of improvement, offering new insights on how to improve these areas, and aiding in the implementation of the solutions. Employees were given ownership, were regarded as important, and were supported in their roles. All these things helped to intrinsically motivate them to achieve extraordinary things in a time of struggle and upheaval.

Research has revealed that motivation in today's workplace is an important factor in determining an organization's success. Therefore, the ability of leaders to influence and motivate employees is crucial to the health of the organization. Transformational leaders, who motivate employees by transforming their beliefs and values to be more in alignment with the organization's values and goals, help create higher levels of intrinsic motivation and are very effective in contemporary healthcare organizations.

Rocio Garcia-Retamero and Stephanie Müller

See also Decision Making and Affect; Heuristics; Patient Rights; Patient Satisfaction; Shared Decision Making; Trust in Healthcare

Further Readings

Bass, B. M. (1998). *Transformational leadership: Industry, military, and educational impact*. Mahwah, NJ: Erlbaum.

Benson, S. G., & Dundis, S. (2003). Understanding and motivating health care employees: Integrating Maslow's hierarchy of needs, training and technology. *Journal of Nursing Management, 11,* 315–320.

Ilies, R., Judge, T., & Wagner, D. (2006). Making sense of motivational leadership: The trail from transformational leaders to motivated followers. *Journal of Leadership and Organizational Studies, 13,* 1–22.

Joffe, M., & Glynn, S. (2002). Facilitating change and empowering employees. *Journal of Change Management, 2,* 369–379.

Jung, D. I. (2001). Transformational and transactional leadership and their effects on creativity in groups. *Creativity Research Journal, 13,* 185–195.

Kalar, T., & Wright, D. K. (2007). Leadership theory and motivation of medical imaging employees. *Radiology Management, 29,* 20–24.

Maslow, A. H. (1954). *Motivation and personality*. New York: Harper.

Maslow, A. H. (2000). *The Maslow business reader* (D. C. Stephens, Ed.). New York: Wiley.

Rudnick, J. D. (2007). Transformational leadership. *Health Progress, 88,* 36–40.

Strickler, J. (2006). What really motivates people? *Journal for Quality and Participation, 29,* 26–28.

Multi-Attribute Utility Theory

Health-related quality of life (HRQL) is a critical element in the eventual outcomes of medical care and public health. Measuring HRQL successfully involves adequate description of health states, distinguishing differences among groups, and detecting change in individuals over time. There are two major approaches to HRQL measurement: psychometric and utility. Psychometric instruments like the SF-36 measure HRQL from a *descriptive* point of view to capture the various dimensions and generate a health profile. Alternatively, with the utility approach, one measures people's *values* for health states, also called *preferences* or *utilities*. Utilities are measured on a scale where 0 = *death* and 1.0 = *perfect or optimal health*. Utilities have a basis in economic theory and decision science and so are useful for calculating the quality-adjusted life years (QALYs) used in cost-utility analysis. Because HRQL is inherently multidimensional (vision, hearing, cognition, mobility, etc.), an extension of utility theory called multi-attribute utility theory (MAUT) has been applied to HRQL in the explicit multidimensional or multi-*attribute* sense. This entry focuses on methods derived from MAUT. The health utilities index (HUI) is the best known example of a MAUT-based HRQL measure.

Basic health utility measurement is implicitly multidimensional. As is detailed elsewhere, *direct* methods like the visual analog scale (VAS), Standard Gamble (SG), and Time Trade-off (TTO) use direct queries about HRQL. With these techniques, people do all the mental processes of weighing multidimensional issues internally and respond with a summary number or point of indifference between choices. For the remainder of this essay, all direct methods are referred to as *utilities* or *utility*, though some are more accurately called *preferences* or *values* (TTO and VAS). Utilities are technically defined as measuring risk under uncertainty (SG).

MAUT allows the use of SG, TTO, or VAS utilities as a basis for modeling an individual's or a population's overall multi-attribute utility structure. Such models are called *indirect* since the end user may complete a simple survey from which utility is later calculated. A MAUT-based model can be used to calculate all possible health states in a comprehensive health status classification system, as defined below.

Components of MAUT-Based Models

In MAUT-based measurement, the following four steps are followed:

1. Develop a health status classification system, defined as incorporating all relevant *attributes* of health and gradations of function or status within each attribute.
2. Obtain utilities for gradations (levels) of function or status within each attribute.
3. Assign relative weights to the attributes.
4. Aggregate the weights of attributes and single-attribute levels of function to obtain an overall utility measure.

The perspective of the population whose utilities are being assessed for the model should be clear. The two most common perspectives are a representative sample of society and a representative sample of clinical patients experienced with certain health states. Health economists generally prefer the perspective of society.

Developing the Health Status Classification System

An example of a health status classification system (HSCS), the Health Utilities Index Mark 3 (HUI3),

is shown in Figure 1. An HSCS is developed through the work of expert panels and focus groups of patients or members of society, depending on the model. The HSCS provides a basis for judging the model in terms of *face validity* (a simple assessment of whether the model seems sensible and appropriate), *content validity* (the extent to which the model represents the health attributes being measured—evaluated with statistical analysis), and *construct validity* (how the model statistically correlates with other measures of similar concepts or varies between known groups). Most MAUT HRQL models address overall or "generic" health. A minority focus on specific diseases or conditions.

For a MAUT-based model, it is desirable for the attributes to be independent of each other in content. Thus, the single-attribute utilities should be conceptually (preferentially) independent of changes in the other attributes. This aspect is important for successfully assigning weights to the individual attributes and obtaining utilities for each level.

Obtaining Utilities for the Levels of Function

Using any of the direct methods mentioned above, one can obtain utilities for all levels of function or morbidity in each attribute, on a utility scale from 0 to 1.0. Some investigators prefer that attribute levels be elicited with a choice-based method such as SG or TTO. However, a VAS, given its simplicity and low cognitive burden, is often favored for such tasks since in most systems there are a significant number of attribute levels to be valued. A compromise can be made, if needed, by also

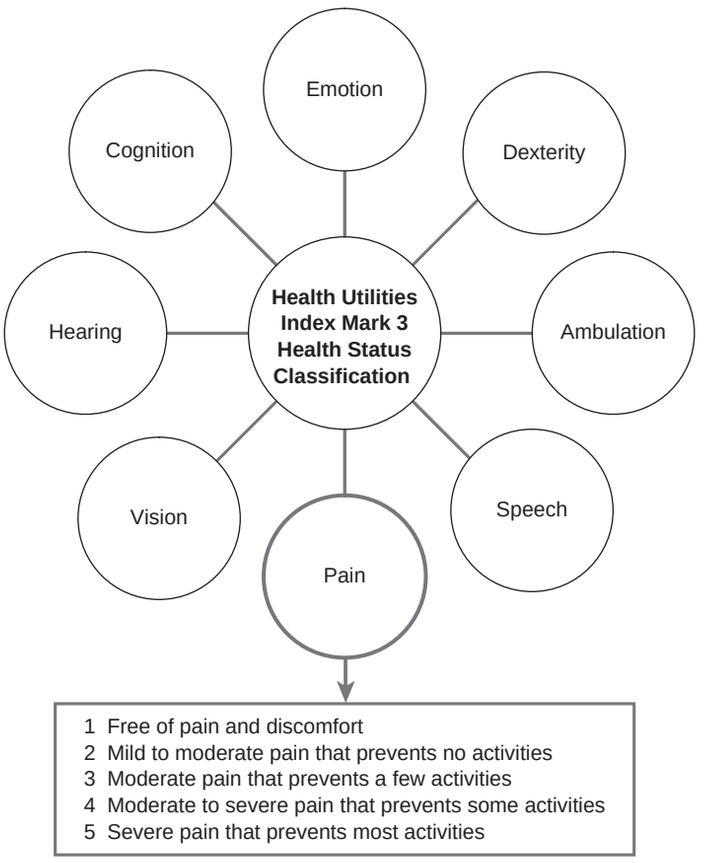


Figure 1 The Health Utilities Index Mark 3 Health Status Classification System

Source: Feeny, D., Furlong, W., Torrance, G. W., Goldsmith, C. H., Zhu, Z., DePauw, S., et al. (2002). Multiattribute and single-attribute utility functions for the Health Utilities Index Mark 3 system. *Medical Care, 40*(2), 113–128.

Note: The details of the 5 levels of the pain attribute in this system are shown. All other attributes have 5 to 6 levels.

selecting a small subset of health states that span the degree of morbidity of the HSCS and obtaining choice-based as well as VAS data. These data can be used for statistical modeling of the relationship of VAS to SG or TTO, followed by transformation of the remaining grouped VAS data to the choice-based method selected. In general, this relationship seems to be nonlinear, with the following forms (Equations 1 and 2) often resulting from experiment:

$$u = v^x \tag{1}$$

or

$$u = 1 - (1 - v)^x, \tag{2}$$

where u is the SG or TTO utility result from experiment and v is the VAS result from experiment.

Assigning Relative Weights to the Attributes

There are four methods of weighting attributes that are most commonly used, which involve rating or ranking. An intuitive approach is to simply have the population rank order the importance of the attributes. Subsequently, 100 importance points are distributed, so that more important attributes have more points. Alternatively, one can use ratios of importance, so that one attribute (often the least important one) is defined as a standard and given a weight (e.g., 10), and all others are given weights (e.g., any multiple of 10) relative to that standard. Swing weighting does not directly assess importance but instead requires the subject to imagine the health state where all attributes are at their worst. The individual is asked which attribute he or she would switch from worst to best before all others, then the next attribute, and so on, until all are ranked. The individual is then asked to assess how much each of those swings was worth, with the most important equal to 100. Weights are normalized by dividing each weight by the sum of all weights. In the fourth method, one can assess utility for the corner state of each attribute (used in the HUI3). Corner states can be defined in two different ways, but the most commonly used approach in health is that state where the attribute of interest is at its worst and all other attributes are at their best, somewhat like swing weighting. Combinations of the above approaches are also used.

Aggregating the Weights of Attributes and Single-Attribute Levels of Function

The assignment of weights to the attributes will determine how or if the attributes of the model interact with one another and, thus, if attributes are to be combined in an additive or multiplicative fashion. The operations to calculate multi-attribute utility models from individual attribute weights and levels can be formalized as the addition or multiplication of simple polynomials. The additive model is the most restrictive and assumes no interactions between attributes and so is based only on the weighted sum of each attribute's contribution to the utility score.

As shown below, when there are no interactions, the interaction term, K (the global scaling constant), is equal to 0 since it is not needed. This also means that that sum of the weights (k_i) for all the attributes is equal to 1. On the other hand, the model is multiplicative if the weights add up to a sum greater than or less than 1. The K term is then needed to adjust the function to a scale between 0 and 1.0. As shown below, the expanded multiplicative version (Equation 4) uses K to express one kind of interaction with all possible combinations of attributes. Note that when $K = 0$, Equation 4 simplifies to the additive model. See Further Readings for a discussion of the interactions allowed in multiplicative models, as well as rarely used multilinear models where more complex interactions are allowed.

Additive Utility Model

The additive utility model is given in the equation

$$u(x) = \sum_{i=1}^{n} k_i u_i(x_i), \tag{3}$$

where

$$\sum_{i=1}^{n} k_i = 1;$$

thus $K = 0$ (K global scaling constant), and an additive model holds. Note $u(x)$ is the utility for health state x, represented by an n-element vector; $\sum$ is summation; k_i is equal to the weight for attribute i;

and $u_i(x_i)$ is the single-attribute utility function for an attribute *i*.

See the SMARTS and SMARTER techniques elsewhere for examples of weighted additive models.

Multiplicative Utility Model

Introducing the interaction term, *K*,

$$u(x) = \sum_{i=1}^{n} k_i u_i(x_i) + \sum_{i<j} K k_i k_j u_i(x_i) u_j(x_j) + \sum_{i<j<m} K^2 k_i k_j k_m u_i(x_i) u_j(x_j) u_m(x_m) + \cdots + K^{n-1} \prod_{i=1}^{n} k_i u_i(x_i). \quad (4)$$

A more compact form of the multiplicative model is shown below, obtained by multiplying each side of the equation by *K*, adding 1, and factoring:

$$(1 + Ku(x)) = \prod_{i=1}^{n} (1 + K k_i u_i(x_i)), \quad (5)$$

where $\prod_{i=1}^{n}$ is the product of all (1 + Kk_i) from k_i to k_n.

Final Scaling Issues

When the model is complete, it should be based on a "Death" to "Perfect Health" utility scale if it is to be used for economic modeling. Some multi-attribute utility models may not be in this form initially, such as those that account for health states worse than death. In that case, the value of death is calculated above some worst possible health state, and the scale is linearly transformed to a Death to Perfect Health scale.

An Alternative: Statistical Modeling

The major alternative to MAUT for measuring multi-attribute utility is statistical modeling. Directly assessed utilities for a subset of levels of function (using only SG or TTO) in a set of health attributes are obtained, and an ad hoc modified linear additive model is generally constructed to predict the remaining levels. This approach has been used for the SF-6D recently and the EURO-QOL and Quality of Well Being Scale (QWB) previously.

Evidence of Reliability, Validity, and Responsiveness

MAUT-based HRQL measures such as the HUI3, as well as non-MAUT measures such as the EURO-QOL, SF-6D, and QWB instruments have been shown to have reasonable test-retest reliability (correlation of scores of the same subjects estimated at two different time points). Because there is no gold standard for HRQL measurement, the degree of construct validity is mainly evaluated. All the existing measures have shown evidence of construct validity in various diseases and at varying levels of responsiveness (the ability to detect meaningful change).

Advantages and Disadvantages of MAUT-Based Measurement

Given the above, the following points summarize advantages of MAUT-based HRQL:

- MAUT is grounded in utility theory.
- The MAUT approach does not make assumptions beforehand of how or if the attributes will interact. The relationships of the weights that are gained by experiment indicate readily whether the model is additive or multiplicative in nature. This aspect could be an advantage relative to non-MAUT statistical methods that assume a certain model structure, usually a linear additive one.
- The MAUT approach informs the utility function like a health profile instrument (SF-36) due to the process that breaks HRQL down into attributes and levels. This is the "decomposed" approach of MAUT, as opposed to direct SG, TTO, or VAS utilities.
- The end user sees a short survey that can be self-administered. Thus, using the final instrument is less labor-intensive than typical utility assessment interviews.
- Differing assessment approaches, interviewers, visual aids, and so on can bias results of SG, TTO, and VAS. MAUT methods avoid these issues since the utility assessments are only done

during the development of the model. Subsequent users have their utilities calculated based on the utilities of those who were interviewed during the model's development, which helps to standardize HRQL utility assessment.

The following are some disadvantages of MAUT-based approaches:

- Development of a MAUT model is resource intensive.
- The valuation process, where subjects are asked to simultaneously assess multiple attributes, can carry a significant cognitive burden.
- Corner state valuations and development of independent attributes may be conceptually difficult in certain situations.
- The number of level assessments required in a MAUT model may be more feasible with the use of a VAS followed by a transformation to SG or TTO, if desired. There is some controversy about the generalizability of such transformations.

Because MAUT is the multi-attribute extension of traditional utility theory, some favor it over statistical modeling for economic analysis of healthcare programs. However, recent work suggests better construct validation of statistical multi-attribute models (by comparing the model with direct SG utilities) as opposed to a similar comparison with a MAUT model. Comparison of these two approaches is an area of ongoing research interest.

J. Shannon Swan

See also Cost-Utility Analysis; Decomposed Measurement; EuroQoL (EQ-5D); Expected Utility Theory; Health Utilities Index Mark 2 and 3 (HUI2, HUI3); SF-6D; SF-36 and SF-12 Health Surveys; SMARTS and SMARTER; Utility Assessment Techniques

Further Readings

Brazier, J., & Roberts, J. (2006). Methods for developing preference-based measures of health. In A. Jones (Ed.), *The Elgar companion to health economics* (pp. 371–381). Northampton, MA: Edward Elgar.

Feeny, D. (2006). The multi-attribute utility approach to assessing health-related quality of life. In A. Jones (Ed.), *The Elgar companion to health economics* (pp. 359–370). Northampton, MA: Edward Elgar.

Feeny, D., Furlong, W., Torrance, G. W., Goldsmith, C. H., Zhu, Z., DePauw, S., et al. (2002). Multiattribute and single-attribute utility functions for the Health Utilities Index Mark 3 system. *Medical Care, 40*(2), 113–128.

von Winterfeldt, D., & Edwards, W. (1986). Multi-attribute utility theory: Examples and techniques. In *Decision analysis and behavioral research* (pp. 259–313). Cambridge, UK: Cambridge University Press.

Multivariate Analysis of Variance (MANOVA)

Multivariate analysis of variance (MANOVA) is a statistical model that generalizes and extends the univariate analysis of variance model. This model is necessary when answers to research questions require the evaluation of multiple outcome measures. While in some cases it may be useful to examine individual outcomes separately, using the univariate model, in many studies the outcomes observed are interrelated. Because of the interrelationship among outcome measures, it is generally more appropriate and meaningful to analyze the outcomes as a composite(s) or weighted combination(s) of the measures using the multivariate model.

While the multivariate model can be applied to a variety of research designs (e.g., between group, repeated measures, mixed model), the focus here is given to posttests-only between-group designs. Furthermore, only a single-factor between-group design is considered. That is, groups are identified based on a single dimension (e.g., drug dosage). Groups may represent existing populations in a nonexperimental study or may be formed through the random assignment of units to the levels of the grouping variable. The same analysis procedures discussed here can be applied to both experimental and nonexperimental studies, with the only difference in application being the inferences that may be drawn from the results. In experimental studies, inferences may be causal, while in nonexperimental studies, only functional relationships may be inferred.

The procedures discussed here can be generalized easily to more complex multifactor designs.

Purpose

When populations are compared, they are generally compared with respect to multiple outcomes or response measures. For example, varying the levels of a vitamin dosage (e.g., 500, 1,000, 1,500, or 2,000 IU) may be the grouping variable under investigation, and the consequences of dosage variation with respect to several outcome measures (e.g., Y_1 = diastolic blood pressure, Y_2 = systolic blood pressure, Y_3 = heart rate, Y_4 = anxiety, Y_5 = mood) may be of interest. Multiple outcomes are often observed because no single outcome measure can adequately capture the intended construct(s) of interest. For example, measures Y_1, Y_2, and Y_3 may be indicators of physical health while Y_4 and Y_5 may be indicators of psychological health. Both physical and psychological health are *latent* constructs that cannot be adequately assessed by any single indicator. However, by combining several indicators, an estimate of a construct can be provided. The purpose of MANOVA is to determine the best combination of indicators to estimate one or more constructs that maximize group differences. Measures are combined by multiplying (i.e., weighting) individual indicators by constants (e.g., $Z = b_1Y_1 + b_2Y_2 + \cdots + b_5Y_5$) to create a composite (i.e., Z) of the outcome measures. The analysis and interpretation of the composites that define the group differences is one of the primary advantages of the multivariate model compared with the univariate model. Other important advantages of the multivariate model include a reduction in the risk of Type I errors and more sensitive (powerful) group comparisons.

Hypothesis Tested

The set of means on the outcome measures within each group is called a *mean centroid*. MANOVA tests the hypothesis that the populations, represented by the groups, have identical centroids: H_0: $\mu_1 = \mu_2 = \ldots = \mu_j (j = 1, 2, \ldots, j)$, where $\mu_j = [\mu_{j1}, \mu_{j2}, \ldots, \mu_{jp}$ $[.]^T$ = transpose of vector, μ_{jm} = mean of population j ($j = 1, \ldots, J$) for outcome measure m ($m = 1, \ldots, p$). Using the vitamin dosage example introduced in the previous section, suppose the mean scores for the five outcome measures observed for the group receiving 500 IU of the vitamin equaled 5, $Y_2 = 75$, $Y_3 = 82$, $Y_4 = 60$, and Y_5 and $Y_5 = 120$; then the sample mean centroid, $\mathbf{Y}_{500}$ = $[115, 75, 82, 60, 120]^T$. To test the hypothesis that population centroids are identical, two matrices are computed, **E** and **H**. The **E** matrix represents a $p \times p$ error matrix of deviations of unit scores around their respective group means on the p outcomes. The **H** matrix represents a $p \times p$ hypothesis matrix of deviations of group means on the p outcomes around the p grand means. The elements on the main diagonal of **E** and **H** are the sum-of-squares within-groups and the sum-of-squares between-groups on the p outcome measures, respectively, used in the univariate model. The off-diagonal elements estimate the interrelationships among the outcome measures. These matrices are used to obtain a very useful statistic called an *eigenvalue*, λ. The number of eigenvalues computed depends on the number of groups and outcome measures studied. The determination of the eigenvalue(s) is a tedious task unless only two outcome measures are examined. These computations are best left to a computer. The General Linear Model (GLM) program in SAS and the MANOVA program is SPSS can provide the necessary calculations. Using the eigenvalues, λ, four different test criteria have been proposed—Wilks, Bartlett-Pillai, Hotelling-Lawley, and Roy—to compute and evaluate a multivariate F statistic. The four criteria provide identical results when only two populations ($J = 2$) are compared. When more than two populations are compared, the four criteria will differ a little but generally lead to the same conclusion regarding the hypothesis tested. All four criteria are reported on SAS and SPSS computer output. The Wilks criterion is the best-known and most frequently cited criterion, but the Bartlett-Pillai criterion is often recommended because of its robustness to assumption violations. The rejection of H_0 means that the population centroids are not identical, and there is some relationship between the grouping variable and the centroids.

Effect Size

Because a trivial association or difference between centroids may be statistically significant, an index

to quantify the degree to which H_0 is false (effect size) is useful. When two population centroids are compared, a popular measure of effect is the distance (difference) between the two centroids, which is provided by the square root of the Mahalanobis D^2 statistic, D. This statistic is a multivariate generalization of Cohen's standardized mean difference, d, which is popular in univariate studies. D is interpreted as the difference between centroids relative to the standard deviation in the total outcome space. What might be considered large or small should be based on the distance between centroids reported in similar previous studies.

When more than two populations are compared, an overall index of the relationship between the grouping variable and the outcome variables is of interest. For the vitamin dosage study, the interest might be in providing an estimate of the overall relationship between vitamin dosage and the outcomes studied. A measure of association has been proposed for each of the four multivariate test criteria. These indices differ as a function of the number of constructs estimated in the MANOVA model. The number of constructs, r, estimated equals the number of eigenvalues computed, which is the lesser of p or $J - 1$ ($r = \min[p, J - 1]$). In the vitamin dosage example, $r = 3$ ($\min[5, 4 - 1]$). Not all estimated constructs may be meaningful, however. In the current example, it is anticipated that two meaningful constructs (e.g., physical health and psychological health) will be identified. In the next section, procedures for determining the number of meaningful constructs represented in the data are discussed. If only one construct is estimated (e.g., $r = 1$) because only two populations are compared, all four criteria provide the same measure of association, called the squared canonical correlation and indicating the proportion of variation in the variable system that is explained by or shared with the grouping variable. The squared canonical correlation is obtained by computing $\lambda/(1 + \lambda)$. If all estimated constructs are meaningful, the measure of association associated with the Bartlett-Pillai criterion provides the average proportion of variation per construct that is explained by the grouping variable.

These multivariate measures of association overestimate the strength of association between the grouping variable and the constructs. The observed measures of effect can be adjusted to reduce the bias. The adjustment suggested by Ronald Serlin is

$$\mathrm{ES}_{\mathrm{adj}} = 1 - \frac{N-1}{N-b-1}(1 - \mathrm{ES}),$$

where

ES = any of the four effect size indices associated with the multivariate test criterion,

N = total sample size, and

$b = \max(p, J - 1)$.

For example, if the measure of association provided by the Bartlett-Pillai criterion for the vitamin dosage study ($p = 5$, $J = 4$, $N = 80$) equaled .136, the adjusted effect size would equal .078

$$\left(= 1 - \frac{80-1}{80-5-1}(1 - .136)\right).$$

That is, on average, vitamin dosage (the grouping variable) explains 7.8% of the variation in each construct estimated. While this correction is not provided in current software packages, it can be computed easily with the available computer output.

Dimensionality

As noted earlier, to test the hypothesis that population mean centroids are identical, at least one eigenvalue is computed. Each eigenvalue is associated with an independent latent construct, but every identified construct may not be meaningful. For the comparison of four vitamin dosages with five outcome variables, three eigenvalues are computed ($r = \min[p, J - 1]$). It is anticipated that two meaningful constructs (e.g., physical health and psychological health) are represented by the five outcome variables. One approach, to determine the number of meaningful constructs present in the variable space defined by the p observed outcome measures, is to order the computed eigenvalues from largest to smallest and then compute the proportion of the total variation in the variable space that is associated with each construct estimated. This is achieved by computing the ratio of each eigenvalue to the sum of the eigenvalues $\left(\mathrm{pct}_v = \lambda_v \Big/ \sum_{v=1}^{r} \lambda_v\right)$. The number of meaningful

nstructs present in the outcome space can then be judged based on the number of eigenvalues needed to "substantially" explain the total variation in the outcome space. This is a subjective judgment that may be based on the researcher's experience, a theoretical model, or previous research findings. Often, the number of constructs assessed can be anticipated accurately by the researcher.

Suppose the three eigenvalues in the vitamin dosage study equaled $\lambda_1 = .261$, $\lambda_2 = .165$, and $\lambda_3 = .063$. Then, the three estimated constructs explain 56.6%, 35.7%, and 7.8% of the total variance in the variable space, respectively. It might then be argued that the first two constructs explain most of the variance in the data set, and so the five outcomes are indicators of two latent variables.

A second strategy for determining the number of meaningful constructs assessed in the variable outcome space is to conduct a series of statistical tests sequentially from the largest to the smallest eigenvalue, using the Wilks test criterion. The first test determines whether at least one construct separates the groups. The second test, based on the second largest eigenvalue, determines whether at least two constructs separate the groups. The testing stops when the *k*th ranked eigenvalue is *not* statistically significant. The number of constructs to be interpreted then equals $k - 1$. These tests can be carried out using either the SAS (e.g., Proc Disc) or SPSS (e.g., MANOVA) software packages. With the vitamin dosage study, it is likely that the first two constructs would be statistically significant, leading to the same conclusion derived in the previous paragraph.

Defining Latent Constructs

Once the number of latent variables has been determined, a final step is needed to complete the analysis of the data set: The constructs need to be defined. Each construct is defined based on the linear composite or weighted sum of the observed outcome measures (i.e., $Z_v = b_{v1}Y_1 + b_{v2}Y_2 + \cdots + b_{vp}Y_p$).

Both SAS and SPSS can provide these weights. The *b*s are referred to as linear discriminant function weights (LDFs), and a unique solution for **b** is not possible. However, the possible solutions are proportional to one another. The weights themselves are not useful for defining the latent constructs because their magnitude is influenced by the scaling of the outcome measures. But these weights may be used in two different ways to provide construct definition.

One way to define the latent constructs is to standardize the discriminant function weights. The relative magnitude of the absolute value of the standardized weights can then be used to judge which outcome measures contribute the greatest to the composite. The variables making the greatest contribution define the construct. Two limitations with this approach are that (1) the collinearity among the outcome measures can reduce the weights for important measures and (2) the addition or reduction of outcome measures can greatly affect the weights.

A second approach for defining a latent construct is to compute the correlation between the composite score, Z_v, with each outcome measure. These correlations are called *structure r*s. Outcome measures with relatively high absolute values for the structure *r*s define the latent construct. A problem with using structure *r*s to define the construct is that the correlations are proportional to the univariate *F* statistics comparing group means on the outcome measures. Consequently, the structure *r*s do not reflect the multivariate criterion. Both standardized discriminant function weights and structure *r* coefficients can be obtained through SAS (e.g., GLM) and SPSS (e.g., MANOVA).

As an example, hypothetical standardized discriminant function weights and structure *r*s are presented in Table 1. Both the standardized weights and structure *r*s indicate that measures Y_4 and Y_5 assess one construct and measures Y_1, Y_2, and Y_3 assess a second construct. LDF3 would not be interpreted based on the previous dimensionality analysis. Based on what the researcher believes that Y_1 though Y_5 measure, the constructs are defined. For example, because Y_4 and Y_5 are psychological measures, the first construct might be labeled psychological health. Measures Y_1, Y_2, and Y_3 are physical measurements, and so the second construct might be labeled physical health. Standardized discriminant function weights and structure *r*s do not always agree, however.

Contrasts

The procedures outlined above have been presented with respect to the simultaneous comparison of

Table 1 Standardized discriminant function weights and structure *r*s for five outcomes comparing four vitamin dosages

Measure	*LDF1*	*LDF2*	*LDF3*
Standardized weights			
Y_1	.023	.362	.057
Y_2	.136	.451	–.102
Y_3	–.201	.582	.206
Y_4	.583	.263	.168
Y_5	–.682	.097	–.321
Structure *r*s			
Y_1	.323	.621	.175
Y_2	.265	.553	.226
Y_3	.320	.498	.126
Y_4	.733	.363	.382
Y_5	–.826	.179	.231

all levels of the grouping variable, or the omnibus hypothesis test. In many contexts, the omnibus hypothesis test does not directly address the researcher's interests. Rather, specific comparisons or contrasts can be specified. For example, a comparison of 500 IU with 1,000 IU of the vitamin might be of interest. Alternatively, the researcher might believe that the vitamin may have diminishing benefits with increasing dosage levels. That is a quadratic relationship between dosage, and mean centroids may be anticipated. To address these interests, the same procedures as those described above may be followed: Test the hypothesis, estimate the effect size, and define the construct. The results, however, are interpreted only in terms of the populations being compared. A test for dimensionality is not needed because all contrasts provide one eigenvalue, $r = 1$, and a single construct is identified for each contrast. It is important to note that the constructs that define the separation among groups in the omnibus test may not be the same constructs that define specific group differences, and different constructs can be obtained for different contrasts. Both GLM in SAS and MANOVA in SPSS allow the specific analysis of group differences through contrast analyses.

Assumptions

To make inferences from the sample data to the populations they are believed to represent, the MANOVA model relies on data assumptions similar to those assumed in the univariate model. That is, the units of analysis are assumed to be independent of each other, the outcome measures have a multivariate normal distribution within each population, and the population covariance matrices, Σ_j, for the J populations are identical, where

$$\Sigma_j = \begin{pmatrix} \sigma_m^2 & \cdots & \sigma_{mp} \\ \vdots & \ddots & \vdots \\ \sigma_{mp} & \cdots & \sigma_p^2 \end{pmatrix},$$

σ_m^2 = variance of measure m ($m = 1, \ldots, p$), and

σ_{mp} = covariance of measures (m, p).

The robustness of the MANOVA model to these assumptions is similar to the univariate model. Independence among units is essential for statistical validity. Violating the multivariate normality assumption generally does not seriously invalidate statistical inference. The effect of violating the homogeneity of covariance matrices does not invalidate the hypothesis test if sample sizes are similar. But if sample sizes are substantially unequal (differ by at least a factor of 2), the Type I error rate may be over- or underestimated, depending on the relationship between the sample sizes and the determinants of the population covariance matrices. A positive relationship can result in a conservative hypothesis test, while a negative relationship can result in a liberal hypothesis test. The Box test reported in both SAS and SPSS is often used to determine whether the assumption is violated, but this procedure is sensitive to multivariate nonnormality and can be extremely sensitive to a minor assumption violation. Alternatively, the log determinants of the group sample covariance and the pooled covariance matrices may be compared visually, and if judged to be similar, robustness may be assumed. If the assumption is judged to be

seriously violated, analysis procedures developed by Ying Yao and Søren Johansen may be useful. When the assumption of equal covariance matrices is violated, methods for defining the latent constructs have not been developed.

Advantages

MANOVA is a very useful and powerful statistical tool for addressing research questions comprehensively and efficiently. Because important constructs can rarely be assessed with a single measure, multiple outcomes must be assessed. In addition, population differences and the consequences of interventions typically have consequences on more than a single construct. MANOVA provides an approach to examining data to determine the number of constructs assessed, test for construct differences between populations, and provide a mechanism for defining the constructs assessed by the multiple indicators. These are all important advantages of the multivariate model compared with a univariate model. Because the multivariate model assesses population differences on constructs, careful selection of the outcome measures to assess is essential for a meaningful and interpretable analysis.

Stephen Olejnik

See also A nalysis of Covariance (ANCOVA)

Further Readings

Box, G. E. P. (1949). A general distribution theory for a class of likelihood criteria. *Biometrika, 36,* 317–346.

Bray, J. H., & Maxwell, S. E. (1985). *Multivariate analysis of variance*. Beverly Hills, CA: Sage.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum.

Huberty, C. J., & Olejnik, S. (2006). *Applied MANOVA and discriminant analysis* (2nd ed.). New York: Wiley.

Johansen, S. (1980). The Welch-James approximation to the distribution of the residual sum of squares in a weighted linear regression. *Biometrika, 67,* 85–92.

Kim, S., & Olejnik, S. (2005). Bias and precision of measures of association for a fixed-effect multivariate analysis of variance model. *Multivariate Behavioral Research, 40,* 401–421.

Olson, C. L. (1974). Comparative robustness of six tests in multivariate analysis of variance. *Journal of the American Statistical Association, 69,* 894–908.

Serlin, R. C. (1982). A multivariate measure of association based on the Billai-Bartlett procedure. *Psychological Bulletin, 91,* 413–417.

Yao, Y. (1965). An approximate degrees of freedom solution to the multivariate Behrens-Fisher problem. *Biometrika, 52,* 139–147.

N

Net Benefit Regression

Net benefit regression describes the activity of doing regression analysis on net benefit data. This entry first describes what net benefit data are and then explains how these data can be analyzed using regression. Finally, this entry concludes with an example illustrating some of the many reasons to use regression analysis to analyze net benefit data.

Overview

Scientific journals and popular media teem with the results of clinical trials proclaiming evidence of more effective new treatments or interventions. Without the resources to be able to provide all new treatments that are more effective, how should decision makers choose? Often, clinical enthusiasm is tempered with economic discipline, in the form of an economic evaluation. In many decision-making contexts, this involves estimating the extra cost of an extra unit of a health outcome. For example, Coyle and colleagues derived the efficacy of erythropoietin (EPO) by conducting a meta-analysis of published randomized trials. In their study, EPO alone led to modest benefits compared with no intervention for orthopedic surgery (.000024 life-year gained per patient) or as an augmentation to preoperative autologous donation (.000006 life-year gained per patient). Based solely on the effectiveness data, decision makers may be ambivalent about covering a drug that leads to modest gains. However, results from the cost-effectiveness analysis (CEA) showed that the extra cost of one more year of life was around $66 million (Canadian) for EPO compared with no intervention and $329 million (Canadian) for EPO to augment preoperative autologous donation. There may be other valid reasons to use EPO in this situation; however, economic efficiency does not appear to be one of them.

Theory

In a clinical trial, CEA estimates the economic efficiency (or the extra cost of an extra unit of patient outcome) of a new treatment by computing the sample means of the cost and effectiveness data for both the new treatment and the usual care groups. The difference in the average costs is called the *incremental cost* (ΔC). The difference in the average effects is called the *incremental effect* (ΔE). The ratio of the incremental cost to the incremental effect is called the *incremental cost-effectiveness ratio* (ICER = $\Delta C/\Delta E$). In the example above, the incremental cost of using EPO compared with no intervention was $1,588, and the incremental cost for EPO to augment preoperative autologous donation was $1,936. Therefore, the ICERs were $66 million per additional year of life ($\Delta C/\Delta E$ = $1,588/.000024 life-year gained) and $329 million per additional year of life ($\Delta C/\Delta E$ = $1,936/.000006 life-year gained), respectively.

In theory, the ICER is a useful statistic because it estimates the rate at which a unit of patient outcome can be purchased with a new treatment

(e.g., $329 million per additional year of life). The ICER can be compared with a decision maker's willingness to pay (WTP). For example, if a new treatment produces better outcomes at a rate of $50,000 per extra year of life, and if the decision maker's WTP is $20,000 per extra year of life, then the new treatment is not a good value. Why would a decision maker spend at a rate of $50,000 per extra year of life (the ICER), when a rate of $20,000 per extra year of life (the WTP) is deemed appropriate? In contrast, if the WTP were $80,000, then the new treatment could be considered cost-effective because the decision maker values what she or he is getting more than what she or he must give up.

In the example presented later in this entry, ΔC = $5,000 and ΔE = .10 life-year. The ICER is $5,000/.10 = $50,000. If WTP = $80,000, then the value of what is gained is .10 life-year × $80,000 or $8,000 (note the assumption that if 1 year of life is worth $80,000, then one tenth of a year of life is worth one tenth of $80,000). It costs an extra $5,000 to realize this gain. The value of the extra benefit outweighs the value of the extra cost by $3,000.

Net benefit (NB) calculation is formulated in the cost-effectiveness literature in two separate ways. Adhering to the method illustrated in the previous paragraph, the incremental net benefit (INB) can be represented as

$$\text{INB} = \text{Extra effect} \times \text{WTP} - \text{Extra cost}.$$

When the assumption is made that WTP = $0, then INB = –Extra cost. When the assumption is made that WTP = ICER, then INB = 0. The value for WTP is a major assumption when using the INB since often only the decision maker knows the decision maker's WTP. As with other values that are unknown to the analyst, the lack of knowledge about WTP need not hamper the analysis. Sensitivity analysis can be used to explore how sensitive findings are to assumptions about WTP. Net benefits can be calculated for both treatment options (i.e., new treatment and usual care); their difference represents the INB (the example that follows illustrates this).

Regression

In 2002, Hoch and colleagues introduced net benefit regression as a way to estimate INB using regression methods. The main idea is that when running a regression such as

$$Y = \beta_0 + \beta_1 tx,$$

where tx = 0 if the patient received usual care and 1 if the patient received new treatment, the estimate of β_1 = Average $Y_{\text{new treatment}}$ – Average $Y_{\text{usual treatment}}$. Therefore, if the dependent variable (Y) is cost, then the estimate of β_1 = Average $\text{cost}_{\text{new treatment}}$ – Average $\text{cost}_{\text{usual treatment}}$ = Extra cost, and if the dependent variable (Y) is effect, then the estimate of β_1 = Average $\text{effect}_{\text{new treatment}}$ – Average $\text{effect}_{\text{usual treatment}}$ = Extra effect.

The authors show that if the dependent variable is Effect × WTP – Cost, then the estimate of β_1 = Average $\text{NB}_{\text{new treatment}}$ – Average $\text{NB}_{\text{usual treatment}}$ = INB.

The reason this is important is that if one creates a net benefit variable for each patient (i.e., computes Effect × WTP – Cost for each person) and then runs the regression

$$\text{NB} = \beta_0 + \beta_1 tx,$$

where tx is defined as before, the coefficient estimate for β_1 equals INB (the INB of the new treatment). If the estimate of β_1 is greater than 0, the new treatment is cost-effective.

The estimate of β_1 and its 95% confidence interval changes as one's choice of WTP changes. This is because both the estimate β_1 and its 95% confidence interval are functions of WTP. Therefore, it is a good idea to run a few net benefit regressions with different WTP values to see how sensitive the results are to one's choices of WTP. The next section illustrates how the process works and how to interpret the results.

Example

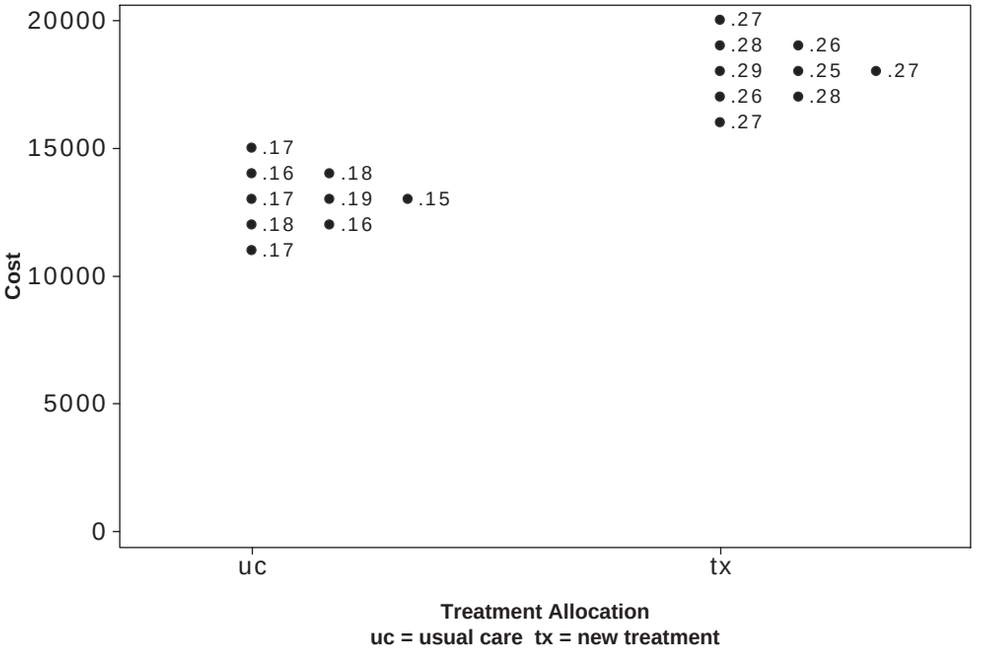
Data

To illustrate net benefit regression, hypothetical data are plotted in Figure 1 and summarized in Table 1 comparing a new treatment (tx = 1) with usual care (tx = 0) for patients with a deadly disease. To facilitate the duplication of methods and reasoning, the hypothetical data have nine people in each treatment group. The patient outcome chosen to represent success was survival (measured in years). In this example, since no one lives

longer than a year, patient outcome is reported as a decimal.

Results

Based on the hypothetical data plotted in Figure 1 and summarized in Table 1, the extra cost is $5,000 and the extra effect is .10 life-year (or about 1.2 months). The ICER estimate is $50,000 per extra life-year (i.e., $\Delta C/\Delta E$ = $5,000/.10 life-year). The lower part of Table 1 provides an example of a sensitivity analysis varying WTP from $0 to $100,000. When the WTP is less than $50,000, the INB is less than 0 (i.e., the net benefits from new treatment are less than those from usual care). Alternatively, when the WTP is greater than $50,000, new treatment is cost-effective since the value of the extra benefits outweighs the extra costs (i.e., INB > 0). The upper graph in Figure 2 illustrates the net benefits for both the usual care



Data stratified by treatment allocation status (tx = 0 is usual care; tx = 1 is new treatment)

obs	effect	cost	tx
1	.15	13000	0
2	.16	12000	0
3	.16	14000	0
4	.17	11000	0
5	.17	13000	0
6	.17	15000	0
7	.18	12000	0
8	.18	14000	0
9	.19	13000	0

obs	effect	cost	tx
10	.25	18000	1
11	.26	17000	1
12	.26	19000	1
13	.27	16000	1
14	.27	18000	1
15	.27	20000	1
16	.28	17000	1
17	.28	19000	1
18	.29	18000	1

Figure 1 Cost data (plotted on the vertical axis) and effect data (plotted near the symbol marker) by treatment status allocation

Table I Sample statistics from the hypothetical economic evaluation data

Variable	*Mean*
Usual care (n = 9)	
Cost	US$13,000
Effect	.17 life-year
New treatment (n = 9)	
Cost	US$18,000
Effect	.27 life-year
Increments	
Cost difference	US$5,000
Effect difference	.10 life-year
Incremental cost-effectiveness ratio	
$5,000/.10 = $50,000 per life-year	

Incremental Net Benefits	*Net Benefits (NB)*		*Incremental*
	New Treatment	*Usual Care*	
WTP = $0	–$18,000	–$13,000	–$5,000
WTP = $20,000	–$12,600	–$9,600	–$3,000
WTP = $40,000	–$7,200	–$6,200	–$1,000
WTP = $60,000	–$1,800	–$2,800	$1,000
WTP = $80,000	$3,600	$600	$3,000
WTP = $100,000	$9,000	$4,000	$5,000

Notes: Net benefits (NB) are calculated as WTP × Average Effect – Average Cost. Incremental Net Benefits (INB) are calculated as $NB_{\text{new treatment}} - NB_{\text{usual care}}$, so INB = WTP × ΔE – ΔC. For example, if WTP = $100,000, $NB_{\text{new treatment}}$ = $100,000 × .27 – $18,000 = $9,000, $NB_{\text{usual care}}$ = $100,000 × .17 – $13,000 = $4,000, and INB = $100,000 × .10 – $5,000 = $5,000.

group (denoted with "◊"s) and the new treatment group (denoted with "+"s). Lines are drawn to connect the *average* net benefit value for each WTP value. The difference between the two average lines is the INB. When WTP = $50,000, the lines intersect, and the net benefit for usual care equals the net benefit for new treatment. The lower graph in Figure 2 also illustrates the average net benefits for both treatment groups. It is clear that when WTP = $60,000, the average net benefits for usual care and new treatment are both negative. However, the average net benefit for new treatment is less negative. Consequently, the INB is greater than 0. The upper graph in Figure 2 confirms this as the line for the "+"s (new treatment's net benefit line) is above the line for the "◊"s (usual care's net benefit line).

Net benefit regression provides an exact estimate of the difference between the two net benefit lines (i.e., the INB). The results of the net benefit regressions, estimating INB, are reported in Table 2. Simple linear regressions of the form

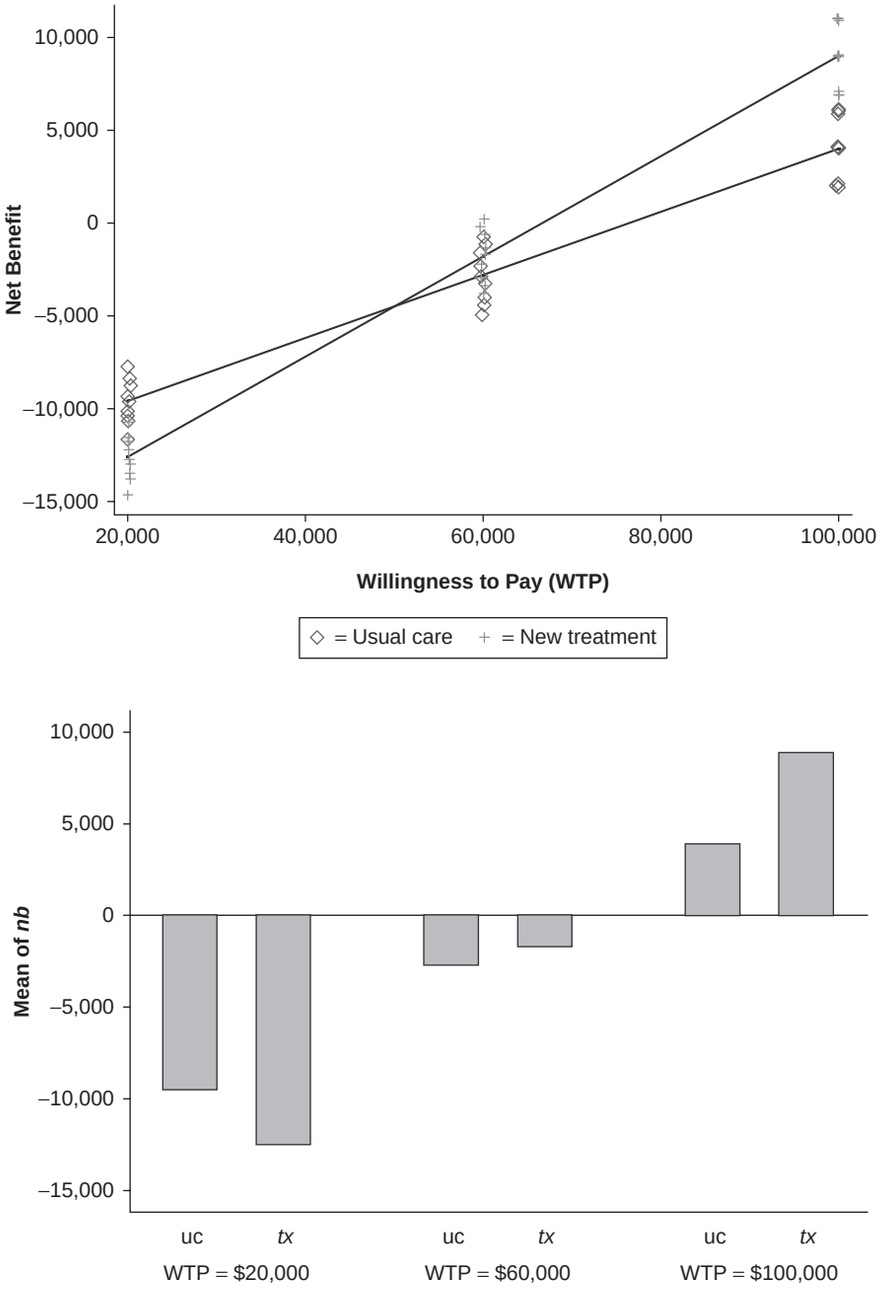


Figure 2 Illustration of net benefits and average net benefits for the hypothetical data

$$nb = \beta_0 + \beta_1 tx$$

were estimated using ordinary least squares (OLS). The dependent variable *nb* was calculated as Effect × WTP – Cost. Net benefit regressions were run for small (WTP = $0 and $20,000), medium (WTP = $40,000 and $60,000), and large (WTP = $80,000 and $100,000) WTP values. The results are discussed next.

Discussion

The results suggest that for small WTP values (WTP = $0 and $20,000), the new treatment is not cost-effective. The INB estimate (the coefficient estimate for the new treatment indicator) is negative, and its 95% confidence interval includes negative values only. For moderate WTP values (WTP = $40,000 and $60,000),

Table 2 Simple linear regression results with hypothetical data (N = 18) with figures rounded to the nearest whole number

	Net Benefit Regression Results by Willingness to Pay (WTP)					
Variables	*Small WTP INB Estimate (95% CI)*		*Medium WTP INB Estimate (95% CI)*		*Large WTP INB Estimate (95% CI)*	
	WTP = $0	*WTP = $20,000*	*WTP = $40,000*	*WTP = $60,000*	*WTP = $80,000*	*WTP = $100,000*
Constant Term	–13,000* (–13,865, –12,135)	–9,600* (–10,483, –8,717)	–6,200* (–7,132, –5,268)	–2,800* (–3,809, –1,791)	600 (–508, 1,708)	4,000* (2,776, 5,224)
New Treatment Indicator (tx = 1 ⇒ Yes, tx = 0 ⇒ No)	–5,000* (–6,224, –3,776)	–3,000* (–4,248, –1,752)	–1,000 (–2,318, 318)	1,000 (–427, 2,427)	3,000* (1,433, 4,567)	5,000* (3,269, 6,731)
R^2 (adjusted)	.8132	.5949	.0853	.0662	.4763	.6822
F Statistic, *F*(1, 16)	75.00	25.96	2.59	2.21	16.46	37.50
Prob > *F*	<.0001	.0001	.1273	.1569	.0009	<.0001

*Statistically significant at the 5% level.

there is high uncertainty about the new treatment's cost-effectiveness; the INB estimate switches signs (it is < 0 at $40,000 and > 0 at $60,000), and the 95% confidence intervals include 0. With high WTP values (WTP = $80,000 and $100,000), the new treatment appears cost-effective. The INB estimate is statistically significantly greater than 0 (the INB estimate is positive, and the 95% confidence interval includes only positive numbers). While a decision maker's WTP may be unknown, the implications are clear for small and large WTP values. For moderate WTP values, the results are sensitive to the assumed WTP value.

In this example, a simple linear regression was run, estimating coefficients using OLS methods. More complex analytical strategies are available for the analysis of cost-effectiveness data, but a recent review of economic evaluations found that many analyses are struggling to present simple, correct conclusions. After the first principles have been mastered, analysts can begin to make use of the substantial methodological work aimed at improving the quality of CEAs done with person-level data. Researchers have studied how best to handle missing data, skewed cost data, between-center differences, multilevel models, and seemingly unrelated regression models. Yet there is still more work to be done. For example, propensity scores have been proposed for CEAs using observational data; however, it is possible that other methods (e.g., instrumental variables) might better ameliorate selection bias. An important advantage of net benefit regression is that it allows all the methods that have been developed for regression analysis (e.g., model-fit diagnostics, advanced estimation and inference techniques) to be directly applied to economic evaluation. Another advantage of net benefit regression in its simplest form is that most researchers are familiar with OLS regression. If key methods can be explained in a familiar regression framework, it seems likely that their widespread adoption might be facilitated. Efforts to explain how to do cost-effectiveness analysis with simple linear regression using OLS represent an attempt at more universal knowledge transfer.

Implications

Economic "duality theorems" show how cost containment can be reframed as efficiency. Some of the earliest mathematical representations of CEA portray healthcare decision makers facing a constrained optimization problem: How does one maximize health given a fixed budget? At least in theory, cost-effectiveness analysis has the potential to improve efficiency and value in healthcare by estimating the extra cost to get one more unit of effect; according to proponents, this is a good way to set priorities. One of the most practical advantages of the net benefit regression approach is being able to use established statistical techniques to analyze cost-effectiveness data (e.g., to adjust for imperfect randomization or to identify important patient subgroups).

Jeffrey S. Hoch

See also Acceptability Curves and Confidence Ellipses; Confidence Intervals; Cost-Effectiveness Analysis; Cost-Utility Analysis; Net Monetary Benefit; Ordinary Least Squares Regression; Pharmacoeconomics; Propensity Scores; Quality-Adjusted Life Years (QALYs); Randomized Clinical Trials; Willingness to Pay

Further Readings

Briggs, A., Nixon, R., Dixon, S., & Thompson, S. (2005). Parametric modelling of cost data: Some simulation evidence. *Health Economics, 14*(4), 421–428.

Coyle, D., Lee, K. M., Fergusson, D. A., & Laupacis, A. (1999). Economic analysis of erythropoietin use in orthopaedic surgery. *Transfusion Medicine, 9*(1), 21–30.

Hoch, J., & Blume, J. (2008). Measuring and illustrating statistical evidence in an economic evaluation: Applying the likelihood paradigm to cost effectiveness analysis. *Journal of Health Economics, 27*(2), 476–495.

Hoch, J. S., Briggs, A. H., & Willan, A. R. (2002). Something old, something new, something borrowed, something blue: A framework for the marriage of health econometrics and cost-effectiveness analysis. *Health Economics, 11,* 415–430.

Hoch, J. S., & Dewa, C. A. (2008). Clinician's guide to correct cost-effectiveness analysis: Think incremental not average. *Canadian Journal of Psychiatry, 53*(4), 267–274.

Mahoney, E., Mehta, S., Yuan, Y., Jackson, J., Chen, R., Gabriel, S., et al. (2006). Long-term cost-effectiveness of early and sustained clopidogrel therapy for up to 1 year in patients undergoing percutaneous coronary intervention after presenting with acute coronary syndromes without ST-segment elevation. *American Heart Hospital Journal, 151*(1), 219–227.

Manca, A., Rice, N., Sculpher, M., & Briggs, A. (2005). Assessing generalisability by location in trial-based cost-effectiveness analysis: The use of multilevel models. *Health Economics, 14*(5), 471–485.

Nixon, R., & Thompson, S. (2005). Methods for incorporating covariate adjustment, subgroup analysis and between-centre differences into cost-effectiveness evaluations. *Health Economics, 14*(12), 1217–1229.

Stinnett, A. A., & Mullahy, J. (1998). Net health benefits: A new framework for the analysis of uncertainty in cost-effectiveness analysis. *Medical Decision Making, 18*(Suppl. 2), S68–S80.

Tambour, M., Zethraeus, N., & Johannesson, M. (1998). A note on confidence intervals in cost-effectiveness analysis. *International Journal of Technology Assessment in Health Care, 14*(3), 467–471.

Torrance, G. W., Thomas, W. H., & Sackett, D. L. (1972). A utility maximization model for evaluation of healthcare programs. *Health Services Research, 7*(2), 118–133.

Weinstein, M., & Zeckhauser, R. (1973). Critical ratios and efficient allocation. *Journal of Public Economics, 2*(2), 147–157.

Net Monetary Benefit

Generally, cost-effectiveness analysis expresses the outcome in the cost-effectiveness ratio. This ratio relates the difference in costs of two alternative healthcare interventions to their difference in health effects. Two alternatives may, for example, be two types of pharmacotherapies—that is, a new drug being compared with the old standard treatment. The difference in health effects may be expressed in life years gained, quality-adjusted life years (QALYs) gained, disability-adjusted life years (DALYs) averted, and

so on. In a formula, the cost-effectiveness ratio (R) may be written down as

$$R = \Delta C/\Delta E,$$

with ΔC the difference in costs and ΔE the difference in health effects.

Obviously, R represents a ratio, limiting its usefulness for understanding the relative sizes of the differences in costs and effects. For example, the ratio does not provide any information on the budget impact. Furthermore, the ratio does not differentiate between the SE and NW quadrants of the cost-effectiveness plane. For example, both the combinations of ($\Delta C = -100$; $\Delta E = 10$) and ($\Delta C = 100$; $\Delta E = -10$) result in the same cost-effectiveness ratio of −10, with the former combination being very acceptable and the latter very unpleasant. Finally, it is well established that statistical analysis on ratios involves some specific problems. Ergo, some drawbacks exist while working with cost-effectiveness ratios.

As one of several options to overcome these drawbacks, the concept of net monetary benefit (NMB) has been developed. To arrive at the NMB, the above equation on the cost-effectiveness ratio is simply rewritten, after having inserted an explicit threshold for cost-effectiveness (often denoted as λ). This threshold explicitly gives the maximum amount of money that society/decision makers want to pay for gaining one unit of health effect, for example, a QALY. So we are interested in the situation that

$$R = \Delta C/\Delta E < \lambda.$$

If we rewrite this as

$$\Delta C - \lambda\Delta E < 0,$$

or

$$\lambda\Delta E - \Delta C > 0,$$

we have formally derived the requirement that the NMB ($= \lambda\Delta E - \Delta C$) should be positive. So, for example, for a new drug, we would require that the monetarized difference in health effects ($\lambda\Delta E$) exceeds the difference in costs (ΔC). It is immediately clear from inserting just the simple example of ($\Delta C = -100$; $\Delta E = 10$) and ($\Delta C = 100$; $\Delta E = -10$) that the NMB does differentiate between SE and NW quadrants.

Furthermore, if calculated on the exact patient populations within specific countries, the NMB provides exact information on the costs (savings) to be net paid (achieved). As such, it gives policy makers information on the socioeconomic impact at the national macrolevel. Finally, calculus of the NMB enhances possibilities for formal statistical tests and analysis of uncertainty around point estimates of cost-effectiveness.

NMB is a versatile tool to evaluate uncertainty in health-economic analyses. For example, derivation of uncertainty intervals around cost-effectiveness ratios is often based on analysis of NMB. Also, NMB can be used for sample size estimation for cost-effectiveness in clinical trials. Furthermore, subgroup analysis using regression techniques on NMB is very straightforward.

NMB allows also for uncertainty evaluation of cost-effectiveness point estimates from health-economic models using cost-effectiveness acceptability curves. In this case, the probabilistic sensitivity analysis is used to generate cost and effect pairs (e.g., 10,000 simulations) for all alternatives included in the model (see Table 1 for a numerical example). In the next step, the threshold for cost-effectiveness (λ) is varied, and the NMB for each cost and effect pair is calculated (Table 2). The number of simulations with the highest NMB represents the probability of cost-effectiveness for each alternative at the specific threshold (bold in Table 2). Finally, the probability is plotted versus the threshold (λ) and the cost-effectiveness acceptability curve (CEAC) is created (Figure 1). Also, value of the information theory is building on the NMB concept.

Maarten J. Postma and René (M) van Hulst

See also Cost-Effectiveness Analysis; Uncertainty in Medical Decisions

Table 1 Cost and effect pairs per 5,000 patients

Simulation	*Alternative A*		*Alternative B*		*Alternative C*		*Alternative D*	
	E (QALY)	*Cost ($)*	*E (QALY)*	*Cost ($)*	*E (QALY)*	*Cost ($)*	***E (QALY)***	***Cost ($)***
1	7,125	–253,733	7,179	–210,720	7,196	–196,688	7,220	–233,653
2	7,676	–264,780	7,781	–229,377	7,748	–203,012	7,739	–240,245
3	7,428	–255,535	7,478	–214,558	7,499	–196,075	7,490	–233,646
						. . .	. . .	. . .
10,000	7,866	–265,846	7,928	–219,885	7,922	–203,103	7,921.52	–242,473

Table 2 Net monetary benefit ($) at cost-effectiveness threshold (λ) of $1,000

Simulation	*Alternative A (in Dollars)*	*Alternative B (in Dollars)*	*Alternative C (in Dollars)*	*Alternative D (in Dollars)*
1	7,378,465	7,389,293	7,392,873	**7,453,521**
2	7,940,399	**8,010,714**	7,950,835	7,979,670
3	7,683,631	7,692,794	7,694,736	**7,723,881**
10,000	8,131,486	8,148,165	8,125,535	**8,163,995**
Highest % NMB	0	6.6	5.6	87.8

Note: The alternative with the highest NMB is in bold.

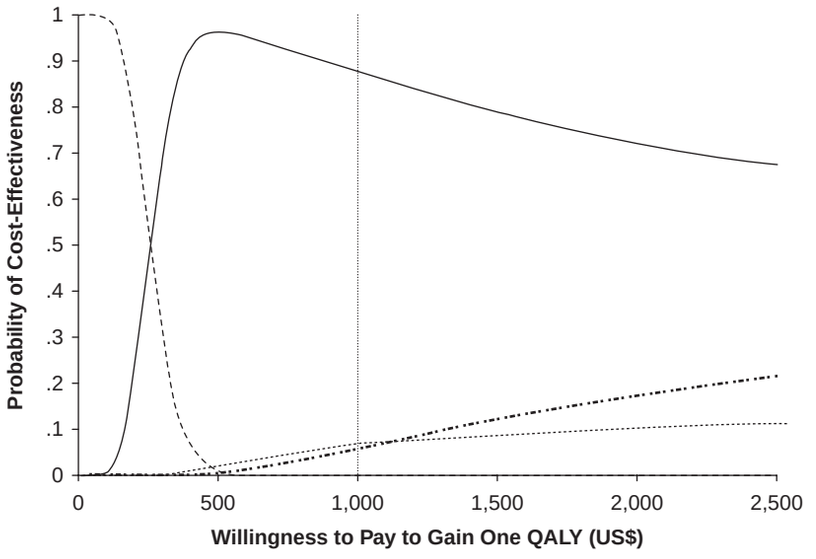


Figure 1 Cost-effectiveness acceptability curve

Note: Probabilities estimated in Table 2 are on the dotted line crossing $1,000. Alternatives A (-----), C (-·-·-·), B (·····), D (——).

Further Readings

Briggs, A. H., O'Brien, B. J., & Blackhouse, G. (2002). Thinking outside the box: Recent advances in the analysis and presentation of uncertainty in cost-effectiveness studies. *Annual Review of Public Health, 23,* 377–401.

Hoch, J. S., Briggs, A. H., & Willan, A. R. (2002). Something old, something new, something borrowed, something blue: A framework for the marriage of health econometrics and cost-effectiveness analysis. *Health Economics, 11*(5), 415–430.

Nomograms

Nomograms are graphical representations of equations that predict medical outcomes. Nomograms use a points-based system, whereby a patient accumulates points based on the levels of his or her risk factors. The cumulative points total is associated with a prediction, such as the predicted probability of treatment failure in the future. Nomograms are attractive as medical prediction tools, because they can consider multiple variables simultaneously to find the best prediction for an individual patient. Nomograms have demonstrated better accuracy than both risk-grouping systems and physician judgment. This improved accuracy should translate into more appropriate patient counseling and medical decision making.

Overview

Making informed medical decisions relies on accurate predictions of the possible outcomes. Paper-based nomograms provide an excellent medium for easily displaying risk probabilities and do not require a computer or calculator. The coefficients used to create the nomogram can be used to create a computer-based prediction tool. The use of nomograms should theoretically help physicians and patients make better treatment decisions. Providing predicted probabilities to patients should reduce the likelihood of regret of treatment choice, particularly when complications arise. However, nomograms are only as good as the data that were used in their creation, and no nomogram can provide a perfect prediction. Ultimately, the best evaluation of a nomogram is made by validating the prediction accuracy of a nomogram on an external data set and comparing the concordance index with another prediction method that was validated using the same data.

Deriving Outcome Probabilities

All medical decisions are based on the predicted probability of different outcomes. Imagine a 35-year-old patient, who presents to a physician with a 6-month history of cough. A doctor in Chicago may recommend a test for asthma, which is a common cause of chronic cough. If the same patient presented to a clinic in rural Africa, the physician may likely test for tuberculosis. Both physicians may be making sound recommendations based on the predicted probability of disease in their locale. These physicians are making clinical decisions based on the overall probability of disease in the population. These types of decisions are better than arbitrary treatment, but they treat all patients the same.

A more sophisticated method for medical decision making is *risk stratification*. Physicians will frequently assign patients to different risk groups when making treatment decisions. Risk group assignment will generally provide better predicted probabilities than estimating risk according to the overall population. In the previous cough example, there are a variety of other factors that may affect the predicted risk of tuberculosis (e.g., fever, exposure to tuberculosis, history of tuberculosis vaccine) that physicians are trained to explore. Most of the risk stratification performed in clinical practice is based on rough estimates that simply order patients into different levels of risk, such as "high risk," "medium risk," or "low risk." Nomograms provide precise probability estimates that generally make more accurate assessments of risk.

Another problem with risk stratification arises when continuous variables are turned into categorical variables. Physicians frequently commit dichotomized cutoffs of continuous laboratory values to memory to guide clinical decision making. Imagine a new blood test for tuberculosis called "serum

marker A." Research shows that patients with serum marker A levels greater than 50 are at an increased risk for tuberculosis. In reality, patients with a value of 51 might have very similar risks compared with patients with a value of 49. In contrast, a patient with a value of 49 would be considered to have the same low risk as a patient whose serum level of marker A is 1. Nomograms allow for predictor variables to be maintained as continuous values while allowing numerous risk factors to be considered simultaneously. In addition, more complex models can be constructed that account for interactions.

Figure 1 illustrates a hypothetical nomogram designed to predict the probability that a patient does not have tuberculosis. Directions for using the nomogram are contained in the figure. One glance at the nomogram allows the user to quickly determine which predictors have the greatest potential impact on the probability of tuberculosis. Fever has a relatively short axis and can contribute less than 25 possible points. In contrast, the exposure to tuberculosis (Tb) variable has a much greater possible impact on the predicted probability.

Nomograms such as the one pictured in Figure 1 are created from the coefficients obtained by the statistical model (e.g., logistic regression or Cox proportional hazards regression) and are only as precise as the paper graphics. However, the coefficients used to create the paper-based nomogram can be used to calculate the exact probability. Similarly, the coefficients can be plugged into a Microsoft Excel spreadsheet or other computer interface that will automatically calculate the probability based on the user inputs.

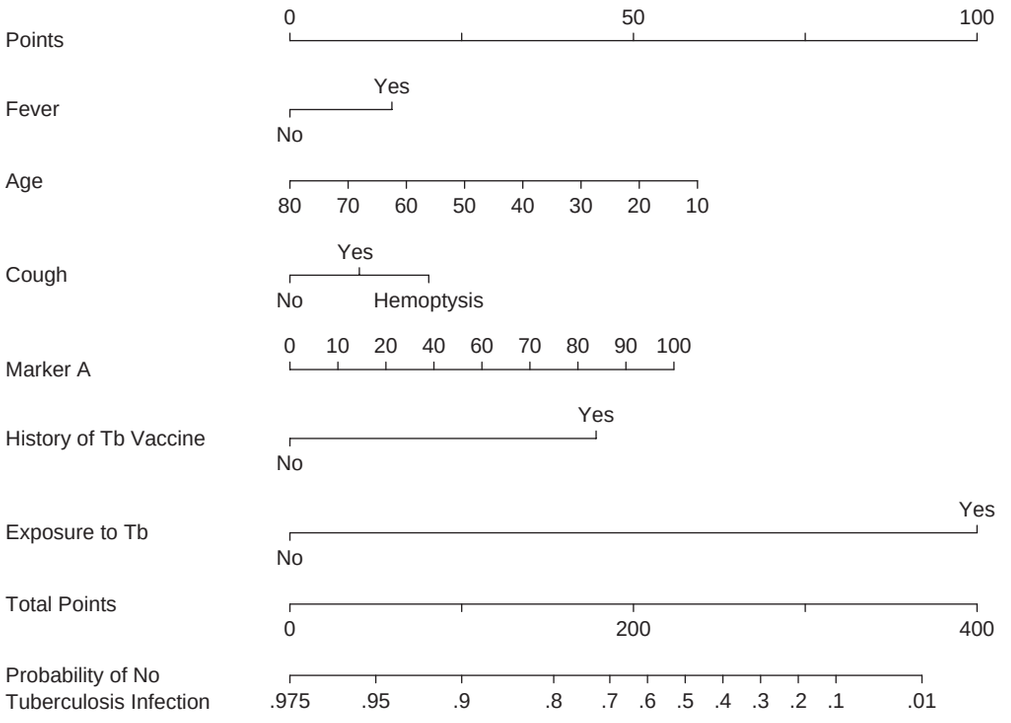
Validation

The estimated probability obtained from nomograms, such as the one in Figure 1, are generally much more accurate than rough probabilities obtained by risk stratification and should help both patients and physicians make better treatment decisions. Predicted probabilities can be graded (validated) on their ability to discriminate between pairs of patients who have different outcomes (discordant pairs). The grading can be performed using either a validation data set that was created with the same database used to create the prediction model (internal validation) or with external data (external validation). Ideally, a nomogram should be validated in an external database before it is widely used in heterogeneous patient populations.

A validation data set using the original data can be created either with the use of bootstrapping or by dividing the data set into random partitions. In the bootstrap method, a random patient is selected, and a copy of the patient's data is added to the validation data set. The patient's record is maintained in the original data set and is available for subsequent random selection. The random selection of patients is continued until a data set that is the same size as the original data set has been formed. The model is applied (i.e., fit) to the bootstrap data, and the model is graded on its ability to accurately predict the outcome of patients in either the original data (apparent accuracy) or the bootstrap sample (unbiased accuracy). Alternatively, the original data can be randomly partitioned. The model is fit to only a portion of the original data, and the outcome is predicted in the remaining subset. The bootstrap method has the added benefit that the sample size used for the model fitting is not reduced.

Evaluating Model Accuracy

As previously mentioned, the models' predictions are evaluated on their ability to discriminate between pairs of discordant patients (patients who had different outcomes). The resultant evaluation is called a concordance index, or *c* statistic. The concordance index is simply the proportion of the time that the model accurately assigns a higher risk to the patient with the outcome. The *c* statistic can vary from .50 (equivalent to the flip of a coin) to 1.0 (perfect discrimination). The *c* statistic provides an objective method for evaluating model accuracy, but the minimum *c* statistic needed to claim that a model has good accuracy depends on the specific condition and is somewhat subjective. However, models are generally not evaluated in isolation. Models can be compared head-to-head either with one another or with physician judgment. In this case, the most



Instructions: Locate the tic mark associated with the value of each predictor variable. Use a straight edge to find the corresponding points on the top axis for each variable. Calculate the total points by summing the individual points for all of the variables. Draw a vertical line from the value on the total points axis to the bottom axis in order to determine the probability that the patient does not have a tuberculosis infection.

Figure 1 Hypothetical nomogram for predicting risk of tuberculosis (Tb) (not for clinical use)

accurate model can generally be identified as the one with the highest concordance index.

However, to fully grade a model, it is also necessary to determine a model's calibration. Calibration is a measure of how close a model's prediction compares with the actual outcome and is frequently displayed by plotting the predicted probability (or value) versus the actual proportion with the outcome (or actual value). The concordance index is simply a "rank" test that orders patients according to risk. A model can theoretically have a great concordance index but poor calibration. For instance, a model may rank patients appropriate while significantly overestimating or underestimating the probability (or value) in all the patients.

Brian J. Wells

See also Artificial Neural Networks; Calibration; Computer-Assisted Decision Making; Decision Rules; Discrimination; Patient Decision Aids; Prediction Rules and Modeling; Risk-Benefit Trade-Off; Shared Decision Making

Further Readings

Harrell, F. E., Jr. (1996). Multivariate prognostic models: Issues in developing models, evaluating assumptions and accuracy, and measuring and predicting errors. *Statistics in Medicine, 15,* 361.

Harrell, F. E., Jr., Califf, R. M., Pryor, D. B., Lee, K. L., & Rosati, R. A. (1982). Evaluating the yield of medical tests. *Journal of the American Medical Association, 247,* 2543–2546.

Kattan, M. W. (2003). Nomograms are superior to staging and risk grouping systems for identifying high-risk patients: Preoperative application in prostate cancer. *Current Opinion in Urology, 13,* 111–116.

Nonexpected Utility Theories

Medical treatments offer health prospects. A health prospect represents a course of action with respect to one's health for which the outcome is generally uncertain. A surgery, initiation of a pharmaceutical treatment, and an exercise program are examples of

health prospects. For simplicity of exposition, this entry considers only binary prospects, prospects involving two outcomes. Such binary prospects are all that is needed to measure health utility. Figure 1 shows a health prospect [.45: 24; 2] that offers different survival durations, 24 years and 2 years with associated probabilities .45 and .55 = –.45.

When utilities are known, it is a widely held view that expected utility is normative and is thus the appropriate approach for assigning value to a prospect such as that in Figure 1. This approach can then inform decisions about optimal treatments and cost-effectiveness.

Utilities, however, are generally not available without some sort of elicitation from a respondent. Elicitation of utility requires that numbers be associated with prospects such that preference for those prospects is faithfully described by the numbers. Behavioral research on choice strongly suggests that people typically do not make choices that conform to expected utility. Thus, when expected utility is used as a measurement tool, it is often the case that the numbers assigned to the prospects do not describe preference well. In addition, research has shown that the effects of expected utility violations are not limited to health utility measurements but also can influence willingness to pay for reductions in health risks and other contingent valuation responses. To address the problems associated with expected utility theory, nonexpected utility theories have been introduced. This entry focuses on the most important of these nonexpected utility theories and prospect theory, first proposed by Daniel Kahneman and Amos Tversky in 1979 and later refined by them in 1992.

Prospect Theory

Prospect theory relaxes those expected utility assumptions that are frequently violated by decision makers. An important advantage of prospect theory is that it offers improvements over expected utility in estimating the value of treatments for application in medical decision models. Once the prospect theory utilities are known, they can then be applied within normative medical decision models to identify treatments that maximize utility. A brief review of expected utility and its relation to prospect theory concepts facilitates an understanding of how prospect theory generalizes and improves expected utility.

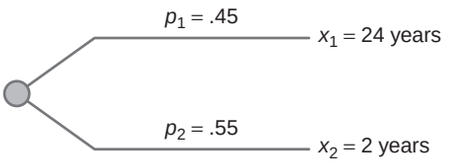


Figure 1 A typical binary health prospect

Expected Utility

Expected utility evaluates prospects by multiplying the utility of each outcome by its associated probability and then summing over this product. For a binary health prospect $[p: x; y]$, the expected utility is $pU(x) + (1 - p)U(y)$, where U is a utility function and an interval scale. This shows that in expected utility, the decision weight assigned to an outcome is equal to its probability.

Nonexpected Utility and Transformation of Probability

Expected utility assumes that preferences are linear in probability. A change in probability from, say, .53 to .54 is given the same weight as a change from 0 to .01 or from .99 to 1. Empirical evidence suggests, however, that people are much more sensitive to the latter two changes than to the former. To model this, prospect theory allows for probability weighting. The probability weighting function w yields a nonlinear transformation of probabilities. The function w is a map from [0, 1] to [0, 1] that is increasing in its argument and for which $w(0) = 0$ and $w(1) = 1$. Empirical work suggests that the function w is often an inverse S shape such that small probabilities of the better outcome are overweighted and large probabilities of the better outcome are underweighted, as in Figure 2. Incorporating probability weighting into expected utility implies that the prospect $[p: x; y]$, $x \geq y$ should be evaluated as

$$w(p)U(x) + (1 - w(p))U(y).$$

This formula corresponds to Quiggin's rank-dependent utility theory, which as shown below is a special case of prospect theory.

Reference Level and Utility

In addition to the utilization of decision weights, empirical studies have also shown that people's

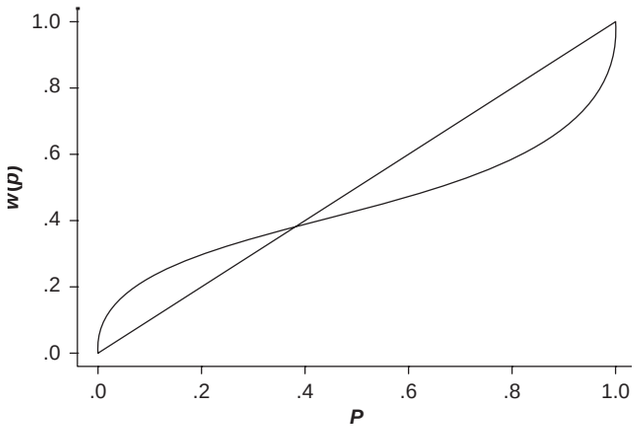


Figure 2 An inverse S-shaped probability weighting function

preferences depend on a reference level. A reference level is a point in the outcome space from which final outcomes are judged. For example, if "fair health" is your reference level, then other outcomes such as "poor health" and "good health" are judged from the "fair health" reference point. Expected utility and rank-dependent utility do not consider reference levels, but prospect theory accommodates reference levels. In prospect theory, those final outcomes preferred to the reference level are coded as gains, and those outcomes not preferred to the reference level are coded as losses. Losses and gains are treated differently. A well-established empirical finding of Kahneman and Tversky is that "losses loom larger than gains," and thus a constant change in outcome represents a larger absolute value difference if it is coded as a loss than if it is coded as a gain. Figure 3 illustrates this point.

Difference in the shape of the value function for gains and losses in risky choices has been observed in health by Lia Verhoef, Anton De Haan, and Willem Van Daal. Furthermore, in prospect theory, probability weighting can be different for gains and for losses.

Because prospect theory distinguishes gains from losses, the evaluation of a prospect depends on the sign of its outcomes. If a prospect $[p{:}x;y]$, $x \geq y$ involves only gains, then its evaluation is identical to rank-dependent utility with probability weighting function w^+, where the + stands for gains. If a prospect $[p{:}\ x;\ y]$, $-x \geq -y$ involves only losses, then its evaluation is the dual of rank-dependent utility:

$$w^-(p)U(x) + (1 - w^-(p))\ U(y),$$

where w^- is the probability weighting function for losses. The most interesting case occurs when the prospect $[p{:}\ x;\ y]$ is *mixed: x* is a gain and *y* a loss. Then, the evaluation is

$$w^+(p)U(x) + \lambda w^-(1 - p)U(y),$$

where λ is a parameter that reflects loss aversion.

The above equations show that if $\lambda = 1$ and $w^+(p) = 1 - w^-(1 - p)$, then prospect theory is equivalent to rank-dependent utility. If $\lambda = 1$ and $w^+(p) = w^-(p) = p$, then prospect theory reduces to expected utility. Hence, rank-dependent utility and expected utility are both special cases of prospect theory.

Functional Forms of the Probability Weighting Function

The numeric value of decision weights may be determined by assuming a parametric form to the probability weighting function. Commonly used probability weighting functions and their parameters are given in Table 1.

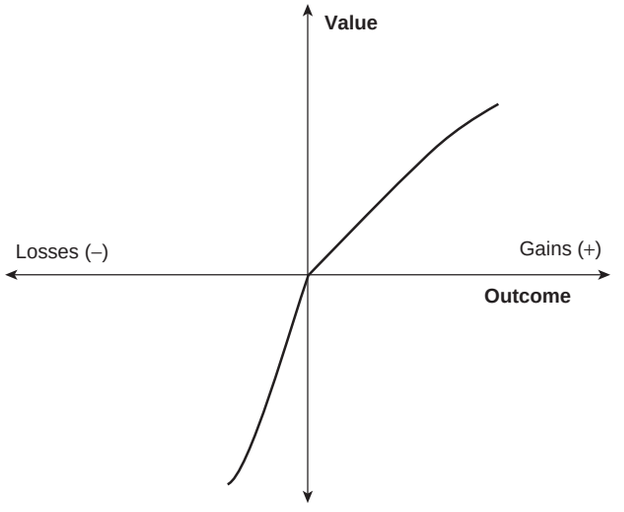


Figure 3 The value function under prospect theory for losses and gains

Source: Adapted from Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica*, *47,* 263–291.

Table 1 Empirical studies on the probability weighting function

Functional Form	*Parameter Estimates*
$w(p) = \frac{p^{\gamma}}{[p^{\gamma} + (1-p)^{\gamma}]^{1/\gamma}}$	Tversky and Kahneman (1992): γ = .61 (gains), γ = .69 (losses) Camerer and Ho (1994): γ = .56 (gains) Wu and Gonzalez (1996): γ = .71 (gains) Abdellaoui (2000): γ = .60 (gains), γ = .70 (losses)
$w(p) = \frac{\delta p^{\gamma}}{\delta p^{\gamma} + (1-p)^{\gamma}}$	Wu and Gonzalez (1996): δ = .84, γ = .68 (gains) Gonzalez and Wu (1999): δ = .77, γ = .44 (gains) Tversky and Fox (1995): δ = .77, γ = .69 (gains) Abdellaoui (2000): δ = .65, γ = .60 (gains) Abdellaoui (2000): δ = .84, γ = .65 (losses)
$w(p) = \exp(-(-\ln p)^{\alpha})$	Wu and Gonzalez (1996): α = .74 (gains)

Source: Adapted from Bleichrodt, H., & Pinto, J. L. (2000). A parameter-free measurement of the probability weighting function in medical decision analysis. *Management Science, 46*(11), 1485–1496.

Framing

Framing refers to how prospects are described. The same prospect can be described in different ways. A framing effect means that two different descriptions of the same prospect occasion different preferences from the respondent. For example, persons are more likely to prefer a medical treatment if they are told that of those who accept the treatment "90% survive" than if they are told "10% die." Prospect theory applies to framed prospects and, thus, is a theory of the valuation process. Rank-dependent utility does not accommodate a framing process. Evidence suggests that in addition to probability weighting, framing may explains differences in utility elicitation methods. Framing has been employed to occasion a greater rate of healthy behaviors, such as smoking cessation, mammography, and sunscreen use among health consumers. However, such applications have been construed as controversial because they may threaten consumer sovereignty.

Prospect Theory and the Standard Gamble

In applied studies, most utility elicitations under risk involve the standard gamble: A method whereby respondents choose between a prospect that will return them to "full health" with probability p, otherwise resulting in immediate "death"; and another prospect that will leave them in some suboptimal health state, Q, with certainty. When a respondent finds the two options equivalent in preference, p is called the *standard gamble equivalent*. Expected utility dictates that $U(Q) = p$, when p is the standard gamble equivalent. Rank-dependent utility dictates that $U(Q) = w(p)$ in this case. With rank-dependent utility, one may choose any of the weighting function and associated parameter estimates in Table 1 to transform p, for the purposes of identifying the utility of Q. Empirical research suggests that under prospect theory, persons frame the certain option, health state Q, as their reference level, and the risky option as a "mixed" prospect, where "full health" is coded as a gain and "death" is coded as a loss. By this analysis, prospect theory requires that when p is the standard gamble equivalent,

$$U(Q) = \frac{w^{+}(p)}{w^{+}(p) + \lambda w \quad (1-p)}.$$

Because the above equation is somewhat complicated to implement in practice, Han Bleichrodt, Jose Luis Pinto, and Peter Wakker have provided a table, such as the one given in Table 2, which converts standard gamble equivalents directly into prospect-theory-based utilities. The utility function U for each of these utility theories is an interval scale and appropriate for cost-utility analysis.

Table 2 Corrected standard gamble utilities as proposed by Bleichrodt et al. (2001) for standard gamble elicitations between 0.00 and 0.99

	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.000	0.025	0.038	0.048	0.057	0.064	0.072	0.078	0.085	0.091
0.1	0.097	0.102	0.108	0.113	0.118	0.123	0.128	0.133	0.138	0.143
0.2	0.148	0.152	0.157	0.162	0.166	0.171	0.176	0.180	0.185	0.189
0.3	0.194	0.199	0.203	0.208	0.213	0.217	0.222	0.227	0.231	0.236
0.4	0.241	0.246	0.251	0.256	0.261	0.266	0.271	0.276	0.281	0.286
0.5	0.292	0.297	0.303	0.308	0.314	0.320	0.325	0.331	0.337	0.343
0.6	0.350	0.356	0.363	0.369	0.376	0.383	0.390	0.397	0.405	0.412
0.7	0.420	0.428	0.436	0.445	0.454	0.463	0.472	0.481	0.491	0.502
0.8	0.512	0.523	0.535	0.547	0.560	0.573	0.587	0.601	0.617	0.633
0.9	0.650	0.669	0.689	0.710	0.734	0.760	0.789	0.822	0.861	0.911

Source: Adapted from Bleichrodt, H., Pinto, J. L., & Wakker, P. P. (2001). Making descriptive use of prospect theory to improve the prescriptive use of expected utility. *Management Science, 47,* 1498–1514.

Note: Row headings represent tenths, column headings hundredths of the uncorrected standard gamble score, and table entries are corrected scores, e.g., the corrected utility for a standard gamble of .15 is .123 (underlined).

Prospect Theory and Quality-Adjusted Life Years

The quality-adjusted life-years (QALYs) model is the most widely used outcome measure in economic evaluations of healthcare. QALYs are computed by adjusting each year of life by the quality of life in which it is spent. An important empirical question is whether QALYs are a valid reflection of people's preferences over health. Previous studies analyzed this question under expected utility. Empirical tests of the QALY assumptions are then confounded with violations of expected utility, however. Prospect theory can also be used as a foundation for QALYs. Tests of the validity of QALYs under prospect theory have tended to be more favorable toward QALYs.

Jason N. Doctor and Han Bleichrodt

See also Allais Paradox; Bias; Choice Theories; Decision Psychology; Decision Weights; Expected Utility Theory; Gain/Loss Framing Effects; Lottery; Prospect Theory; Rank-Dependent Utility Theory; Risk Aversion; Utility Assessment Techniques; Value Functions in Domains of Gains and Losses

Further Readings

Bleichrodt, H., & Eeckhoudt, L. (2006). Willingness to pay for reductions in health risks when probabilities are distorted. *Health Economics, 15,* 211–214.

Bleichrodt, H., & Pinto, J. L. (2005). The validity of QALYs under nonexpected utility. *The Economic Journal, 115,* 533–550.

Bleichrodt, H., Pinto, J. L., & Wakker, P. P. (2001). Making descriptive use of prospect theory to improve the prescriptive use of expected utility. *Management Science, 47,* 1498–1514.

Camerer, C., & Ho, T. H. (1994). Nonlinear weighting of probabilities and violations of the betweenness axiom. *Journal of Risk and Uncertainty, 8,* 167–196.

Doctor, J. N., Bleichrodt, H., Miyamoto, J. M., Temkin, N., & Dikmen, S. (2004). A new and more robust test of QALYs. *Journal of Health Economics, 23,* 353–367.

Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica, 47,* 263–291.

Quiggin, J. (1981). Risk perception and risk aversion among Australian farmers. *Australian Journal of Agricultural Economics, 25,* 160–169.

Stalmeier, P. F. M., & Bezembinder, T. G. G. (1999). The discrepancy between risky and riskless utilities. *Medical Decision Making, 19,* 435–447.

Starmer, C. (2000). Developments in non-expected utility theory: The hunt for a descriptive theory of decision under risk. *Journal of Economic Literature, 28,* 332–382.

Tversky, A., & Kahneman, D. (1992). Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and Uncertainty, 5*(4), 297–323.

Verhoef, L. C. G., De Haan, A. F. J., & Van Daal, W. A. J. (1994). Risk attitude in gambles with years of life: Empirical support for prospect theory. *Medical Decision Making, 14*(2), 194–200.

Wakker, P., & Stiggelbout, A. (1995). Explaining distortions in utility elicitation through the rank-dependent model for risky choices. *Medical Decision Making, 15*(2), 180–186.

Noninferiority Testing

See Equivalence Testing

Number Needed to Treat

The number needed to treat (NNT) is defined as the number of patients that would need to be treated to prevent an adverse outcome in one additional patient compared with control treatment over a specified time period. The term was first introduced in 1988 by Andreas Laupacis, David Sackett, and Robin Roberts as the *number needed to be treated* but has been shortened to *number needed to treat* or its abbreviation NNT.

In medicine and other clinical practices, the well-performed randomized clinical trial is often considered the gold standard for judging the effectiveness of therapeutic interventions. The outcome measures from such studies are reported in many ways that involve both group and individual patient outcomes. An example of the former might be the mean reduction in blood pressure in the group of patients receiving one antihypertensive medication compared with the mean reduction in the group receiving another drug or a placebo. Such grouped data outcomes have often been the evidence for making clinical decisions. It makes sense, at first glance, that if all other factors (e.g., cost, access, side effects) are equal between the two antihypertensive medications, patients in the group with the average lower blood pressure would benefit from the medication compared with patients given the other drug. Missing from such analysis, however, is whether, in fact, individual patients benefited in clinically meaningful ways and how many individuals benefited. The researchers could have alternatively reported the results as the number or percentage of individuals in each medication group with outcome blood pressures in the normal range.

Largely through the development of the field of evidence-based medicine, there has been and continues to be an understanding of the unfavorable impact on the practice of medicine that can result not only from methodologically flawed clinical trials but also from the failure to report treatment effects in clinically relevant outcomes and in terms of individual patient responses.

In the current literature, when clinically relevant outcomes are reported for individual patients, randomized controlled trials and systematic reviews frequently report the treatment effect as relative risk (RR), relative risk reduction (RRR), absolute risk reduction (ARR), or the number needed to treat (NNT). This entry focuses on NNT and its clinical utility relative to the other measures, as well as the strengths and weakness of the NNT as a measure of clinical effect. Additionally, useful resources for calculating the NNT and its precision are provided.

Calculations

The NNT is the reciprocal of the ARR, where the ARR is the simple mathematical difference between the control event rate and the experimental event rate. The ARR has been termed the *benefit* of the treatment. Some researchers use the term *rate difference* or *risk difference* for the same calculation. Table 1 summarizes the calculation of event rates, the ARR, and the NNT.

The NNT when calculated in a clinical study or review provides a point estimate or average of the number of patients that, if given the new treatment, would result in a reduction of one adverse event *over and above the control event rate*. The precision (or variability) of this point estimate can be calculated as confidence intervals (CI) around this point estimate and provides additional information for the clinician as to whether to recommend (and for the patient as to whether to accept) the treatment.

The confidence interval for the ARR is calculated from the standard error for the two proportions (control event rate and experimental event rate). The formula is

$$95\%\ \text{CI} = \text{ARR} + 1.96\ \text{sq root}\ [p_1(1 - p_1)/n_1 + p_2(1 - p_2)/n_2)],$$

Table 1 Calculation of event rates, absolute risk reduction (ARR), and the number needed to treat (NNT)

		Poor Outcome Event	
		Present	*Absent*
Treatment Group	Experimental	*A*	*B*
	Control	*C*	*D*

Experimental event rate (EER) = $A/(A + B)$
Control event rate (CER) = $C/(C + D)$
Absolute risk reduction (ARR) = CER − EER
Number needed to treat (NNT) = 1/ARR

where p_1 is the CER; p_2 is the EER; n_1 is the number of subjects in the control group; and n_2 is the number of subjects in the experimental group.

The confidence interval for the NNT is calculated as the reciprocal of the upper and lower limits of the confidence interval for the ARR. These numbers provide the clinician with a range from the fewest to the largest number of patients needed to be treated to see treatment effect in one. In practice, 95% confidence intervals are routinely calculated, and the upper limit of the 95% confidence interval of the NNT is the clinically more conservative value; that is, the clinician can intuitively be 95% sure that if he or she treats this many patients, one patient will have benefit over and above the controls. The values for the confidence interval around the ARR also indicate if the result is statistically significant; that is, the values do not include zero.

Table 2 gives a numerical example from a study of oral ondansetron given to children in an emergency room setting to reduce vomiting during oral rehydration in uncomplicated gastroenteritis. In this example, the NNT is 5 (usually rounded up to the closest whole number) with a 95% confidence interval from 3 to 11 patients.

In an effort to extend the concept of NNT from randomized controlled trials to systematic reviews and cohort analysis, there have been formulas developed for calculating the NNT from RR and odds ratios (ORs). Sackett and colleagues introduced the term *patient-expected event rate* (PEER) to facilitate these calculations for an individual patient. The PEER is an estimate of an individual's risk of the adverse event if that patient is in the control therapy group. The formulas for calculating the NNT for results expressed as RR or ORs are shown below.

$$\text{For RR} < 1\text{: NNT} = 1/(1 - \text{RR}) \times \text{PEER}.$$

$$\text{For RR} > 1\text{: NNT} = 1/(\text{RR} - 1) \times \text{PEER}.$$

$$\text{For OR} < 1\text{: NNT} = 1 - [\text{PEER} \times (1 - \text{OR})]/(1 - \text{PEER}) \times (1 - \text{OR}).$$

$$\text{For OR} > 1\text{: NNT} = 1 + [\text{PEER} \times (\text{OR} - 1)]/(1 - \text{PEER}) \times (\text{PEER}) \times (\text{OR} - 1).$$

In the literature, there are a number of terms that have been used and are identical to ARR and NNT in calculation but not in terminology. If improvement (benefit) is measured, rather than an adverse event, the term *absolute benefit increase* is often used, and its reciprocal the NNT is then descriptively the number of patients needed to treat to see improvement in one, over and above the control event rate. Similarly, the NNT to prevent one adverse outcome has been extended to looking at the harmful effects from comparative treatments. When the adverse event rates of two treatments are measured, the number needed to harm can be calculated from the reciprocal of the difference in the harmful event rates between experimental and control groups.

Why Number Needed to Treat?

Although not universally accepted as the gold standard for clinical effect, the NNT has become the

Table 2 Calculation of the number needed to treat: A clinical example

		Vomiting During Oral Rehydration	
		Present	*Absent*
Treatment Group	Ondansetron	15	92
	Placebo	37	70

Experimental event rate (EER) = $A/(A + B)$ = 15/107 = 14%
Control event rate (CER) = $C/(C + D)$ = 37/107 = 35%
Absolute risk reduction (ARR) = CER – EER = 35% – 14% = 21% [9.4%, 31.7%]
Number needed to treat (NNT) = 1/ARR = 1/.21 = 4.9 [3.2, 10.6]

Source: Freedman, S. B., Adler, M., Seshadri, R., & Powell, E. C. (2006). Oral ondansetron for gastroenteritis in a pediatric emergency department. *New England Journal of Medicine, 354,* 1698–1705.

standard for those who subscribe to classical evidence-based medicine, for both appraisal of primary randomized clinical trials and for reporting results of systematic reviews. For clinicians (and patients), the appeal of the NNT is that it calculates a result that is intuitively understandable, since it is reported in a number of patients, unlike other measures of treatment effect such as RR, RRR, and OR. Formulas for calculating the RR, RRR, and OR are given in Table 3. The RR and the RRR are both calculated percentages of the control event rate, and as such they are anchored and dependent on that value.

Relative values (RR and RRR) are dependent on the control event rate and may be misleading, even if numerically large, regarding the absolute or attributable benefit. Table 4 gives an illustration showing how the NNT varies, while RRR and RR are constant. As illustrated in Table 4, without an explicit understanding of the control event rate, the RR and RRR do not provide the clinician an understanding of the impact of alternative treatments for his or her patients.

A clinician when proposing a treatment (and the patient when deciding on treatment options) seeks clarity on how much more effective one treatment is compared with other alternatives, and relative efficacy does not fully address this issue, particularly when cost, access to care, risk of harm, and other factors from alternative treatment options are not equivalent. The NNT provides a more intuitive, contextual understanding of the impact of comparative treatments in these situations.

The same understanding has been used by public health practitioners for calculating both efficacy and harm and by third-party payers to assess cost as a factor for expanding or restricting treatment options.

Limitations

The value of the NNT in clinical practice is limited both by the specific characteristics of the calculated value and also by penetration of this concept into the clinical practice literature. Three specific criteria define each NNT value: (1) the particular outcome measured, (2) the baseline risk (control event rate), and (3) the time period for measuring the outcome. If any one of these criteria changes, then the NNT will also vary. The NNT therefore cannot be compared, one study with another, unless these three criteria are similar.

The specificity of the outcome measure, in terms of an individual's response, forces researchers to make choices regarding the presentation of their data. The clinician then is faced with a decision regarding the applicability of that outcome for his or her patient. For example, a number of drug efficacy studies for treating migraine headaches use an outcome defined as 50% reduction in the number of migraine headache events in a specified period of time. If, as in one study, the NNT is 4, the clinician reading the research article may wonder whether that number reflects both the participants in the study with baseline (control event rate) headache frequency of 2 per month and also

Table 3 Calculation of relative risk, relative risk reduction, and odds ratio

		Poor Outcome Event	
		Present	*Absent*
Treatment Group	Experimental	*A*	*B*
	Control	*C*	*D*

Relative Risk (RR) = $(A/A + B)/(C/C + D)$ = EER/CER
Relative Risk Reduction (RRR) = (CER − EER)/CER = ARR/CER
Odds Ratio (OR) = $(A/B)/(C/D)$

Source: Adapted from Straus, S. E., Richardson, W. S., Glasziou, P., & Haynes, R. B. (2005). *Evidence-based medicine: How to practice and teach EBM* (3rd ed.). New York: Elsevier.

Table 4 A comparison of absolute and relative measures of treatment effect

EER	*CER*	*RR*	*RRR*	*ARR*	*NNT*
10%	30%	33%	67%	20%	5
1%	3%	33%	67%	2%	50
.1%	.3%	33%	67%	.2%	500

Source: Adapted from Table 2 in Laupacis, A., Sackett, D. L., & Robert, R. S. (1988). An assessment of clinically useful measures of the consequences of treatment. *New England Journal of Medicine, 318,* 1728–1733.

those who experience headaches with greater frequency, such as 10 headaches per month, as well as what the NNT would be for a lesser reduction, such as 25%, or greater reduction, such as 75%, from a host of other measures.

Understanding the time period for the specific outcome is particularly important for outcomes of a chronic disease or events that occur late in the natural course of a disease, for example, the development of a stroke in patients with mild hypertension.

There is no absolute, definitive clinically useful NNT. It always depends on the adverse outcome, the treatment (and its attendant risks), and time of disease progression to adverse outcome.

An important practical limitation of the NNT has to do with the current editorial practice of journals regarding clinical articles. Although the CONSORT (Consolidated Standards on Reporting Trials) group has made a recommendation to report results as NNT or ARR, few journals have adopted such reporting. In one review by Jim Nuovo and colleagues, only 18 of 359 randomized clinical trials reported NNT or ARR. As a result, few practitioners have the opportunity to consider clinical decisions in terms of NNTs.

It is often possible to extract the information needed to calculate the NNT from journals if it is not explicitly provided. Although the calculations are relatively simple, busy practitioners probably are not going to make these calculations themselves, particularly with the confidence intervals. To aid clinicians, there are a number of resources both Web-based and for PDA (Personal Digital Assistant) that are available to perform and interpret results from randomized clinical trials, systematic reviews, and cohort studies.

Jerry Niederman and Jordan Hupert

See also Complications or Adverse Effects of Treatment; Effect Size; Evidence-Based Medicine; Informed Decision Making; Odds and Odds Ratio, Risk Ratio; Randomized Clinical Trials; Treatment Choices

Further Readings

Centre for Evidence-Based Medicine. EBM Calculator v1.2: http://www.cebm.utoronto.ca/palm/ebmcalc

EBM and Decision Tools by Alan Schwartz: http://araw.mede.uic.edu/~alansz/tools.html

Guyatt, G., Rennie, D., Meade, M. O., & Cook, D. J. (2002). *Users' guides to the medical literature: Manual for evidence-based clinical practice.* Chicago: American Medical Association Press.

Heneghan, C., & Badenoch, D. (2006). *Evidence-based medicine toolkit.* Malden, MA: Blackwell.

Laupacis, A., Sackett, D. L., & Roberts, R. S. (1988). An assessment of clinically useful measures of the consequences of treatment. *New England Journal of Medicine, 318,* 1728–1733.

McQuay, H. J., & Moore, R. A. (1997). Using numerical results from systematic reviews in clinical practice. *Annals of Internal Medicine, 126,* 712–720.

Nuovo, J., Melnikow, J., & Chang, D. (2002). Reporting number needed to treat and absolute risk reduction in randomized controlled trials. *Journal of the American Medical Association, 287,* 2813–2814.

Straus, S. E., Richardson, W. S., Glasziou, P., & Haynes, R. B. (2005). *Evidence-based medicine: How to practice and teach EBM* (3rd ed.). New York: Elsevier.

Numeracy

Numeracy, also known as numerical or quantitative literacy, refers to an ability to reason with numbers and other mathematical concepts. The word was first introduced in 1959 by the UK Committee on Education, presided over by Sir Geoffrey Crowther. Innumeracy is a lack of numeracy.

Definitions and Prevalence

In 2000, the Department of Health and Human Services defined health literacy as the skills needed to "obtain, process, and understand basic health information and services for approximate health decisions." The National Center for Education Statistics (NCES), as part of the U.S. Department of Education's Institute of Education Sciences (IES), collects, analyzes, and publishes statistics on education and public school district finance information in the United States. The NCES in its Adult Health Literacy Survey in 1993 defines numeracy or quantitative literacy as "the knowledge and skills required to apply arithmetic operations, either alone or sequentially, using numbers embedded in printed material (e.g., balancing a checkbook, completing an order form)." In 2003, more than 19,000 adults participated in the national- and state-level assessments, representing the entire population of U.S. adults who are aged 16 and older, most in their homes and some in prisons from the 50 states and the District of Columbia. Approximately 1,200 inmates of federal and state prisons were assessed to provide separate estimates of literacy for the incarcerated population. The National Adult Literacy Survey (NALS), a nationally representative household survey administered by the NCES, showed that 22% of all American adults surveyed exhibited the lowest level of numeracy. The prevalence of below basic numeracy skills is associated with race/ethnicity: 13% of whites, 47% of blacks, and 50% of Hispanics exhibit below basic numeracy skills.

Health Numeracy and Measures of Numeracy

Health numeracy is emerging as an important concept and a component of health literacy. Use of health numeracy is increasing in health communication, for much of the health information has been provided to patients and written in numbers—such as diagnostic test results and treatment, prognosis, and medication regimens. For the majority of the medical decision making, when discussing risks and benefits, providers use simple ratios, probabilities, and estimates to communicate with patients. Numeracy is assessed by different measures and has been associated with poor outcomes (e.g., diabetes, nutrition, obesity, and asthma). Lower numeracy has been associated with increased medicine errors and increased hospitalizations. The original Test of Functional Health Literacy in Adults (TOFHLA) in 1995 included 17 questions that measured numeracy. Lisa Schwartz and colleagues and Isaac Lipkus and colleagues introduced and modified comprehensive numeracy scales, which measured simple numeracy skills of percentages, proportions, and frequencies.

Individual quantitative competencies can be categorized into three basic levels: (1) basic computation, (2) estimation, and (3) statistical literacy.

Basic computation includes number recognition and comparison, arithmetic, and use of simple formulas. The TOFHLA screening test and the NALS include quantitative problems ranging from abstract problem solving to proportions and frequencies using medical scenarios, such as pill count, understanding sliding scale prescriptions, and nutrition labels. Estimations are used for quick calculations to estimate medication dosage, such as insulin units in diabetes. Statistical literacy is an understanding of concepts such as chance and uncertainty, sampling variability, and margins of errors. A large variety of medical scientific information is in forms of graphics with scales, bars, ratios, and so on, which might help patients understand epidemiological distribution of any disease, chances and uncertainty associated with a disease occurrence, and outcome or use in medical decision making for choice of treatment.

Use in Medical Decision Making

Evaluating Patients

Providers in healthcare use quantitative values when interviewing patients to understand the symptoms. For example, to understand a simple chief complaint of pain, providers use phrases such as *How much pain? How frequent is the pain?* and *How intense is this pain?* To answer these questions, patients are supposed to use simple or abstract values to rate the severity/frequency or intensity of pain using a number *X*/10.

Discussion of Disease Process

In a healthcare setting, providers often use numbers in the form of probabilities and frequencies to explain the likelihood of a disease or its progression and prognosis when a patient presents with a sign or a symptom. For example, women with low numeracy may be unable to comprehend the concept of breast cancer occurrence, risk perception, and breast cancer screening.

Risk and Benefit

Data on assessing the accuracy of a laboratory test, certainty of a diagnostic imaging, potential risks and benefits of a drug, or comparing treatments are frequently presented using odds ratios, risk ratios, relative ratios, or confidence intervals. Patients with low numeracy may get overwhelmed with all the numbers and statistical data and may have greater difficulty understanding recommendations. For example, patients with diabetes and obesity having low educational attainment can struggle to interpret food labels and can underestimate or overestimate the nutrient facts. With advancement in technology and increased use of insulin pumps, it is of the utmost importance for patients to be able to comprehend food labels and compute calorie intake based on nutrient facts.

People with variable numeracy may need different decision tools. Hence, it is necessary to assess the numerical capacity or quantitative literacy of patients to be able to deliver appropriate healthcare information and to help them make a genuine informed decision.

Arpita Aggarwal

See also Models of Physician–Patient Relationship; Patient Decision Aids; Patient Rights

Further Readings

Lipkus, I. M., Samsa, G., & Rimer, B. K. (2001). General performance on a numeracy scale among highly educated samples. *Medical Decision Making, 21,* 37–44.

The National Assessment of Adult Literacy: http://nces.ed.gov/naal

Parker, R. M., Baker, D. W., Williams, M. V., & Nurss, J. R. (1995). The test of functional health literacy in adults: A new instrument for measuring patients' literacy skills. *Journal of General Internal Medicine, 10,* 537–541.

Paulos, J. A. (1990). *Innumeracy: Mathematical illiteracy and its consequences*. New York: Hill & Wang.

Rothman, R. L., Montori, V. M., Cherrington, A., & Pignone, M. P. (2008). Perspective: The role of numeracy in healthcare. *Journal of Health Communication, 13*(6), 583–595.

Schwartz, L. M., Woloshin, S., Black, W. C., & Welch, H. G. (1997). The role of numeracy in understanding the benefit of screening mammography. *Annals of Internal Medicine, 127,* 966–972.

U.S. Department of Health and Human Services. (2000). *Healthy people*. Washington, DC: Government Printing Office.

Odds and Odds Ratio, Risk Ratio

The main goal of many medical studies is to evaluate the effect of a treatment or the risk of disease under given conditions. This entry introduces and discusses measures of effect and/or risk when the factors or variables of interest are categorical (i.e., nominal) in nature. The discussion focuses on the case of dichotomous variables (factors), that is, those that take only two values that often indicate the presence or absence of a characteristic (disease, treatment, exposure, etc.). The data consist of information on n individuals who have been categorized according to the presence or absence of two factors. The information is presented in a 2×2 contingency table like the one in Table 1.

The numbers in the table denote the frequency of each cell. For example, a is the number of individuals for whom both factors were present, and b those for whom A was present and B was absent. Thus, the total number of patients is $n = a + b + c + d$. Common types of factors are exposure, treatment, and disease. The statistical significance of the association between the two factors is tested using the chi-square test (or the likelihood test). The result of the test is a p value, which measures the chance of the observed relationship under the assumption that there is none. Accordingly, a small p value (e.g., $< .05$) leads to a "significant" result and the rejection of the assumption of no association. The size of the p value is determined, in a critical way, by the sample size and can't be used to assess the strength of the association. When the association (or effect, when one factor is the "cause" of the other) is significant, the measure of its strength is critical. Below are the most common measures of association.

Definition 1. The relative risk (RR) is defined as

$$\mathrm{RR} = \frac{\Pr[\text{B present if A is present}]}{\Pr[\text{B present if A is absent}]} = \frac{Risk(\text{A present})}{Risk(\text{A absent})} = \frac{a/(a+b)}{c/(c+d)} = \frac{a(c+d)}{c(a+b)},$$

where Pr[ ] indicates probability.

Definition 2. The risk difference (RD), also called attributable risk, is defined as

$$\mathrm{RD} = Risk(\text{A present}) - Risk(\text{A absent}) = \frac{a}{a+b} - \frac{c}{c+d}.$$

Definition 3. The odds ratio (OR) is defined as

$$\mathrm{OR} = \frac{Odds(\text{Present})}{Odds(\text{Absent})} = \frac{a/b}{c/d} = \frac{ad}{bc}.$$

Table 1 The 2×2 contingency table

		Factor B		
		Present	*Absent*	*Total*
Factor A	Present	a	b	$a + b$
	Absent	c	d	$c + d$

Comparing Measures

As mentioned, the significance of the association is tested using the chi-square test (or the likelihood test). However, when it comes to measuring the strength of the association, there is not only a lack of consensus but also considerable confusion. The three measures defined above assess the strength of the relationship but do so in different ways. The confusion in the interpretation of these measures is caused by the fact that each one is based on ratios (relative frequencies or odds). So, in the case of RR and OR, the reader is faced with the daunting task of interpreting ratios of ratios. It is therefore crucial to understand the differences and, consequently, the correct way to interpret these measures.

The RR is a relative measure interpreted as a percentage. It is important to note that both RR and 1/RR are relative risks and assess the strength of an association from a different point of view. The choice of denominator, which is not always obvious, determines the value and the interpretation. For example, RR = 1.30 describes a risk increase of 30%, whereas 1/RR = .77 is interpreted as a 23% decrease in risk. The key is to realize that RR describes a 30% risk increase of exposure (i.e., A present) relative to nonexposure (i.e., A absent), whereas 1/RR represents a 23% risk decrease in nonexposure relative to exposure. Without careful description, the interpretation of the value will depend on whether RR or 1/RR is used, when, in fact, they are measuring the same association. In general, relative measures such as RR and OR are not easy to interpret and can be misleading. The first two columns of Table 2 contain combinations of risks that yield the same value of RR, but represent very different situations.

The RD is easier to understand because it removes a layer of complexity by simply computing the difference between the risks. The RD is the amount the risk of disease increases (or decreases) by the exposure. The advantage of the RD is that it is measured in the risk scale, which makes it more intuitive. The RD can only vary between −1 and 1, and is not distorted when the risks are very high or very low. Another property of the RD that makes it attractive in medical decision making is that its reciprocal is the number needed to treat (NNT). The NNT is defined as the number of patients that need to be treated to achieve one success. Care must be taken to interpret the reciprocal of RD as the NNT, since doing so assumes that the factor investigated is the sole cause of the increase in risk (i.e., that it really is the attributable risk). The RD and the NNT are given in Table 2 for the combinations of risks considered before. Table 2 shows that an exposure that doubles the risk (RR = 2) can be associated with a very small increase in the probability of disease (RD = .00005). This represents a rate of 1 in 10,000 in the exposed group versus 1 in 20,000 in the unexposed group.

The comparison is further complicated by the fact that the points of reference of the two measures are different. That is, an increase in risk is associated with both, RR > 1 and RD > 0. The value of RD is not directly related to that of RR. Actually, the value of one can't be calculated from the other unless the value of at least one risk is known. Sometimes, particularly in the media, only the value of RR (or RD) is published, frequently resulting in faulty interpretation of the results. Whenever possible, both risks (e.g., exposure and nonexposure) should be reported to avoid confusion. Confidence intervals for RR are simple to compute using properties of the logarithm of the ratio of two proportions.

The OR is another common measure of association. The OR is defined as the ratio of odds of exposed over nonexposed. The odds is itself the ratio of the risk of disease over the risk of no disease. The interpretation of odds is not intuitive. For instance, a risk of .3 (3 in 10) is equivalent to odds of .43 (= .3/.7). One advantage of the OR is that it has the same point of reference as the RR. In other words, the OR and the RR are always on the same side of 1, either both greater than 1 or both less than 1, but the OR always overestimates the strength of the association relative to the RR. When the prevalence of the disease is low, the risk of no disease is close to 1, causing the odds to be similar to the risk and thus, the OR to be similar to the RR. The last column of Table 2 illustrates this fact. In practice, the assumption of low prevalence is made, but not always justified, to interpret the OR as the RR. The last three rows of Table 2 show combinations of risks that show that even for low risks (e.g., .2 and .133), the OR can substantially overestimate the RR (in this case from 50% to 63%). Confidence intervals for the OR are easy to

Table 2 Combinations of risks and associated measures

Risk(A Present)	*Risk(A Absent)*	*RR*	*RD*	*NNT*	*OR*
.98	.49	2	.49	2.04	51
.9	.45	2	.45	2.22	11
.7	.35	2	.35	2.9	4.3
.5	.25	2	.25	4	3
.3	.15	2	.15	6.7	2.4
.1	.05	2	.05	20	2.1
.01	.005	2	.005	200	2.01
.001	.0005	2	.0005	2000	2.001
.0001	.00005	2	.00005	20000	2.0001
.3	.2	1.5	.1	10	1.71
.2	.133	1.5	.0867	11.5	1.63
.1	.067	1.5	.033	30.3	1.54

compute and are based on the fact that the logarithm of a ratio is the difference of the logarithms.

Why Use Odds?

After the previous discussion, the obvious question is, "Why use odds and the OR at all?" This is a valid question since these measures are difficult to interpret and do not give any extra insight. One of the reasons for their use is that, in certain situations, the risks cannot be estimated. The case-control design is used frequently to study low prevalence diseases or situations in which random allocation of patients is not possible. This design consists of selecting a sample of patients with the disease (cases) and another sample of patients without the disease (controls). The number of cases and controls are determined by the researcher and thus provide no information about frequencies in the population, including prevalence and risks, and thus, it is not possible to calculate meaningful values for the RR and RD. The OR does not depend on the number of cases and controls and can be computed. Furthermore, as previously discussed, when the prevalence is low, the OR can be used to estimate the RR.

The other reason for the frequent use of odds and the OR is their superior mathematical properties. Among these properties (that risks and RR lack) are the following:

- The range of possible values is from 0 to infinity regardless of the prevalence.
- Reversing the outcomes (good by bad or vice versa) simply changes the OR to its reciprocal.

As a result of these properties, logistic regression—the ubiquitous statistical technique used to analyze categorical responses—works with odds and reports ORs. Logistic regression is powerful because it allows the estimation of adjusted effects, that is, effects of factors when other—possibly confounding—factors are taken into account. The results of a logistic regression analysis are usually reported in terms of adjusted odds and ORs. Clearly, odds and ORs are measures that will continue to be used in the analysis of categorical responses.

Examples

Peberdy and colleagues compared outcomes from in-hospital cardiac arrest during nights/weekends

with those during weekdays/evenings. A total of 86,748 events from the National Registry of Cardiopulmonary Resuscitation from January 2000 to February 2007 were retrieved. Table 3 shows the data on survival to discharge.

The results are reported in terms of the risks, thus avoiding the problem of interpreting relative measures. Also provided is an estimate and confidence interval of the OR (OR = 1.43 [95% CI: 1.38, 1.49]). The results indicate that the discharge survival rate was significantly higher during days/evenings relative to nights/weekends. The best form in which to report the findings is either the RD (= 5.1%) or the RR (= 1.35). The RR represents a 35% increase in the probability of surviving to discharge in days/evenings relative to nights/weekends. Using logistic regression, the article reports an OR = 1.18 (95% CI: 1.12, 1.23) adjusted by multiple factors, including sex, age, race, and illness category.

In another study, D'Souza and colleagues used a case-control design to investigate the relationship between the presence of human papillomavirus (HPV) and oropharyngeal cancer. The hospital-based study included 100 patients with newly diagnosed oropharyngeal cancer and 200 patients without cancer. The results of the presence of any oral HPV infection and oropharyngeal cancer appear in Table 4.

The article reports a significant association between oropharyngeal cancer and oral HPV infection with OR = 12.3 (95% CI: 5.4, 26.4). This OR was adjusted by age, sex, tobacco use, alcohol use, dentition and toothbrushing, and a family history of head and neck cancer. It is clear that the proportion of cases in either the positive or negative classification depends directly on the ratio of cases to controls, which is determined by the researcher. In this case, the assumption of low prevalence is reasonable, so the OR can be interpreted by saying, "The presence of oral HPV infection increases the risk of oropharyngeal cancer by a factor of 12."

Table 3 In-hospital cardiac arrest outcomes by time of day

		Survival to Discharge		
		Yes	*No*	*Total*
Time n (%)	Day/evening	11604 (19.8)	46989 (80.2)	58593
	Night/weekend	4139 (14.7)	24016 (85.3)	28155

Source: Peberdy et al. (2008).

Table 4 Association between the presence of any HPV oral infection and oropharyngeal cancer

		Oropharyngeal Cancer	
		Cases	*Controls*
Any oral HPV infection n (%)	Negative	63 (63)	189 (94)
	Positive	37 (37)	11 (6)
Totals		100	200

Source: D'Souza et al. (2007).

Interpretations

Absolute and relative risks are effective measures of effect or association when the variables involved are dichotomous. However, their interpretation is not always straightforward and can be misleading. The recommended way to report the results is by presenting estimates of the actual risks to avoid confusion. The odds and odds ratio are more difficult to interpret and should be used only when the data are generated by a case-control study or are obtained using logistic regression.

Esteban Walker

See also Case Control; Logistic Regression; Number Needed to Treat

Further Readings

Davies, H. T. O., Crombie, I. K., & Tavakoli, M. (1998). When can odds ratios mislead? *British Medical Journal, 316,* 989–991.

Deeks, J. (1998). When can odds ratios mislead? *British Medical Journal, 317,* 1155–1156.

D'Souza, G., Kreimer, A. R., Viscidi, R., Pawlita, M., Fakhry, C., Koch, W. M., et al. (2007). Case-control study of human papillomavirus and oropharyngeal cancer. *New England Journal of Medicine, 356,* 1944–1956.

Peberdy, M. A., Ornato, J. P., Larkin, G. L., Braithwaite, R. S., Kashner, T. M., Carey, S. M., et al. (2008). Survival from in-hospital cardiac arrest during nights and weekends. *Journal of the American Medical Association, 299,* 785–792.

Oncology Health-Related Quality of Life Assessment

Oncology was one of the first disease areas where trade-offs between quality of life and quantity of life were recognized. Over time, a variety of health-related quality of life (HRQOL) measures have been used to quantify the impacts of cancer and its treatments on the quality of life, and several measures have been developed to assess HRQOL in cancer specifically.

Studies of HRQOL in cancer populations have been used to inform decision making at several levels, including policy making and population monitoring, clinical trials and observational studies, and individual patient–clinician interactions. Various HRQOL measures have been used within and across studies at the different levels of decision making. Thus, researchers and clinicians who are faced with the task of selecting which HRQOL measure to use in a given study or application may be unsure how to proceed.

There has been consideration of identifying a core set of measures that would be appropriate for use across a range of oncology studies. Most recently, the National Cancer Institute's Cancer Outcomes Measurement Working Group (COMWG) investigated the possibility of identifying core measures for oncology HRQOL assessment. However, it was evident early on that the experts in the COMWG did not think that defining such a core set of measures was advisable for several reasons. First, the COMWG felt strongly that the measure selected for a given study or application should be tailored to the specific objectives of that particular study or application. For example, the appropriate measure for population monitoring and policy making may be quite different from the appropriate measure to compare treatment options in a randomized controlled trial. In addition, the COMWG felt that nominating a core set of measures may be premature, given the relative youth of the field of HRQOL assessment. Furthermore, several measures have been used often and successfully in oncology applications, making it difficult to select one particular "winner." Finally, as interest in item response theory and other modern measurement techniques increases, alternative approaches for assessing HRQOL outcomes may be on the horizon.

This entry first reviews the range of HRQOL measures available for use in oncology studies and describes several commonly used and emerging HRQOL measures. Then, considerations for selecting a measure for a given study or application are discussed.

Types of Measures

As mentioned above, a wide range of HRQOL measures are available for use in cancer studies and applications. While, in general, HRQOL measures are categorized as either generic (appropriate for use

across a range of disease and healthy populations) or disease-specific (appropriate for use in patient populations with a specific disease), in cancer, the categories are a bit more complex. Specifically, cancer is not one disease but many. Thus, in cancer, a disease-specific measure could refer to a measure appropriate for use across cancers or a measure developed for a specific cancer (e.g., breast cancer). To differentiate these latter two categories, the term *general cancer measure* is used to refer to a measure appropriate for use across cancer types, and *cancer site-specific measure* is used to refer to a measure designed for use in a specific cancer type. At least two measurement systems have been designed so that general cancer core measures can be supplemented with cancer site-specific measures, the so-called modular approach.

Another dimension for categorizing HRQOL measures is whether they are profile or preference-based measures. Profile measures are scored using psychometric scaling processes (e.g., computing mean scores based on patient ratings on numerical scales), whereas preference-based measures are scored through incorporating either direct or indirect valuations of different health states (e.g., using weightings that reflect the relative values raters ascribe to the different health states). While profile measures are more commonly used in cancer studies than preference-based measures, preference-based measures may be the best approach for certain situations.

HRQOL is generally considered to be a multidimensional concept that includes measures of physical, psychological, and social functioning; however, some measures that are patient-reported focus on a particular outcome (e.g., pain) and are thus unidimensional. For the purposes of this entry, the discussion is limited to multidimensional HRQOL measures, but it is important to keep in mind that unidimensional patient-reported outcome measures may be used instead of, or in addition to, multidimensional HRQOL measures.

When selecting the HRQOL measure for a given study or application, it is not necessary to use only one type of measure. For example, one might choose to use both a profile measure and a preference-based measure, and it is common to combine a generic measure or a general cancer measure with a cancer site-specific measure, potentially using one of the modular systems. Furthermore, as noted above, unidimensional measures may be used to supplement multidimensional measures in cases where certain outcomes are not covered by a multidimensional measure or are of particular importance and warrant more comprehensive or precise measurement. In some cases, researchers may choose to use a battery of unidimensional measures rather than a multidimensional measure.

Commonly Used and Emerging Measures

Some of the more commonly used generic measures in oncology studies include the Medical Outcomes Study Short Form-36 (SF-36) and the Sickness Impact Profile (SIP). Although not yet widely applied in oncology studies, the most commonly used preference-based measures include the EuroQol EQ-5D, the Health Utilities Index, and the Quality of Well-Being Scale.

As mentioned above, there are two modular systems that are commonly used to assess HRQOL in oncology populations. The European Organization for Research and Treatment of Cancer (EORTC) system includes the Quality of Life Questionnaire-Core 30 (QLQ-C30), which can be supplemented with modules for several cancer types. The QLQ-C30 is a 30-item questionnaire that assesses five function domains (physical, role, emotional, social, cognitive), eight symptoms (fatigue, pain, nausea and vomiting, dyspnea, insomnia, appetite loss, constipation, diarrhea), plus financial impact and a global health/quality of life rating. The core measure can be supplemented with modules, which are currently available for the following cancers: breast, lung, head and neck, esophageal, ovarian, gastric, cervical, and multiple myeloma. Modules for other cancer types are currently under development. In addition, the EORTC has developed questionnaires to assess patient satisfaction and palliative care.

The second modular system is the Functional Assessment of Cancer Therapy (FACT) program (which later expanded beyond cancer to be the Functional Assessment of Chronic Illness Therapy—FACIT). The core measure of the FACT system is the FACT-General (FACT-G). It has 27 questions that assess four domains: physical well-being, social/family well-being, emotional well-being, and functional well-being. The FACT-G can be supplemented with modules for the following cancer types: breast, bladder, brain, colorectal, central nervous system, cervix, esophageal, endometrial, gastric, head and

neck, hepatobiliary, lung, leukemia, lymphoma, melanoma, ovarian, prostate, vulva, and brain cancer survivors. The FACT system also has symptom indices for prostate, bladder, brain, breast, colorectal, head and neck, hepatobiliary, kidney, lung, and ovarian cancers. In addition, modules are available to assess the impact of certain treatments (e.g., bone marrow transplantation, biologic response modifiers, and taxanes) and to assess specific symptoms (e.g., anorexia/cachexia, diarrhea, fatigue, lymphedema). A variety of non-cancer-specific measures are also included in the FACIT system, for example, palliative care, spiritual well-being, and treatment satisfaction. The FACT/FACIT system has several additional measures currently under development.

An emerging option for researchers and practitiossners interested in assessing HRQOL in oncology and other disease areas are the measures under development by the National Institutes of Health Patient-Reported Outcomes Measurement Information System (PROMIS). PROMIS is currently creating item banks for various patient-reported outcomes, including emotional distress, physical function, social role participation, pain, and fatigue. Item banks are a collection of items (i.e., questions) from various different measures that assess a particular domain. Items within a bank are calibrated in terms of their properties (e.g., how likely patients are to select a response option based on their level of functioning on the domain). This calibration allows comparison of scores between populations even if they responded to different items from the item bank. Item banks may be used to develop fixed-item measures (questionnaires where all respondents answer the same set of questions) or to develop computer-adaptive assessments (dynamic assessments in which the questions asked and sequence in which the questions are asked are determined by respondents' previous answers). Because of this calibration, fixed-item measures can be tailored to the level of functioning expected to be applicable to the population being studied. With computer adaptive assessments, the calibration facilitates targeting of questions based on a given respondent's level of functioning using his or her previous responses. Thus, these PROMIS item banks, and the fixed-item and adaptive assessments that can be generated from them, represent a powerful new option for assessing HRQOL. However, the PROMIS item banks are just being released now so testing is required to determine how well the PROMIS measures work in practice.

Measure Selection

With this vast array of measure options for assessing HRQOL in oncology, it can be daunting for the researcher or clinician to determine which measure to use in a given study or application. Table 1 provides general suggestions for selecting a measure based on the study or application's objectives and also lists an example. These considerations and how they have been applied in the example studies are discussed in more detail below.

Population Comparisons

When the goal of a study or application is to compare the HRQOL of different population groups for monitoring or policy-making purposes, use of a generic measure is generally advised. For example, the Medicare Health Outcomes Survey is conducted by the Centers for Medicare and Medicaid Services with the National Committee for Quality Assurance to monitor the care and outcomes of Medicare managed care enrollees. This survey uses the SF-36 as an HRQOL measure.

Recently data from the Medicare Health Outcomes Survey have been linked with cancer registry data from the Surveillance, Epidemiology and End Results (SEER) program. This linkage has allowed comparisons of HRQOL across a variety of cancer types and population groups. In some cancers (e.g., colorectal, lung, urinary, kidney, and non-Hodgkin's lymphoma), decrements were found in both physical and mental component summary scores, whereas breast, uterine, and prostate cancer only exhibited decrements on the physical component summary score. In this case, while the SF-36 is not specifically targeted to cancer or the impacts of specific types of cancers, the generic measure provides useful data for comparing different cancer types and potentially for comparisons of cancer patients to the general population.

Intervention Comparisons

In cases where the objective of the study is to compare the impact on HRQOL of different

Table I Considerations and examples of matching the measure to the study objectives

Objective	*Suggestions*	*Example Study*
Comparing HRQOL across different population groups for population monitoring or policy making	Consider commonly used generic measures because of the availability of normative data and the ability to map an array of health conditions within and across populations onto a common continuum	Reeve, B. B., Arora, N. K., Clauser, S. B., Haffer, S. C., Han, P. K., Hays, R. D., et al. (2007). Prospective evaluation of cancer diagnosis and initial treatment on health-related quality of life in cancer patients. 2007 International Society for Quality of Life Research meeting abstracts [www.isoqol.org/2007mtgabstracts.pdf]. *Quality of Life Research supplement,* A-7, Abstract #1278.
Comparing interventions in clinical trials and observational studies	Consider HRQOL measures that are specifically targeted to the outcomes of interest and sensitive to differences between groups	*Clinical Trial*: Wenzel, L. B., Huang, H. Q., Armstrong, D. K., Walker, J. L., & Cella, D. (2007). Health-related quality of life during and after chemotherapy for optimally debulked ovarian cancer: A gynecologic oncology group study. *Journal of Clinical Oncology, 25,* 437–443. *Observational Study*: Hu, J. C., Elkin, E. P., Krupski, T. L., Gore, J., & Litwin, M. S. (2006). The effect of postprostatectomy external beam radiotherapy on quality of life. *Cancer, 107,* 281–288.
Informing comparisons through decision-analytic models and cost-utility studies	Consider preference-based measures because they incorporate the valuation of various outcomes and can be used to calculate quality-adjusted life years	van den Hout, W. B., Kramer, G. W. P. M., Noordijk, E. M., & Leer, J. W. H. (2006). Cost-utility analysis of short- versus long-course palliative radiotherapy in patients with non-small-cell lung cancer. *Journal of the National Cancer Institute, 98,* 1786–1794.
Informing and guiding individual patient-clinician interactions	Consider HRQOL measures that provide clear and interpretable data	Velikova, G., Booth, L., Smith, A. B., Brown, P. M., Lynch, P., Brown, J. M., et al. (2004). Measuring quality of life in routine oncology practice improves communication and patient well-being: A randomized controlled trial. *Journal of Clinical Oncology, 22,* 714–724.

interventions through either randomized controlled trials or observational studies, the recommended approach is to use HRQOL measures that are targeted to the outcomes of interest and sensitive to differences between groups. For example, a recent randomized controlled trial of Stage III ovarian cancer patients compared intraperitoneal with intravenous chemotherapy and included an assessment of patient HRQOL. The intraperitoneal arm had significantly longer survival compared with the intravenous arm but was associated with worse HRQOL. Specifically, this study used the FACT-G, supplemented with the 12-item FACT-Ovarian (FACT-O) subscale, the 11-item FACT/Gynecologic Oncology Group Neurotoxicity subscale, and two items to assess abdominal discomfort that were developed for this study and combined with two items from the FACT-O that assessed concerns relevant to intraperitoneal treatment.

This study found that patients who received intraperitoneal therapy had worse physical and

functional well-being, ovarian cancer symptoms, and abdominal distress during treatment. At 3 to 6 weeks following treatment, intraperitoneal arm patients continued to experience deficits in physical and functional well-being and ovarian symptoms, which were accompanied by worse neurotoxicity. While the neurotoxicity persisted to 12 months posttreatment, most other HRQOL differences between groups had resolved by then. Thus, intraperitoneal chemotherapy is associated with improved overall survival but worse short-term HRQOL. The authors concluded that patients and clinicians should discuss these trade-offs between survival and HRQOL when deciding on a treatment strategy.

The measurement strategy employed in this study was specifically targeted to the outcomes relevant to a comparison of intraperitoneal and intravenous chemotherapy administration in ovarian cancer patients, but also assessed general cancer HRQOL. The results of the study indicated significant and clinically important differences between groups on these HRQOL measures, thus providing important data for patients and clinicians to consider when discussing the treatment options.

HRQOL is also commonly assessed in observational studies comparing treatment options, and again, use of measures targeted to the outcomes of interest and sensitive to differences between groups is recommended. For example, the Cancer of the Prostate Strategic Urologic Research Endeavor (CaPSURE) provides a national, longitudinal database of prostate cancer patients, including measures of their HRQOL. The measures of HRQOL included in CaPSURE are a generic measure (the SF-36) and a prostate-cancer-specific measure (the University of California at Los Angeles Prostate Cancer Index). These HRQOL measures are collected at baseline and every 6 months thereafter. Thus, the CaPSURE database provides measures that allow comparison with the general population as well as a prostate-specific measure that assesses urinary function and bother, bowel function and bother, and sexual function and bother.

Using the CaPSURE data, a 2006 study compared the impact of salvage radiotherapy with radical prostatectomy and primary radiotherapy on HRQOL. In this study, men who underwent salvage radiotherapy had significantly worse changes in their sexual and bowel function compared with men who underwent radical prostatectomy alone; however, compared with primary radiotherapy, men who underwent salvage radiotherapy had significantly less worsening in their sexual function and bother. There were no differences between groups on the SF-36 physical or mental component summary score changes. As with the randomized controlled trial example above, this study demonstrates how using measures specifically targeted to the outcomes of interest in a given study can provide important information on trade-offs between different treatment options.

Decision-Analytic Models and Cost-Utility Studies

The above examples have demonstrated how profile HRQOL measures can be used in population comparisons through large-scale surveys and in treatment comparisons through randomized and observational studies. When the objective of a study is to inform decision making using decision-analytic modeling and cost-utility studies, preference-based measures are the recommended approach. An example is a cost-utility analysis of short- versus long-course radiotherapy for palliative purposes in non-small-cell lung carcinoma patients. To assess patients' preferences for the various health states, the authors used the EuroQol EQ-5D classification system, which assesses mobility, self-care, usual activities, pain/discomfort, and anxiety/depression. They did not find statistically significant differences between the short- and long-course groups on the average valuations of the modeled health states, but the long-course patients tended to survive longer. This resulted in a relative advantage for the long-course radiotherapy. This study used a preference-based HRQOL measure to obtain health-state valuations that could then be used in the calculation of quality-adjusted life years for a cost-utility analysis. While, in this case, the values of the expected health states did not differ between groups, this approach allows for explicit evaluation of trade-offs between quality and quantity of life.

Individual Patient–Clinician Interactions

Recently, there has been increasing interest in using HRQOL measures in routine clinical practice

for individual patient management. For these types of applications, clear and interpretable HRQOL measures are needed. In a 2004 randomized study evaluating the impact of using HRQOL assessment in clinical practice, intervention patients completed the EORTC QLQ-C30 and the Hospital Anxiety and Depression Scale with feedback of the data to their physician; attention-control patients completed the same questionnaires but the data were not provided to their physicians, and control patients did not complete any questionnaire. Some key findings from the study were that both the intervention and attention-control groups had better HRQOL compared with the control group (but not significantly different from each other) and that the intervention group was more likely to have had discussions with their clinicians about chronic nonspecific symptoms. However, this trial did not detect any differences in how patients were managed.

This study identified more positive effects of HRQOL assessment in clinical practice than many of the other similar early studies, in that it showed differences in HRQOL outcomes based on incorporation of standardized HRQOL assessment in individual patient management. In general, investigations of using HRQOL for individual patient management have shown improvements in communication but not changes in management or outcomes. It is likely that the field has yet to identify the best measures for use in individual patient management. Additional research will build on these early studies and inform the selection of measures that meet the criteria of being clear, interpretable, and actionable.

Future Outlook

It is tempting to think that selecting a small number of HRQOL measures and requiring their use across studies and applications would provide sure benefits to the quality and interpretability of cancer HRQOL, and certainly improved comparability and interpretability are worthy goals. However, it is important that the measure selected for a given study or application match its particular objectives.

Indeed, as shown in the examples above, there are a variety of ways in which an HRQOL measure can be used, and the best measure for a study or application depends heavily on what that study or application aims to accomplish. The examples above include the following measurement strategies: (a) a generic measure; (b) a modular approach with a general cancer core measure supplemented with a disease-specific module, an additional subscale, and two items developed for the study; (c) a generic measure used in combination with a cancer site-specific measure; (d) a generic preference-based measure; and (e) a general cancer measure from a modular approach that was not used with a disease-specific module but was used with a measure of anxiety and depression. In each of these cases, the measurement strategy selected was well matched to the matter being investigated. In the future, advances in measurement science (e.g., item response theory, item banking, and computer-adaptive assessment) may be able to bridge the sometimes competing goals of having a measure tailored to the study or application's objectives, while promoting comparability across studies and applications.

Claire F. Snyder, Carolyn C. Gotay, and Joseph Lipscomb

See also Cost-Utility Analysis; EuroQoL (EQ-5D); Health Outcomes Assessment; Health Status Measurement, Generic Versus Condition-Specific Measures; Health Utilities Index Mark 2 and 3 (HUI2, HUI3); Quality of Well-Being Scale; SF-36 and SF-12 Health Surveys; Sickness Impact Profile

Further Readings

Clauser, S. B., Ganz, P. A., Lipscomb, J., & Reeve, B. B. (2007). Patient-reported outcomes assessment in cancer trials: Evaluating and enhancing the payoff to decision making. *Journal of Clinical Oncology, 25,* 5049–5050.

Erickson, P. (2005). Assessing health status and quality of life in cancer patients: The use of generic instruments. In J. Lipscomb, C. C. Gotay, & C. Snyder (Eds.), *Outcomes assessment in cancer: Measures, methods, and applications* (pp. 31–68). Cambridge, UK: Cambridge University Press.

European Organisation for Research and Treatment of Cancer group for research into quality of life: http://groups.eortc.be/qol

Feeny, D. H. (2005). The roles of preference-based measures in support of cancer research and policy. In J. Lipscomb, C. C. Gotay, & C. Snyder (Eds.), *Outcomes assessment in cancer: Measures, methods,*

and applications (pp. 69–92). Cambridge, UK: Cambridge University Press.
Food and Drug Administration. (2007). *Guidance for industry: Patient-reported outcome measures: Use in medical product development to support labeling claims (DRAFT)*. Retrieved March 27, 2008, from http://www.fda.gov/cder/guidance/5460dft.htm
Functional Assessment of Chronic Illness Therapy: http://www.facit.org
Gotay, C. C., Lipscomb, J., & Snyder C. (2005). Reflections on the COMWG's findings and moving to the next phase. In J. Lipscomb, C. C. Gotay, & C. Snyder (Eds.), *Outcomes assessment in cancer: Measures, methods, and applications* (pp. 568–589). Cambridge, UK: Cambridge University Press.
Lipscomb, J., Gotay, C. C., & Snyder, C. (Eds.). (2005). *Outcomes assessment in cancer: Measures, methods, and applications*. Cambridge, UK: Cambridge University Press.
Patient-Reported Outcomes Measurement Information System: http://www.nihpromis.org
Snyder, C. F., Watson, M. E., Jackson, J. D., Cella, D., & Halyard, M. Y. (2007). Patient-reported outcome instrument selection: Designing a measurement strategy. *Value in Health, 10*(Suppl. 2), S76–S85.

Ordinary Least Squares Regression

The treatment of errors has a long tradition with attempts to combine repeated measurements in astronomy and geodesy in the early 18th century. In 1805, Adrien-Marie Legendre introduced the method of least squares as a tool for using models with specification errors to fit data collected to determine the shape and circumference of the earth. Specifying the earth's shape to be a sphere, he had to estimate three parameters using five observations from the 1795 survey of the French meridian arc. With three unknowns and five equations, any estimate of the unknown parameters led to errors, when fitted to the five observations. He then proposed to choose those estimates that make "the sum of squares of the errors a minimum" (Legendre, 1805, pp. 72–73).

A formal statistical theory of errors was developed by Gauss in 1809 and Laplace in 1810. The method of least squares was shown to possess many desirable statistical properties. For more than 200 years, a method invented to deal with experimental errors in the physical sciences has become universal and is used, with practically little or no modification, in the biological and social sciences.

A scientific method in the biological sciences often involves statement of a causal relationship between observable variables and a statistical model to estimate the relation and test some hypotheses.

Three common medical decision problems involving statistical methods are screening, diagnosis, and treatment. Data used in statistical analysis include medical history, clinical symptoms, and laboratory tests. For many medical conditions, there are no perfect tests such as an X-ray to detect the fracture of a bone. Decisions have to be made using one or more associated, observable factors.

Two problems arise with this approach: (1) How does one formulate a decision rule using the associated factors? and (2) Since no decision rule will be perfect, how is one to compare the decision rules, that is, the errors associated with these rules?

The ordinary least squares regression (OLS) method provides a solution. Suppose the medical condition is type 2 diabetes, and the gold standard is the oral glucose tolerance test. For a screening rule, we want to use readily available data for risk factors such as age, gender, body mass index, race, and so on, to predict the blood glucose and identify individuals with high risk for follow-up tests. Any function of the risk factors will provide an estimate of the blood sugar and hence be useful in diagnosing diabetes. Errors associated with the estimates are calculated using the observed blood sugar. The OLS method can be used to select a set of weights to combine the risk factors and estimate the blood sugar as follows: For every set of weights, there will be corresponding predicted values of blood sugar. Prediction errors can be calculated using the observed blood sugars. One can then calculate the sum of squares of the errors and choose the set of weights with the least sum.

Why square the errors and sum? Why not simply sum the errors? A simple sum of errors will be 0 if the positive errors add up exactly to the sum of the negative errors and hence will be misleading. On the other hand, the sum of squares of errors will be 0 if and only if all the errors are 0, that is, only if there are no errors.

This is true for other simple functions of errors also. For example, sum the magnitudes of the errors ignoring the signs, that is, the absolute values. This is the least absolute deviation (LAD) introduced by Galileo Galilei in 1632. The essential difference between OLS and LAD is in the treatment of large errors: OLS gives greater weights to large errors than LAD when the errors are greater than 1 and conversely.

What is the method of estimating the unknown parameters of the model? There is a precise mathematical formula to calculate the estimates. An intuitive explanation of the method is the following. Pick an arbitrary set of values for the parameters and use those values to compute the fasting plasma glucose (FPG) for each person. The deviations from the observed values of FPG are the estimated errors. Take another set of parameter values and do the same. Using a computer, you can do this for a million different sets easily. Find the set with the least error sum of squares. That will be quite close to the OLS estimate of the parameters obtained by using a mathematical formula.

Example

The following example illustrates the method. According to the American Diabetes Association, a patient is diagnosed to have type 2 diabetes if the patient's FPG is 126 mg/dl or greater on two occasions. This criterion requires fasting for at least 8 hours. An alternative biomarker of the disease is glycated hemoglobin known as HbA1c, or simply A1c. This is an average of the blood sugar levels over 90 to 120 days. Since this does not require fasting, this value could be obtained during any office visit. This will be a good screening tool if there is a good correlation between A1c and FPG. This correlation may vary biological characteristics such as age, gender, body mass index, and race. Let the statistical relation between FPG, A1c, and other variables be given by

$$\begin{aligned}\text{FPG} = \beta_0 &+ \beta_1\ \text{A1c} + \beta_2\ \text{DM} + \beta_3\ \text{UDM} \\ &+ \beta_4\ \text{gender} + \beta_5\ \text{age.cat} + \beta_6\ \text{race} \\ &+ \beta_7\ \text{bmi.cat} + \varepsilon,\end{aligned}$$

where

DM = 1 if the patient has physician-diagnosed diabetes and 0 otherwise (referent),

UDM = 1 if the patient has undiagnosed diabetes and 0 otherwise (referent),

gender = 1 for male and 0 for female (referent),

age.cat = 20+ (referent), 30, 40, 50, 60, 70, 85,

race = white (referent), black, Mexican American, Hispanic, others,

bmi.cat = underweight, normal weight (referent), overweight, and obese,

ε = error (normal, mean = 0, common variance = σ^2),

and the βs and σ^2 are parameters to be estimated by OLS.

The parameters are estimated using the National Health and Nutrition Examination Survey data, a nationally representative probability sample of the noninstitutionalized U.S. civilian population. We use 8,350 observations for the years 1999 to 2005. The results of OLS are given in Table 1.

Some natural questions concerning the estimated model are as follows:

1. Is there a good relation between FPG, A1c, and other covariates?
2. Does the model explain the variations of FPG adequately?
3. Is there evidence from the estimated residuals that assumptions such as constant variance (homoscedasticity) and normality are satisfied?
4. Are there outliers?

Note that the R^2 for the unweighted model is .77. This means that the model explains 77% of the variance of FPG, which is very good, given that we have a large sample representing different ethnic and age-groups. This is one general way to assess the performance of a linear regression model. We will suggest another performance measure relating to the purpose of the model, namely, screening for diabetes using A1c after examining the estimated residuals.

A plot of the residuals (estimated errors) is shown in Figure 1. Most residuals are concentrated around the estimated FPG of 100 since there are many nondiabetic persons. Due to clustering of these residuals, they appear as closed circles.

Table 1 Results of ordinary least squares regression

```
Coefficients:
                                  Value      Std. Error      t value    Pr(>|t|)
             (Intercept)        -18.3719       1.3921       -13.1973     0.0000
                     A1c         21.2307       0.2616        81.1622     0.0000
                      DM         41.2534       0.9875        41.7774     0.0000
                     UDM         -9.8355       1.3214        -7.4432     0.0000
                    male          2.9534       0.3734         7.9096     0.0000
    age.cat30+ thru  40           0.1164       0.6121         0.1902     0.8492
    age.cat40+ thru  50           1.2396       0.6282         1.9732     0.0485
    age.cat50+ thru  60           0.9280       0.6711         1.3827     0.1668
    age.cat60+ thru  70           0.2979       0.6568         0.4535     0.6502
    age.cat70+ thru  85          -0.5233       0.6456        -0.8106     0.4176
                   Black         -5.4793       0.5052       -10.8457     0.0000
              Mex.Amer.          -0.7945       0.4829        -1.6454     0.0999
             Other Hisp          -1.1862       0.9671        -1.2266     0.2200
                  Other          -3.6443       1.0112        -3.6040     0.0003
                under.Wt         -0.5819       1.4840        -0.3921     0.6950
                over.Wt.          1.3081       0.4640         2.8191     0.0048
                   Obese         -0.0591       0.4820        -0.1227     0.9023

Residual standard error: 16.88 on 8333 degrees of freedom
Multiple R-Squared: 0.7735
F-statistic: 1778 on 16 and 8333 degrees of freedom, the p-value is 0
18086 observations deleted due to missing values
```

It is clear from Figure 1 that the dispersion of the residuals is greater for observations with large FPG, suggesting that the variance of FPG of the diabetic group is larger than that of the nondiabetic group. The standard deviations of FPG calculated from the data for the nondiabetic and diabetic groups are 10 and 70. Hence, the assumption of homoscedasticity (common variance) is violated. In other words, there is heteroscedasticity.

One way to take this into account is to scale (weight) the observations from the diabetic and nondiabetic groups using the reciprocal of the standard deviations and estimate a weighted least squares regression. The results of such a regression are given in Table 2.

The estimated variance of the errors is 14.95. There are some small changes in the estimated coefficients.

We will now compare the screening performance of the two models. From Table 3, we see that among 8,350 persons, there are 240 with undiagnosed diabetes, that is, persons with an FPG of 126 or more who are not aware that they have diabetes. We use our regression model and predict their FPG using A1c, gender, race, and so on and giving a value 0 for the diabetes indicator variable, DM, and an estimated prevalence of UDM = .02. We then classify a subject to have diabetes if the estimated FPG is greater than or equal to 126 mg/dl. This can lead to two types of errors: false positives and false negatives. The performance can be judged by the true positives (TP) and true negatives (TN). From Tables 3 and 4, we see that they are 228 and 7,448 for the unweighted model and 7,494 and 240 for the weighted model. The TP for the weighted model is perfect and TN is greater. Thus, the weighted model performs better.

It should be noted that our model uses DM and UDM to explain variations in FPG. If this is to be used in practice for screening, estimates of these two variables must be obtained, for example, by estimating another two logistic regression models using DM and UDM as dependent variables and other variables as covariates.

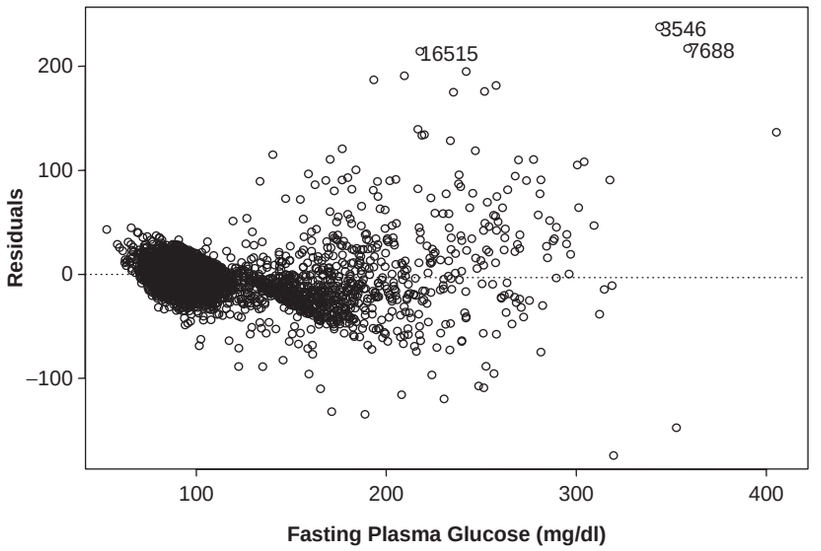


Figure 1 Plot of the residuals (estimated errors)

Table 2 Results of weighted least squares regression

Coefficients:

	Value	Std. Error	t value	Pr(>|t|)
(Intercept)	-16.8759	1.3989	-12.0637	0.0000
A1c	20.8917	0.2651	78.8000	0.0000
DM	52.0719	1.0626	49.0021	0.0000
UDM	-7.2858	1.4614	-4.9854	0.0000
male	1.7278	0.3909	4.4206	0.0000
age.cat30+ thru 40	-0.3726	0.5870	-0.6347	0.5256
age.cat40+ thru 50	-0.1875	0.6271	-0.2990	0.7649
age.cat50+ thru 60	-1.4407	0.6925	-2.0803	0.0375
age.cat60+ thru 70	-2.0585	0.6821	-3.0181	0.0026
age.cat70+ thru 85	-3.3402	0.6751	-4.9478	0.0000
Black	-5.2163	0.5132	-10.1641	0.0000
Mex.Amer.	-1.2379	0.5055	-2.4490	0.0143
Other.Hisp	-0.0805	1.0011	-0.0804	0.9359
Other	-4.1099	1.0676	-3.8497	0.0001
under.Wt	-1.0916	1.4225	-0.7674	0.4429
over.Wt.	0.8238	0.4713	1.7481	0.0805
Obese	-1.4669	0.4970	-2.9517	0.0032

Residual standard error: 14.95 on 8333 degrees of freedom
Multiple R-Squared: 0.8174
F-statistic: 2331 on 16 and 8333 degrees of freedom, the p-value is 0

Table 3 Screening performance: Sensitivity and specificity—unweighted model

N N/RowTotal N/ColTotal N/Total	no.dx.dm | fpg.hat.dm 0	1	RowTotl
0	7448 0.92 1 0.89	662 0.082 0.74 0.079	8110 0.97
1	12 0.05 0.0016 0.0014	228 0.95 0.26 0.027	240 0.029
ColTotl	7460 0.89	890 0.11	8350

Table 4 Screening performance: Sensitivity and specificity—weighted model

no.dx.dm | fpg.wtd.hat.dm 0	1	RowTotl
0	7494 0.92 1 0.9	616 0.076 0.72 0.074	8110 0.97
1	0 0 0 0	240 1 0.28 0.029	240 0.029
ColTotl	7494 0.9	856 0.1	8350

Advantages and Limitations

Some of the advantages of OLS are as follows:

1. It is easy to compute using exact mathematical formulae.
2. When used to predict future observations, it minimizes the mean squared errors.
3. It has the maximum correlation between the predicted and observed values of the outcome variable.
4. When the errors are distributed normally, OLS provides the most efficient estimators of the unknown parameters (shortest confidence intervals) in a linear regression model.

Some of the limitations of OLS are as follows:

1. All observations are given equal weight.
2. The expected value of the dependent variable is assumed to be a linear function of the covariates.
3. Statistical properties of the estimated coefficients are very sensitive to departures from the basic normality of the distribution of errors and also to the presence of outliers.

The heteroscedasticity in the data in our example illustrates the first limitation and the use of weights to solve the problem. All statistical software such as SAS, Stata, SPSS, and S-PLUS have functions to incorporate weights.

Regarding the other limitations, the assumptions of the Gaussian law of errors and the linear relation between the mean of the dependent variable and the independent variables, methods are now available to include more general distributions for the errors and various functional forms for the relation between the mean of the dependent variables and covariates.

The generalized linear models allow the distribution of errors to include the following five: Gaussian, binomial, Poisson, gamma, and inverse Gaussian. A generalized linear model is not restricted to linear relations. It is enough if some known function of the outcome variable is linearly related to the risk factors

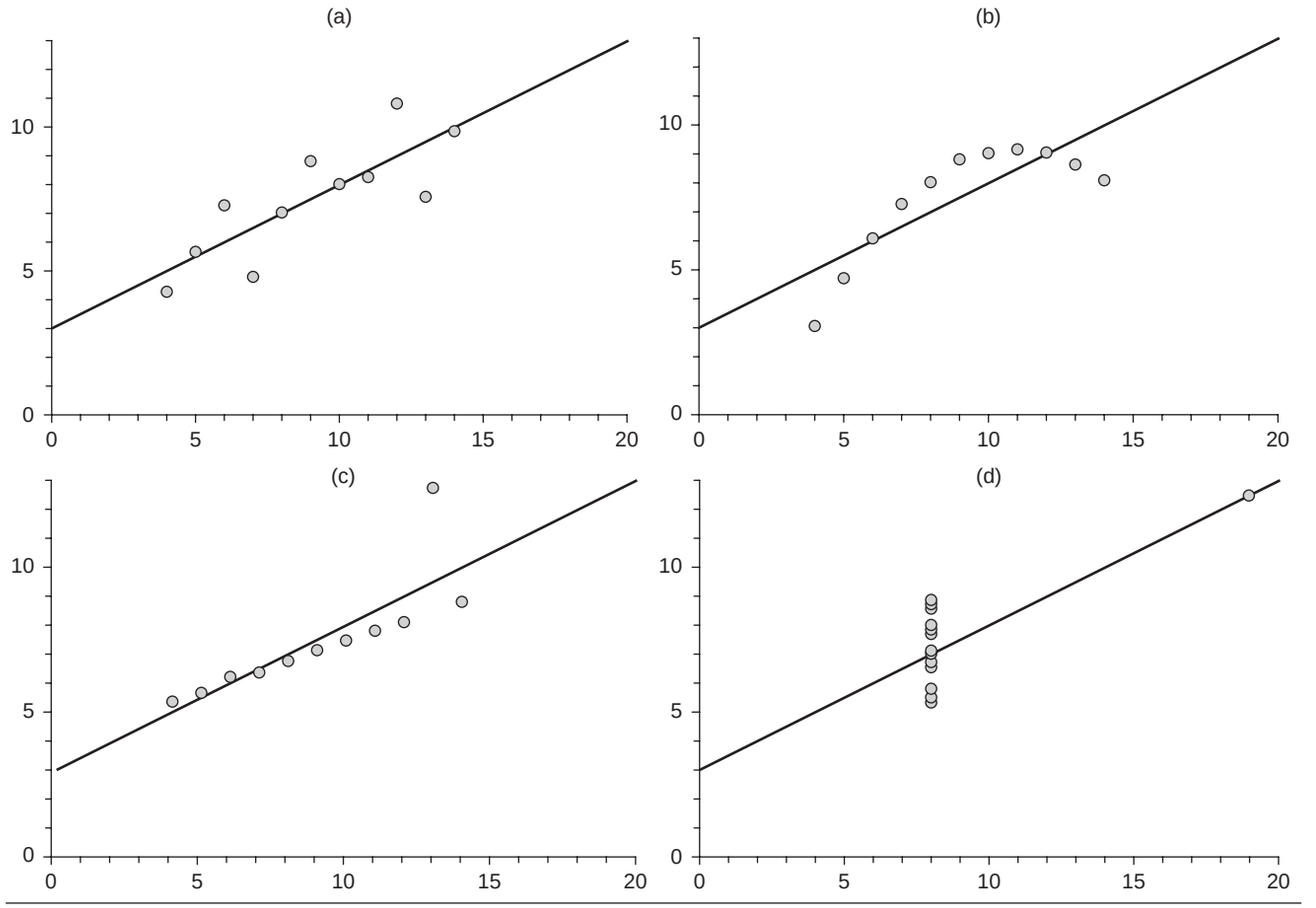


Figure 2 Graphical analysis of errors

Source: Anscombe (1973), pp. 17–21.

(independent variables). In another class of models, the generalized additive models, a linear relation between the transformations of both the dependent (outcome) variable and the independent variables is estimated. The least squares method is then applied locally to the observations. These are referred to as locally weighted error sum of squares (LOWESS). A linear or a polynomial function is fitted to the data points in the vicinity of each observation and then a smoothed value is used as a predictor of that observation. The advantage is that the nearby data points get greater weights.

With modern software and fast computers, it is now easy to visualize the data and the errors associated with various statistical models and estimation and prediction methods. Many diagnostic methods to detect departures from the assumptions of the least squares method are now available. It is useful to visualize the data and the estimated errors as the following figures from Anscombe show. Four quite different data sets are used to estimate a linear relation using the ordinary least squares method. The printed outputs of the regression coefficients, standard errors, R^2, and so on are exactly the same for all the four data sets. In other words, the error sum of squares is the same when the same line is fitted to all the data sets. Any other line will increase the error sum of squares. Yet a look at the figures shows that it is a good fit for the data in Figure 2a but a poor fit for the remaining three data sets shown in Figure 2b, c, and d.

V. K. Chetty

See also Analysis of Covariance (ANCOVA); Analysis of Variance (ANOVA); Weighted Least Squares

Further Readings

Anscombe, F. J. (1973). Graphs in statistical analysis. *The American Statistician, 27*(1), 17–21.

Gauss, C. F. (1809). *Theoria motus corporum coelestium.* Hamburg, Germany: Perthes et Besser.

Legendre, A. M. (1805). *Nouvelles methods pour la determination des orbites des comets.* Paris: Courcier.

Rao, C. R., & Toutenburg, H. (1995). *Linear models: Least squares and alternatives.* New York: Springer.

Outcomes Research

In the past 20 years, outcomes research has proliferated throughout the medical, academic, and health technology communities. But the definitions of outcomes research vary widely depending on who is conducting the research, or using the findings. The origins of outcomes research reach well into the 1960s and are rooted in the evaluation of the quality of medical care. The knowledge gained from outcomes research is critically needed to provide evidence about benefits, risks, and results of treatments to inform decision making by patients, providers, payers, and policy makers.

What Is Outcomes Research?

The term *outcomes research* is a vague, nontechnical term that is used to describe a wide-ranging spectrum of medical care research today. Broadly stated, outcomes research seeks to understand the end results of particular healthcare practices and interventions. Outcomes research has taken on so many different meanings to so many different constituencies or stakeholders in the business model of medicine that a clear definition is thought to be lacking. Depending on who is asked, a different answer will arise. A managed care organization CEO may respond with an economic example related to costs of new technologies that consume limited resources. A physician may be concerned with how well a certain procedure is performed for a certain patient with a given medical problem. A patient, now the consumer of healthcare services, may be equally concerned with how well a procedure or test is performed, but may also want to know if the right thing was done to begin with, or if the best thing was done that takes into account his or her own preferences.

Origins

Arguably, it was Avedis Donabedian who coined the term *outcome* as a component in his paradigm for quality assessment. His view of outcomes as a means to examine the quality of medical care is a foundation of outcomes research.

Traditionally, outcomes have been classified into three types: economic, clinical, and humanistic, dating to a classification by Kozma and others in the early 1990s. This economic, clinical, or humanistic outcomes (ECHO) model is a useful way to organize a framework around the concept of medical outcome. The model is built on the traditional medical model of an individual that develops disease or symptoms and seeks acute or preventive care. Healthcare professionals assess the needs of the patient, and clinical parameters that are modifiable by treatments can be monitored. Clinical outcomes can be medical events, such as heart attack, stroke, complications, or death, that occur as a result of disease or medical treatment. Recognizing that these concerns alone do not take into account quality-of-life measures such as functional status or patient preferences or the ever-increasing emphasis on costs of care, the model includes humanistic and economic outcomes, respectively. Therefore, depending on whose perspective is engaged, the term *outcome* can refer to various types of outcomes, but they are all aspects of the same construct: the result of disease or medical care treating the disease. As healthcare systems worldwide become more complex, reflecting the perspectives common to today's environment of the payer, provider, patient, policy maker, or regulating organization, research on these outcomes must necessarily involve a multidisciplinary approach. For example, determination of the value of a new pharmaceutical intervention will require data on all three types of outcomes to conduct cost-effectiveness, -benefit, -utility, or decision analysis.

In the early 1980s, John Wennberg and colleagues at Dartmouth began to discover differences in the rates of medical procedures and events that were not explainable by differences in disease status alone when examining populations in different locations. The phrase *geography is destiny* was coined to describe the odd finding that a variety of procedures such as hysterectomy or hernia repair were performed much more frequently in some areas than in others, even when there were no differences in the underlying rates of disease. And there was limited information about the end results of these procedures, the outcomes. The genesis of outcomes research began when stakeholders of all types pounced on these small area variations to try to explain them.

There are many advantages to studying outcomes directly. Many outcomes are relatively concrete and easily measured, such as mortality, and few would question the validity or importance of restoration of function, survival, and economics. The key reason to measure outcomes, to begin with, is to gain some tractable status for the outcomes of interest to establish a baseline. Once a baseline measure is made, quality assessment can occur to determine if the outcome of interest is occurring at a level that is desired. Questions such as "Is this rate high or low" are seen in literature from the 1990s onward. After assessments are made, monitoring the outcomes over time can then inform quality improvement efforts. Remembering that these outcomes of interest are the result of medical care for the treatments of disease, a necessary prerequisite to studying the outcomes is to define and examine the process or cause of these outcomes, the medical care.

Methods and Developments

The overriding paradigm that currently is the most accessible for the study of outcomes is the one proposed by Donabedian in the 1960s: structure, process, outcome. In this conceptual model, outcomes are the direct result of processes of care. The processes of care can be thought of as any activity that takes place between a care provider and a patient. For example, a nurse takes blood pressure, or temperature; a physician orders a diagnostic test or prescribes a medication. There are technical and qualitative aspects to process of care. Was the blood pressure cuff correctly placed and the systolic and diastolic readings carefully recorded? Was the patient comfortable and treated in a respectful manner while this procedure was performed? Processes of care are ubiquitous, and each and every one has the opportunity to affect outcomes. One cannot begin to even think about studying outcomes without first understanding and at least acknowledging the precedent of process that leads to them.

Outcomes research is not defined by a particular methodology. An array of study designs, including clinical trial approaches to test interventions, experimental or quasi-experimental studies, and nonrandomized or nonexperimental observational studies, are all available and used by the outcomes researcher. Such researchers draw from a wide-ranging multidisciplinary perspective from fields of clinical epidemiology, biostatistics, health services research, behavioral sciences, economics, and management sciences.

Outcomes research generates knowledge about safety and effectiveness of medical care. Disciplines such as pharmacoepidemiology and pharmacoeconomics employ outcomes research approaches to study effects of drugs in everyday use in healthcare systems. Beneficial effects, both expected and unexpected, as well as adverse events from drug-drug or drug-disease interactions are examples of clinical outcomes. Economic analyses, including cost-identification, cost-effectiveness, and cost-benefit can all be thought of within the framework of costs as outcomes. Cost-effectiveness studies of pharmaceuticals examine the incremental cost of one treatment over another with the incremental benefit in terms of effectiveness of a particular clinical outcome.

With the advent and explosion of the computer technology age, a wide variety of information sources now exist to support outcomes research. Health information sources such as administrative databases, clinical databases, disease registries, and trial databases are often linkable to other information such as census or survey studies for healthcare systems worldwide. In the United States, the federal government oversees the Centers for Medicare and Medicaid Services, which produce very extensive health information sources in the form of medical claims for both Medicare and state-administrated Medicaid. The U.S. federal government also administers a relatively complete electronic medical record for all beneficiary users of the Department of Veterans Affairs Health

Administration (VHA), which employs a national cadre of health services researchers engaged in a variety of outcomes research.

Uses

Decision analysis draws deeply on outcomes research. The modeling of a clinical problem to guide decision making requires inputs of prevalence of disease, effectiveness of interventions, incidence of adverse events, costs of care, a plethora of other outcomes measures, and patient preferences or utilities. Sensitivity analyses for decision models may examine risks for a range of outcomes or a range of effectiveness measures in subpopulations of interest at different risks for different outcomes.

Healthcare managers and purchasers can use findings from outcomes research studies to identify the best practices of potentially effective and cost-effective strategies that may be appropriate for their healthcare systems to take up and implement in order to improve the quality of care for their members or beneficiaries. The U.S. Agency for Healthcare Research and Quality (AHRQ) has long been a leader in conducting and funding outcomes research studies, as well as in the translation of findings from these studies into practice. AHRQ Patient Outcomes Research Teams (PORT) studies of the 1980s and 1990s helped identify variation in quality in a variety of acute and chronic medical conditions and led to Centers for Education and Research on Therapeutics (CERTs), Evidence-Based Practice Centers, and much more. Ideally, collaborations of academic researchers, third-party payers, and private-industry developers and innovators such as pharmaceutical and device manufacturers could make research, development, and implementation of new interventions more efficient to improve quality and reduce costs.

Future Directions

A variety of challenges and opportunities can be foreseen for outcomes research in the near future. An increasing emphasis on comparative effectiveness of interventions will result in the United States from the enactment of the Medicare Modernization Act of 2006, which mandated research into not only effectiveness of drugs in everyday use but also comparative effectiveness. Such information is crucially needed to inform patients and providers to best choose pharmaceutical treatments for the rapidly aging U.S. population. Findings from these studies can be expected to have worldwide influence. Dissemination of research findings for rapid adoption and implementation into practice will remain a developing and much needed science. Standards for outcomes research methodologies and training of new investigators is a growing need as evidence from clinical studies must be valid, reliable, and trustworthy if stakeholders in all arenas are to value and depend on these studies. Outcomes research theories and methodologies that originated in the 1960s will certainly continue to generate knowledge to improve quality of care in the 21st century.

Michael L. Johnson

See also Decision Analyses, Common Errors Made in Conducting; Economics, Health Economics; Health Outcomes Assessment; Mortality; Pharmacoeconomics; Risk Adjustment of Outcomes

Further Readings

Donabedian, A. (1966). Evaluating the quality of medical care. *Milbank Memorial Fund Quarterly, 44,* 166–206.

Donabedian, A. (1980). Explorations in quality assessment and monitoring. In A. Donabedian (Ed.), *The definition of quality and approaches to its assessment* (pp. 1–31). Ann Arbor, MI: Health Administration Press.

Epstein, R. S., & Sherwood, L. M. (1996). From outcomes research to disease management: A guide for the perplexed. *Annals of Internal Medicine, 124*(9), 832–837.

Freund, D., Lave, J., Clancy, C., Hawker, G., Hasselblad, V., Keller, R., et al. (1999). Patient outcomes research teams: Contribution to outcomes and effectiveness research. *Annual Review of Public Health, 20,* 337–359.

Stryer, D., Tunis, S., Hubbard, H., & Clancy, C. (2000). The outcomes of outcomes and effectiveness research: Impacts and lessons from the first decade. *Health Services Research, 35*(5, Pt. 1), 977–993.

Overinclusive Thinking

Defining the term *overinclusive thinking* depends on consideration of a *thinking episode*.

Thinking Episode

Questions of thinking may start out with the consideration of an episode of thinking (or an episode of thought), and one can ask the question regarding what is considered within that thinking episode, what is excluded from that thinking episode, what cannot be included in a thinking episode, and what cannot be excluded from a thinking episode.

When an individual is presented a choice between two options A and B and asked to choose between the options, it is key to recognize that some thinkers may well ask themselves the basic question of why one of the alternatives was selected as label A and presented first and the other was selected as B and presented second. While this may be termed by some as *overthinking* the problem, there may well be reasons why the principal investigator (researcher) called one A and presented that option first and called the other B and presented that option second. Even if one tells the study volunteer that the labels A and B were randomly selected as the labels of the alternatives, the study volunteer may still have a lingering question in his or her mind about the ordering and may assume that the researcher is telling him or her about the ordering being randomly selected to distract the study volunteer from the "real reason" the labeling and ordering were selected as they were.

The above considerations bring up the notion of thinking and what is to be included within an episode of thinking and what is to be excluded from such an episode. Here, the capacity to include elements within thought or to exclude elements from thought may be difficult to do purely mentally and may need to rely on tools such as decision trees to structure decisions and keep track of information that goes into a decision and the information that is excluded, or pruned away, from a decision. Even decision scientists can be legitimately criticized for what information is included in a decision and what information is excluded from that same decision.

Positive Choice Versus Rejection

When the thoughts under consideration are thoughts about choices among a set of items or options, the very notion of choice comes into question. What is a positive choice (choosing Option A over Option B) versus what is a negative choice (rejecting Option B and thus choosing Option A by default)?

Positive choice of one alternative over another or rejection of one alternative in preference of another can both be influenced by underinclusive and overinclusive representation of information in decision making affecting consumer choice in economics and patient choice and preference in medical decision making even when the structure of that decision is being overseen by a decision scientist.

Underinclusive Versus Overinclusive Thinking

If someone were a betting man or woman, he or she might bet that most individuals in most economic and medical decision-making situations are more underinclusive in their thinking (not including enough or including fewer pieces of information in their thought processes than other reasonable people would include) rather than being overinclusive (including too much information in their thought processes or including more information than other reasonable people would include). This is why physicians have particular roles in helping patients understand what is going on in their care, and why professional decision scientists have a role in and are recompensed for their work in economics and medical decision making. But physicians and decision scientists can misrepresent choice to a patient in decision making by underinclusive and overinclusive representation of information in decision trees causing underinclusive and overinclusive thinking on the parts of patients.

Underinclusive Thinking

While one may think of decision scientists as expert decision makers, one can also call to mind the types of decisions made by physicians and decision scientists that exhibit underinclusive thinking (removing information from a decision that other reasonable persons would keep in the decision).

Decision analysts or other experts in decision making could be chided by other experts, for example, ethicists, for underinclusive thinking regarding the risk information that a reasonable patient needs and should be given by his or her surgeon regarding a decision to accept or reject a surgical operation.

Underrepresenting key information is one dynamic cause of underinclusive thinking.

Underinclusive Thinking About a Surgical Intervention

Let us consider an example of underinclusive thinking in the care of a patient. Here we will contrast the thinking of a decision scientist with that of an ethicist on what information a patient needs to know to make a decision about whether to accept a physician-recommended surgical intervention, that is, a surgical operation in the patient's care.

The ethicist would recognize that there is an inherent conflict of interest in any decision recommended by a surgeon that a surgery should or should not be recommended in the patient's care. Here, the ethicist alerts us to the recognition that the physician is a surgeon who is recommending a surgical intervention; and what the patient may need to have is a physician, perhaps the patient's internist, to also consider the surgical recommendation as an option in the patient's care at this time to get a fuller range of points of view of the recommended intervention.

After the ethicist has completed his or her initial analysis of conflict of interest, a decision scientist examines the outcomes (benefits and risks) and the chance (probabilities) of their occurrence in the decision that the patient faces. The decision scientist bases his or her structuring of the decision the patient faces on review of the peer-reviewed medical literature and discussion with experts going beyond the surgeon and physician caring for the patient. The decision scientist then develops a framework such as a decision tree to structure the decision.

But as the tree gets larger with more decision branches, the decision scientist's whiteboard can no longer hold the informational content of the decision, and the decision scientist foreshortens the outcomes that need to be considered by the patient (prunes the decision tree) by excluding all rare adverse events where rare is defined in terms of events that occur at or below a rate of 1 in 10,000 surgical operations of the same or similar type as the operation under consideration. Here, the decision analyst argues that his or her exclusion of rare events at a level of chance of occurrence that is less than 1 in 10,000 is being done to simplify the construction of the decision tree, and the decision scientist argues that this exclusion is reasonable from his or her decision scientific perspective.

The ethicist comes by to visit the decision scientist and sees the decision tree on the analyst's whiteboard and asks the following question: "Where are the severe adverse outcomes at low probability that are associated with the surgery under consideration?" The ethicist continues,

> Shouldn't all the severe adverse outcomes such as death and cognitive, motor, memory, and sensory disability be explicitly included in the tree because these events may in fact occur if the patient undergoes the surgical operation; and if these events are systematically excluded from the decision tree, the patient may not even understand that the medical intervention in question is serious enough that he or she may die from the intervention or that he or she may sustain an irreversible severe adverse event—like a stroke resulting in major motor paralysis or a severing of a nerve supplying an organ causing a loss of organ function—that the patient will have to live with the rest of his or her life?

Here, the decision analyst is in effect being chided by the ethicist for underrepresenting information to the patient, information that a reasonable patient needs to have to make a decision about whether he or she wants to undergo the surgical intervention in question. In particular, the patient needs to know the chances of dying during the intervention or within 30 to 60 days after it and the chances of sustaining an irreversible injury that will remain with the patient for the rest of his or her life.

The above example of underinclusive representation by the decision scientist yielding underinclusive thinking by the patient on the fact of severe adverse outcome occurrence at low chance of occurrence can be exemplified in the real world by consideration of consent and informed consent cases in courts throughout the globe: If the patient had been given the information that death and severe disability were part of the risks of the surgical intervention and that information was given to the patient prior to the intervention, then the patient would have better understood the seriousness of the physician-recommended

surgical intervention and could not claim after the fact of injury occurrence (where the nondisclosed severe adverse outcome actually materialized in the patient's case) that he or she did not understand that the surgical operation to which he or she consented was as dangerous as it turned out to be.

Underinclusive and Overinclusive Thinking in Prescription Medicine Discussions

Patient-physician discussions about prescription medicines provide an opportunity to examine both underinclusive thinking (and its companion concept, underrepresenting risk information to patients in structuring the choice the patient has to consider) and overinclusive thinking (and its companion concept, overrepresenting risk information to patients in structuring the choice the patient has to consider).

Here, we take the case that the prescription medicine has been on the market for 20 years, and hence its risks are well understood across a broad range of patients with different medical conditions. In attempting to secure information about the risks of a prescription medicine that has been on the market for 20 years, one can readily assess the risk information from drugs or drug compendia—like the *Physicians' Desk Reference* (*PDR*) or textbooks—or from searching and review of the peer-reviewed medical literature on the topic. Often, what one finds in a review of the risks of a drug therapy is risk information that has been pared down considerably to a listing that shows only "significant adverse reactions" under a set of ranges of chance of the adverse outcome occurring in a population of patients. The following is a review of the nature of tables whose authors and developers have used underinclusive principles (Table 1) and overinclusive principles (Table 2) to represent risk of a drug, Prescription Medicine A.

Table 1 Underinclusive representation of risks of Prescription Medicine A

Significant adverse outcomes (adverse reactions) occurring
>10%
• Neuromuscular discomfort
• Skeletal discomfort

Table 1 is underinclusive in that it intentionally leaves out consideration of severe adverse outcomes (adverse reactions) at both the 1% to 10% level and the less than 1% level. In comparison, Table 2 provides more information but is also at risk (as judged by the patient, the physician, and the decision scientists) as it in a way still underrepresents certain types of information and, in other ways, overrepresents other types.

Table 2 Overinclusive representation of risks of Prescription Medicine A

Significant adverse outcomes (adverse reactions) occurring
>10%
• Neuromuscular discomfort
• Skeletal discomfort
1% to 10%
• Headache, dizziness, rash, abdominal discomfort, constipation, diarrhea, dyspepsia, flatulence, nausea, myalgia, weakness, muscle cramps, blurred vision
<1%
• Abdominal discomfort; depression; dermatomyositis; dizziness; fatigue; headache; hypotension (low blood pressure); insomnia; lichen planus; muscle pain, soreness, tenderness, weakness, myopathy; photosensitivity; pruritus; thrombocytopenia; vertigo; weakness

Both tables also lack the following elements of definition and conceptual development of terms:

- Lack of information of relationship of elements of the list to
  - Whether there is an impact on patient survival over time
  - Whether an injury is reversible or irreversible
  - What impact the element will have on the patient's quality of life
- Lack of specification of the origin of the information
- Lack of definition and meaning of symptoms (such as myalgia), laboratory test abnormality (such as thrombocytopenia), disease process (such as dermatomyositis), and physical sign and disease (such as lichen planus)
- Lack of clarity as to whether all risks are included in the greater than 10% category (for example, in the category of "nerve and muscle discomfort" and "bone discomfort," are there also issues related to cartilage, ligaments, and tendons?)
- Lack of development of criteria to be used to distinguish "significant" risk from "insignificant" risk
  - Are there risks that are still reported that are not included in the "significant" and "insignificant"risk categories?
- Lack of development of what the numbers represent
  - Greater than 10% of what? 1% to 10% of what? Less than 1% of what?
  - What is the numerator and what is the denominator of the fractional range of greater than 10%? Of 1% to 10%? Of less than 1%?
- What does the appearance of one term in both the 1% to 10% category and the less than 1% category mean; for example, why does "myopathy" appear as both a 1% to 10% event and a less than 1% event?

In addition, even experts would have difficulty answering a patient's questions of the following types:

1. Why do "abdominal discomfort," "depression," "dermatomyositis," and "insomnia," among others, occur with the drug (even if a physician would attempt such a discussion of how these events could be associated with the drug therapy in question)?
2. What does "dermatomyositis" feel like, and how will it affect the life of a patient should it occur?
3. What will be the level of severity of the adverse reaction if it should occur?
4. What is the mechanism of action by which the adverse reaction is causally associated with the drug? How does the drug "cause" the particular adverse outcome in a patient?

Determining the best ways to effectively describe the risks and the benefits and their chance of accruing in a particular patient's care and the best ways to represent these outcomes and their chance of occurrence while attempting to minimize underinclusive thinking and to minimize overinclusive thinking is an active area of research in all areas of the world at the present time.

Dennis J. Mazur

See also Cognitive Psychology and Processes; Decision Psychology; Deliberation and Choice Processes; Evaluating Consequences; Informed Decision Making; Unreliability of Memory

Further Readings

Colombo, L., Nicotra, E., & Marino, B. (2002). Preference reversal in decision making: The attraction effect in choice and rejection. *Swiss Journal of Psychology, 61,* 21–33.

Redelmeier, D. A., Shafir, E., & Aujla, P. S. (2001). The beguiling pursuit of more information. *Medical Decision Making*, *21,* 376–381.

Schwartz, J. A., & Chapman, G. B. (1999). Are more options always better? The attraction effect in physicians' decisions about medications. *Medical Decision Making*, *19,* 315–323.

Shafir, E. (1993). Choosing versus rejecting: Why some options are both better and worse than others. *Memory and Cognition, 21,* 546–556.

Shafir, E., Simonson, I., & Tversky, A. (1993). Reason-based choice. *Cognition*, *49,* 11–36.

P

Pain

Pain is a universal experience. How pain is experienced and the clinical management that may be employed in its diagnosis and treatment are complex and multidimensional. When a patient presents with pain, the decision making of the clinician may be simple or complex, intuitive or analytical and is susceptible to multiple errors in assessment, investigation, and treatment. In this, pain is no different from any other area of medicine. Nevertheless, there are unique aspects of pain and its management that raise challenges to the quality of clinical decision making. Multiple medical, sociocultural, and religious values exist in all aspects of pain and its management. The heuristics of pain management are complex, and the possible cognitive dispositions to respond are ever present. Good clinical management requires a solid foundation in the science and practice of pain medicine, careful attention to detail, meticulous communication, and a vigilant awareness of potential biases.

The Multidimensional Nature of Pain

The International Association for the Study of Pain defines pain as "an unpleasant sensory and emotional experience associated with actual or potential tissue damage, or described in terms of such damage." Pain is inherently subjective. Pain is experienced as a result of a highly complex interaction of physical, biochemical, physiological, cognitive, emotional, behavioral, and sociocultural factors. The brain integrates information from multiple sources to form the experience that is pain. That multiplicity of sources, both aggravating and ameliorating the final experience, makes clinical decision making in relation to pain challenging. The recognition of this complexity has led to the development of multidisciplinary pain teams and the broadening of pain management to include a range of nonpharmacological interventions.

The major aspects of pain are physical, psychological, and environmental. Thus, the experience of pain is a combination of local biochemical changes, sensory information from somatic and visual receptors, visual and other sensory information, intrinsic neural inhibitory inputs, phasic cognitive and emotional inputs (e.g., anxiety) and tonic cognitive and emotional inputs (memory and cultural experiences), and inputs from the body's stress regulation system.

The sociocultural dimension of pain includes demographic characteristics; ethnic background; and cultural, religious, and social factors that influence an individual patient's perceptions of and response to pain. The search for meaning in pain is universal.

The adequacy of the treatment of pain varies enormously around the world. The reasons for inadequacy of treatment include lack of or suboptimal training of clinicians in pain assessment and management, inadequate attention to pain as a symptom, the presence of very restrictive domestic opioid laws, infrastructure weaknesses preventing patients' access to analgesia, opiophobia

of clinicians, and medical neglect. In response to this universal challenge, there has been a growing recognition that pain management is a basic human right that places clear obligations, through the international right to healthcare, on national governments. Other legal sources of this right emerge from statutory law, elder abuse law, and the law of medical negligence.

Heuristics and Pain

The spectrum of decision making in pain medicine lies along a continuum from simple to complex and is largely related to the level of uncertainty. That uncertainty may arise in all aspects of management, from the presenting history through to its treatment. In response to that uncertainty, clinicians employ a variety of conscious and unconscious tools from careful reasoning to intuition. One response to the irreducible element of uncertainty in clinical decision making is the use of heuristics. Heuristics are simple rules of thumb or judgments borne out of years of individual and collective experience. Generally, such heuristics are effective, but occasionally, they fail. Good decision making in pain management may be impeded by various cognitive biases or cognitive dispositions to respond.

Heuristics and History Taking in Pain

When a patient presents with pain, it is extremely important that a careful and meticulous history be taken. A critical threshold is the recognition that pain is a subjective symptom. This foundation in subjectivity means that two aspects of history taking by the clinician are vital: (1) a clarification of all aspects of the pain experienced by the patient and (2) meticulous listening to that description by the clinician. If either is missing or inadequate, pain management will be flawed. The time of onset of pain, its severity, its character, and the aggravating and ameliorating factors are all significant aspects of the history. For example, the site, character, and severity of acute pain usually allow a clinician to employ heuristics to rapidly reach a differential diagnosis of the cause of that pain. Aspects of the history of chest pain are, for instance, more suggestive of ischemic heart disease, pulmonary emboli, pericarditis, or a dissecting aortic aneurysm. Clearly, investigations will also contribute to decision making. Equally in the presentation of chronic pain syndromes, there are important aspects of the history of, for instance, carpal tunnel syndrome or fibromyalgia. Similarly, in malignant pain, a sudden onset of very severe bone pain suggests a pathological fracture, or a pain radiating directly from upper abdomen to upper lumbar spine may represent infiltration of a known pancreatic malignancy to the celiac plexus. Instruments containing elements of a pain history have been developed to increase the precision of diagnosis of pain associated with nerve damage (e.g., PainDetect and S-LANSS).

Heuristics, however, are not infallible. In the context of pain and history taking, multiple issues may apply: a patient speaking without an interpreter, a stoical patient, a patient fearful of hospitalization who minimizes the extent of his or her pain, a patient who feels that his or her pain is an inevitable part of age or illness, a patient fearful of opioids, an infant or child in pain, or a patient with an intellectual disability or cognitive impairment. Equally, there may be issues for the clinician: A doctor who is rushed, exhausted, impatient with the pace of reply, or overly suspicious of the patient's veracity may experience an impact on this critical initial step in pain management.

Examples of failed heuristics in the taking of a pain history include the following:

1. *Representativeness*—the heuristic that underlies pattern recognition: This is the assumption that something that seems similar to other things in a certain category is itself a member of that category. *Representativeness restraint* is an error that occurs when a clinician makes an incorrect judgment on the basis that the patient does not fit a representative class. Atypical presentations of pain are examples.

2. *Anchoring*—the tendency to fixate on specific features of a presentation too early in the therapeutic encounter, leading to a premature closure of thinking: Examples would include prematurely diagnosing nonmalignant osteoarthritic lumbar back pain in an elderly patient who is ultimately found to have metastatic bone disease or prematurely dismissing the pain history of a patient as

opioid-seeking behavior, where there is a genuine organic reason for rapidly increasing the opioid dose, in response to the pain. A premature character judgment may irreparably compromise both pain management and the trust between clinician and patient.

3. *Duration neglect*—the tendency to neglect a symptom or sign: Patients with pain may attempt to ignore it for some time. A patient's single report of pain onset, especially in the context of chronic pain, will often lead to an underestimation of chronicity.
4. *Posterior probability error*—the tendency to assume that history repeats itself: Even though a patient may present multiple times with pain due to a specific cause, it does not exclude the possibility that the patient's pain in the current presentation has another cause.

Heuristics and the Treatment of Pain

The heuristics of pain management are both advanced and primitive. They are advanced in the sense that the understanding of the actions, efficacy, and use of analgesia in all contexts from acute to chronic pain, from simple to strong analgesia, and in all routes of delivery from topical to intraspinal have improved significantly in the modern era. While the levels of evidence for all analgesia in all contexts vary considerably, overall the modern clinician has a strong armamentarium to respond to the pain of patients. Evidence-based guidelines for pain management of adults and children exist. The key theme of those guidelines is the importance of a calibrated response to the severity of the pain, that pain should never be neglected or inadequately treated, that analgesia is best given regularly rather than intermittently, and that the cause of the pain should be carefully considered throughout.

What remains primitive is the decision-making capacity of many doctors to manage moderate to severe pain. This lack of capacity arises partly from a lack of knowledge of the broad range of options that are now available. However, even partial knowledge may not result in the clinician responding, because of lack of familiarity with the treatment options and fear of making mistakes or even embarrassment in making a referral to an expert ("loss of face") who is capable of providing the help that is needed. An example is the critical bias that lies in attitudes to and knowledge about opioids. That bias represents a significant barrier to effective analgesia. Fears about the medical use of opioids (opiophobia), while lessening, remain ubiquitous and are founded on myths that retain currency to the present. These fears, in part, are based on opioignorance—inadequate education on the safety of opioid medication. These myths include the belief that all opioids in all contexts are addictive, that opioids are inevitably and perpetually sedating, that the commencement of opioids should be based on the extent of the disease and not the extent of the pain, and that opioids proportionately given will hasten death. Doctors are part of a wider society, and similar attitudes to opioids reside among the general population of patients.

Multiple biases may flow from these preconceptions and myths. Opiophobia and opioignorance may lead to a cascade of flawed decision making in treatment. Despite clear evidence-based guidelines on pain management, clinicians may be susceptible to *aggregate bias*, where clinicians rationalize treating an individual patient differently from guidelines on the basis of a mistaken belief that variables representing group averages reflect what is true for a particular patient rather than a group of patients. This bias may lead to pain management that is idiosyncratic and simply ignores guidelines. Other sources of flawed decision making in pain management include *omission bias* (temporizing or reluctance to treat), the *availability heuristic* (the tendency of clinicians to overestimate the risk of addiction when prescribing opioid analgesics for pain relief and to consequently undertreat pain), and *overconfidence* (clinicians rating their ability to manage pain highly even though they have serious shortcomings in attitudes and knowledge).

Another common bias in the use of opioids is the *illusory correlation*, the tendency to incorrectly perceive two events as causally related: opioid use and sedation or confusion. The clinician needs to be careful not to reflexly ascribe these symptoms to opioid analgesia when other pathological processes such as sepsis, electrolyte abnormalities, or cerebral metastases are in fact the true cause.

Heuristics may be both practically useful and especially problematic when the patient cannot give an adequate pain history. In a study of nursing

decision making in pediatric pain management after surgery by Shannon Scott-Findlay and Carole Estabrooks, three heuristics were commonly used—representativeness, availability, and anchoring. *Representativeness* was employed to justify an assumption that certain surgical procedures for children were associated with certain levels of pain. The error lay in allowing that assumption to dictate a clinical response over careful pain assessment. As the authors stated, the management of the children's pain was "more generalized and less individualized" (p. 217).

In adults, such representativeness has been used to characterize some patients as "having a low pain threshold" or "exhibiting excessive pain behavior." Such labels inevitably lead to a negative attitude, which may result in inadequate pain control. Scientific data contrast starkly with the foregoing myths. As Geoffrey Gourlay and colleagues showed, the minimum effective blood concentration of opioid analgesics varies by a factor of 6 to 10 among patients receiving the same operation by the same surgeon. In addition, there is a strong genetic basis for the difference in pain response among patients to a similar noxious stimulus.

The experience of pain may have a profound influence on the interaction between patient and clinician. It may influence all aspects of medical decision making—pain assessment; choice of investigations; pain management; and, from the perspective of the patient, compliance with treatment. How patients experience pain, how sensitive clinicians are to that experience, and how both clinician and patient respond to that pain are each critical components of medical decision making. Good clinical decision making in pain management requires good heuristics. However, biases may undermine objective reasoning. Those biases may impede all aspects of pain management, from assessment to treatment. Pain is complex and is susceptible to multiple interpretations in medicine, culture, and religion. Opiophobia and opioignorance are critical and explain, at least in part, the myriad biases that may influence pain management. Pain is ubiquitous. Excellence in decision making in relation to pain is not necessarily so.

Frank Brennan and Michael Cousins

See also Heuristics; Patient Rights

Further Readings

Baruch, J. M. (2008). Why must pain patients be found deserving of treatment? *Virtual Mentor, 10*(1), 5–12. Retrieved February 21, 2009, from http://www.virtualmentor.ama-assa.org/2008/01/pdf/ccas1-0801.pdf

Brennan, F. P., Carr, D. B., & Cousins, M. C. (2007). Pain management: A fundamental human right. *Anesthesia and Analgesia, 105,* 205–221.

Croskerry, P. (2005). The theory and practice of clinical decision-making. *Canadian Journal of Anesthesia, 52*(6), R1–R8.

Gourlay, G. K., Kowalski, S. R., Plummer, J. L., Cousins, M. J., & Armstrong, P. J. (1988). Fentanyl blood concentration: Analgesic response relationship in the treatment of postoperative pain. *Pain, 67,* 329–337.

International Association for the Study of Pain. (2009). *IASP pain terminology*. Retrieved January 7, 2009, from http://www.iasp-pain.org/AM/Template.cfm?Section=Home&template=/CM/HTMLDisplay.cfm&ContentID=6648#Pain

Larue, F., Colleau, S. M., Fontaine, A., & Brasseur, L. (1995). Oncologists and primary care physicians' attitudes toward pain control and morphine prescribing in France. *Cancer, 76,* 2375–2382.

Potter, M., Schafer, S., Gonzalez-Mendez, E., Gjeltema, K., Lopez, A., Wu, J., et al. (2001). Opioids for chronic nonmalignant pain: Attitudes and practices of primary care physicians in the UCSF/Stanford Collaboration Research Network. University of California, San Francisco. *Journal of Family Practice, 50,* 145–151.

Scott-Findlay, S., & Estabrooks, C. A. (2006). Knowledge translation and pain management. In A. G. Finley, P. McGrath, & C. T. Chambers (Eds.), *Bringing pain relief to children: Treatment approaches* (pp. 199–228). Totowa, NJ: Humana Press.

Parametric Survival Analysis

Parametric survival analysis is a subset of mathematical and statistical methods for characterizing the time relatedness of the occurrence of events such as death. It is distinguished from nonparametric and semiparametric methods by using a mathematical formula, termed a *model*, to summarize times to an event. At a minimum, the model contains the variable *time* and at least one

constant, called a *parameter*, whose value is estimated statistically from event-time data.

Alternative Philosophies in Formulating Parametric Survival Models

Biomathematically, parametric models characterize the rate at which time-related events occur, called the *force of mortality* or *hazard function*. Survival is derived by applying this rate to living subjects such that their number diminishes as deaths increase. Biostatistically, parametric models are empirical, convenient characterizations of the distribution of event times.

Although their utility is the same, parametric survival models emanating from these two perspectives are quite different. Biomathematical models tend to be formulated along mathematical lines of physical and chemical processes, whereas biostatistical distribution models arise from point-process, stochastic underpinnings. For example, survival after cancer diagnosis often decreases in exponential fashion. This suggests a biomathematical analogy to a unidirectional chemical reaction with a constant rate of transformation of substrate to product, which results in an exponential depiction of substrate. Such biochemical reaction rates are known to depend on variables such as temperature, so it is not difficult to imagine that a similar rate of death could be influenced by factors such as cancer stage. A biostatistician may appreciate this same exponential decrease in survival but would not think in terms of a mortality rate. Instead, he or she may logarithmically transform the survival function, find it to be a linear decline, and think about what factors may be associated with changes in the slope of log survival.

Alternatives to Parametric Methods

Parametric survival methods are distinguished from nonparametric ones, by which distribution of event times is estimated directly without an underlying model. A cumulative distribution curve turned upside down (called the *complement*) is a nonparametric representation of the distribution of event times. However, because data are often incomplete (think of the distribution of ages of people, some of whom refuse to reveal their age except that it is greater than 29 years), a method is needed that can estimate at least part of the cumulative distribution of event times. The product limit method of Kaplan and Meier is an example of a nonparametric method for generating a cumulative distribution, in part or as a whole, from incomplete data (called *censored data*). Parametric survival methods are also distinguished from semiparametric ones that do not explicitly model the underlying risk function (called the *hazard function*) but only the factors modulating it. The most commonly encountered semiparametric method is the one proposed by Cox, a time-related multivariable regression model. The parametric portion of semiparametric methods often relates the logarithm of the hazard function to a linear (additive) combination of risk factors (variables) whose values are weighted (multiplied) by statistically estimated constants (parameters). Such a logarithmic function of risk factors gives rise to what is known as *proportional hazards*.

Nature of Survival Analysis

Common Threads

All survival methods assume that time-related events occur at an instant in time. This assumption holds only approximately for many events. Often, what is called an event is a protracted process that may better be analyzed by what is called *longitudinal data analysis*.

A second common thread is that the event time is not yet known for some subjects. All that is known is that at follow-up time, the event has not yet occurred. For death, this means that some subjects remain alive at follow-up. This incomplete information about event time is called *censoring* (a term borrowed from census nomenclature).

Importance

The importance of survival analysis in evidence-based medicine is that it addresses *appropriateness* of therapeutic decisions by assessing long-term benefits and risks. For example, an interventional procedure (e.g., an appendectomy) that increases short-term risk may be appropriate because this early risk is far outweighed by long-term benefit.

Essential Data

Survival analyses require three pieces of data: (1) a clear definition of the time-related event,

(2) a "time zero" at which all individuals become exposed uniformly to the possibility of experiencing the event, and (3) a time thereafter when the individual either has or has not experienced the event or has ceased to be at risk. Parametric survival models are particularly vulnerable to lapses in these three essential data elements. For example, if one is interested in bioprosthetic heart valve deterioration, which follows an accelerating failure pattern, one can observe an apparently complex accelerating-then-decelerating pattern if some subjects are not at risk because they received nondeteriorating mechanical heart valves.

These essential data elements must be gathered by a formal, nonopportunistic follow-up mechanism. Again, parametric survival analysis is particularly sensitive to improper follow-up. Lapses tend to generate temporal patterns of risk that appear artificially to require complex (high order) mathematical models to characterize. For example, a clinical trial may require systematic yearly follow-up of all subjects but notification within 48 hours of any death. If in a survival analysis, one includes deaths beyond the point of last systematic follow-up, these deaths have no denominator. Therefore, the survival curve will appear anomalously to fall precipitously.

Some Parametric Survival Models

Compartmental Models

The simplest parametric survival model is the one-parameter constant-hazard model. A constant hazard implies that events occur randomly across time. This results in exponentially decreasing survival, much like radioactive decay, which has a constant decay rate.

It is useful to think of many parametric survival models in the same framework as radioactive decay or biochemical reactions, something that healthcare workers will find familiar. Indeed, the genesis of the constant-hazard model was the analysis of Bills of Mortality during the Black Plague by the merchant John Graunt in 1662. He assumed a constant birth rate, analogous to goods being delivered to a store, and a constant death rate, analogous to goods being bought. He thought of survival as being analogous to inventory on the shelves, whose quantity reflected the balance of these two rates. Graunt called the death rate the *hazard* rate, which was a technical term for a form of dicing that had crept into common usage and meant "calamity."

Especially during the 19th century, numerous mathematical models were developed for physical and biological phenomena. In the early 20th century, others were developed for industrial events (e.g., the Weibull model, a generalization of constant hazard). In more modern times, complex machines and devices, such as semiconductors, were found to follow a bathtub-shaped hazard, giving rise to the familiar terms *burn-in, random failure*, and *wearout*. This bathtub-shaped hazard holds for human survival: high infant death rate, low childhood and middle-age death rate, and accelerating old-age death rate.

Whether it be a population of semiconductors or people, the pattern of failure is usually simple and can be characterized by simple mathematical formulae with a small number of parameters (a low-order model). Therefore, identifying an appropriate mathematical model for a time-related event is not an onerous task, although the relative rarity of biomedical (as opposed to industrial) investigators using such models may lead one to think otherwise. These days, simple parametric survival models are available in standard statistical software packages, but these may not fit the data well. Thus, some investigators have compiled systems of simple models from which it is relatively easy to select statistically an appropriate model (e.g., see Blackstone, Naftel, & Turner, 1986).

Distribution Models

The survival curve is a complete or partial cumulative distribution function (CDF) of event times. A typical function used to characterize such data is the Gaussian distribution of the logarithm of event times (because their values are strictly positive). It has two parameters, commonly called the mean and the standard deviation.

Those familiar with distributions, however, may think only of the CDF and its first derivative, the probability density function (PDF). The idea of a force of mortality (hazard function) is foreign to this framework. It is helpful to recognize that the hazard function is the ratio of PDF to CDF, so it is a conditional PDF.

Modulating Survival by Risk Factors

A natural extension of parametric survival models is incorporating into one or more parameters a model (e.g., a log-linear model) of risk factors. These act to change the value of model parameters and thereby the contour of time-related survival. Some model parameters modulate survival considerably more per unit change in value than others (sensitivity), and these are often the parameters targeted to carry risk factor information.

Advantages of Parametric Survival Analysis

Why go to the trouble of characterizing either the hazard or the survival function in mathematical or distributional terms? The following are several reasons for doing so.

Portability

The most compelling reason for using parametric survival analysis relates to portability and ease of manipulating a mathematical equation to understand the underlying phenomenon, assess the impact of risk factors, and facilitate strategic medical decision making.

Understanding Phenomena

Particularly in examining the hazard function, one can discover easily what is difficult nonparametrically. For example, one can find that the overall hazard function for prosthetic valve infection peaks early after valve replacement, then falls to a low constant level; actually, however, the peak occurs earlier for some organisms, and the hazard function does not peak at all for others. Another example: Crossing the lines of survival for medical and surgical therapy for left main trunk disease reflects an initially elevated risk at the time of surgery that falls to a substantially lower risk than with medical therapy.

Assessing Risk Factors

One can assess the impact of a continuous risk factor such as age at intervention simply by varying age and solving for survival at a fixed time point, such as 10 years, keeping the value of all other risk factors constant. Similarly, one can solve the mathematical formula for presence or absence of some risk factor, again holding all other factors constant to isolate one risk factor at a time. These graphic displays are termed *nomograms*.

Decision Making

Mathematical formulae for alternative treatments can be solved for an individual patient's characteristics and the resulting survival curves compared. Where do the lines cross, if they do? What is the magnitude of short- versus long-term risk? How much lifetime can be saved by one versus another therapy (obtained by mathematical integration of the area between survival curves)? Use of parametric survival models for such decision making was particularly encouraged in the 1991 American College of Cardiology/American Heart Association guidelines for coronary artery surgery.

Prediction in Future Groups

Parametric survival models can readily predict survival in future groups of similar patients. Each patient can be given a personalized survival curve as a solution of the mathematical equation for the patient's characteristics. To determine if these predictions are valid, individual survival curves can be averaged and compared with observed survival after the new patients have been followed. In addition, cumulative hazard (negative logarithm of survival) at the time of each new patient's death or end of follow-up can be summed, and this number should correspond with the total number of observed deaths (a measure of *prediction error*).

Extrapolations

As long as sufficient data are available, parametric survival models can be extrapolated to end of life. Although these extrapolations are useful in estimating length of life, such estimates, as with any extrapolation, must be viewed with caution.

Adaptability to Nonproportional Hazards

Parametric survival models can be readily adapted to nonproportional hazards. For example, early risk after intervention for a serious disease is usually modulated by the patient's immediate preprocedure condition. After patients recover, subsequent survival usually depends on coexisting

chronic factors, such as age, weight, and comorbid conditions. Low-order mathematical models, properly formulated, can accommodate risk factors for these different time frames of risk.

Generation of Smooth Hazard Function

Parametric survival models generate an inherently smooth hazard function. Nonparametric estimation of hazard functions, which may be of more interest than survival, is inherently noisy, although smoothing techniques can filter out some of this noise.

Immunity to Completion Effect

Well-formulated parametric survival models are relatively immune to the "completion effect" inherent in nonparametric survival estimates, which results in underestimating survival. This occurs at the tail end of the survival curve, when few subjects remain.

Accurate Reflection of Shape

Parametric estimates of survival, particularly when the number of subjects is small, may reflect the shape of survival more accurately than less "smooth" nonparametric models, providing better prediction.

Limitations of Parametric Survival Analysis

The limitations of parametric survival analysis are typical of any mathematical model used to summarize data:

1. The data may not fit model assumptions.
2. Time must be spent in fitting the underlying hazard function or survival distribution, usually requiring nonlinear (iterative) estimation procedures.
3. In theory, an infinite number of mathematical models can fit a set of data. However, systems of flexible parametric models will yield generally similar hazard and survival functions and incorporate risk factors (often without assuming proportional hazards).
4. Although the underlying structure of hazard or survival is usually simple, its modulation by risk factors may be complex. Yet risk factors are usually incorporated into parametric (and semiparametric) models as an additive function. This limitation is common to all statistical regression models.

Eugene H. Blackstone

See also Cox Proportional Hazards Regression; Hazard Ratio; Nomograms; Survival Analysis

Further Readings

Blackstone, E. H., Naftel, D. C., & Turner, M. E., Jr. (1986). The decomposition of time-varying hazard into phases, each incorporating a separate stream of concomitant information. *Journal of the American Statistical Association, 81,* 615–624.

Cox, D. R. (1972). Regression models and life tables. *Journal of the Royal Statistical Society: Series B, 34,* 187–220.

Cox, D. R., & Oakes. D. (1984). *Analysis of survival data*. London: Chapman & Hall.

Diggle, P. J., Heagerty, P. J., Liang, K. Y., & Zeger, S. L. (2002). *Analysis of longitudinal data* (2nd ed.). New York: Oxford University Press.

Graunt, J. (1939). *Natural and political observations made upon the Bills of Mortality*. Baltimore: Johns Hopkins University Press. (Original work published 1662)

Harris, E. K., & Albert, A. (1991). *Survivorship analysis for clinical studies*. New York: Marcel Dekker.

Kalbfleisch, J. D., & Prentice, R. L. (2002). *The statistical analysis of failure time data* (2nd ed.). Hoboken, NJ: Wiley.

Kaplan, E. L., & Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association, 53,* 457–481.

Weibull, W. A. (1951). A statistical distribution function of wide applicability. *Journal of Applied Mechanics, 18,* 293–297.

Patient Decision Aids

The goal of evidence-based healthcare is to integrate clinical expertise with patients' values using the best available evidence. However, many decisions fall in the "gray zone," because the benefit-harm ratios either are unknown or depend on how patients value them. For example, should

patients take a more aggressive treatment when simpler therapies fail to control moderate or severe symptoms of menopause, osteoarthritis, benign prostatic hyperplasia, back pain, benign uterine bleeding, or stable angina? Other complex decisions include genetic testing, reproductive choices, treatment of early breast and prostate cancer, and the intensity and location of care at the beginning and end of life.

To prepare patients for discussion of these options, patient decision aids (PtDAs) have been developed as adjuncts to counseling. This entry defines PtDAs, describes their efficacy, and highlights the challenges in implementing them in clinical care.

What Are Patient Decision Aids?

According to the International Patient Decision Aids Standards Collaboration, PtDAs are tools to help patients participate in their healthcare decisions in the ways they prefer. They supplement rather than replace a practitioner's counseling. PtDAs provide facts about options and outcomes, communicate probabilities of benefits and harms, clarify which benefits and harms matter the most, and guide patients in the steps of deliberation and communication.

PtDA development has been guided by several different decision theory and transactional frameworks from economics, psychology, and sociology. The mode of delivering PtDAs varies (print, audiobooklet, videotape, DVD, and, more recently, interactive multimedia Web-based tools). PtDAs are self-administered or practitioner administered; they are used in one-to-one or group situations; and they are used before, during, or after the clinical encounter.

Regardless of the framework, medium, or implementation strategy, there are three key elements common to their design:

1. *Information provision:* For a given clinical condition, PtDAs include high-quality, up-to-date information about the condition or disease stimulating the need for a decision, the available healthcare options, the likely outcomes for each option, the probabilities associated with those outcomes, and the level of scientific uncertainty. The information is clearly presented as a "choice situation," in a balanced manner so as not to persuade the viewer toward any particular option and in sufficient detail to permit choosing among the options.

2. *Values clarification:* A range of methods may be used to help patients consider the personal value or desirability/undesirability of options. First, patients are better able to judge the value of options when they are familiar and easy to imagine. Therefore, PtDAs describe what it is like to experience the physical, emotional, and social consequences of the procedures involved and the potential benefits and harms. Second, patients are asked to consider (either implicitly or explicitly) the positive and negative features that matter most to them. Although there is no clinical trial evidence that explicit methods are always needed, some developers directly engage patients in rating the personal value or importance of each attribute of the options. They argue that these exercises foster engagement, insight, and communication with the others involved. Other developers argue that implicit methods work just as well and are simpler. Moreover, decision scientists have demonstrated in nonmedical contexts that people are not good at predicting the intensity and duration of their feelings regarding future losses or gains. As has been demonstrated in the case of choosing household goods (jam, posters), people may also be more dissatisfied with their choices if they explicitly consider each individual attribute of options rather than make an overall holistic judgment. This debate should be resolved in clinical trials involving practitioners and patients facing real medical decisions.

3. *Guidance in deliberation and communication:* PtDAs are designed to improve patients' confidence and skills by guiding them in the steps involved in decision making regarding the specific choices and showing them how to communicate values and personal issues to their families and practitioners. The PtDAs' structure may provide this guidance implicitly in a step-by-step process, or there may be additional question lists or worksheets for patients to use while discussing the options with their practitioners.

An international group of researchers, known as the Cochrane Review Team of Patient Decision

Aids, has registered more than 500 PtDAs in various stages of development. From this registry, publicly available PtDAs have been described in more detail and evaluated.

Do Patient Decision Aids Improve Decision Quality?

Over the past decade, there has been considerable debate about the definition of a "good decision," when there is no single "best" treatment and choices depend on how patients value benefits versus harms. Recently, the International Patient Standards Collaboration has reached agreement that PtDAs should improve the match between the chosen option and the features that matter most to the informed patient.

The PtDA helps patients recognize that a decision needs to be made, know options and their features, understand that values affect the decision, be clear about the option features that matter the most, discuss their values with their practitioner, and become involved according to their preferred style of interaction.

Angela Coulter has summarized 10 reviews of PtDAs. She concluded that they improve patients' participation, increase knowledge of their treatment options and probable outcomes, and improve agreement between patients' values and subsequent treatment decisions. The use of discretionary surgery decreases without apparent adverse effects on health outcomes. A more recent update of the ongoing Cochrane Collaboration review supports Coulter's conclusions. PtDAs were also shown to reduce the proportion of people who remained undecided postintervention and to improve feelings of being informed and clear about personal values. Exposure to PtDAs reduced not only the rates of discretionary surgery but also the rates of prostate cancer screening (using the prostate-specific antigen [PSA] blood test) and hormone therapy at menopause. PtDAs appeared to do no better than comparison interventions in affecting patients' anxiety, satisfaction, and health outcomes. The degree of detail PtDAs require for positive effects on decision quality should be explored. The effects on persistence with chosen therapies and cost-effectiveness need further evaluation.

How Does One Judge the Quality of Patient Decision Aids?

The International Patient Decision Aid Standards (IPDAS) is a collaboration of more than 100 researchers, practitioners, stakeholders, and policy makers from 14 countries. These collaborators have arrived at a consensus on an internationally approved set of criteria to assess the quality of PtDAs. The criteria focus on the following: (a) essential content (providing information, presenting probabilities, clarifying values, guiding deliberation and communication); (b) development (systematic development process, balance, evidence base, plain language, disclosure); and (c) evaluation (decision quality). The endorsed criteria are summarized in a PtDA users' checklist, which can be used as a guide to payers, practitioners, developers, and researchers.

The collaboration also identified areas for further research. For example, patients' stories (first-person narratives) are commonly used in PtDAs, but there was no consensus on whether they should be considered an essential element. Although some patients may find stories more meaningful than factual information, they may bias patients' decisions, divert them from the facts, or underrepresent different patients' points of view.

Another issue is the use of coaching, which is provided in-person, one-on-one, by a trained person who is supportive but neutral in the decision. It may be given before or after using a PtDA or as part of its delivery. Although one trial demonstrated that coaching added value and cost-effectiveness when women faced options for benign uterine bleeding, there was limited evidence of its incremental benefit.

A final issue involves the application of the IPDAS criteria in evaluating PtDAs. This includes the feasibility, efficiency, and standardized assessment of these criteria. As a placeholder until these issues are resolved, the top-rated criteria from the checklist are being used in rating available PtDAs.

How Are Patient Decision Aids Used in Clinical Practice?

In North America, the use of PtDAs in call centers and public or health plan portals has expanded rapidly. For example, high-volume PtDA producers

estimate that PtDAs were accessed about 9 million times in 2006, mostly via the Internet.

The use of PtDAs as part of clinical care has had a much slower rollout. A recent systematic review identified health professionals' most commonly perceived barriers to implementing shared decision making (SDM): (a) lack of applicability due to patient characteristics, (b) time constraints, (c) lack of applicability due to the clinical situation (e.g., emergency situations), and (d) perceived patient preferences for a model of decision making that does not fit an SDM model. Identified factors that facilitated implementation of SDM were as follows: (a) the perception that SDM will lead to a positive impact on patient outcomes, (b) the perception that SDM will lead to a positive impact on the clinical process, (c) patients' preferences for a model of decision making that fit an SDM model, (d) the motivation of health professionals, (e) the perception that SDM is useful/practical, and (f) the characteristics of the patient.

There are a few notable examples of implementation in clinical care, including the following: (a) primary-care centers at Massachusetts General Hospital, the Dartmouth Hitchcock Medical Center (DHMC), the White River Junction Veterans Administration, 10 community-based primary-care practices at the University of California at Los Angeles (UCLA), and the University of North Carolina; (b) numerous cancer care centers in Massachusetts and at the University of California at San Francisco, Allegheny General Hospital, and DHMC; and (c) orthopedic centers at DHMC and the Ottawa Hospital (Canada).

Delivery of decision support may be some combination of clinical consultation, counseling, provision of PtDAs, and coaching. The sequence, combination, and professionals involved depend on the type of decision, the population, and the service context in which care is provided; these should be spelled out in clinical and care pathways. For example, in the United Kingdom, several urology centers in the National Health Service have care pathways for benign prostatic hyperplasia and early-stage prostate cancer treatments that involve (a) a medical consultation with the urologist to confirm the diagnosis and to clarify the options and roles in decision making; (b) referral to the urology nurse specialist who provides a PtDA about relevant treatments and a personal decision form that elicits decision quality (knowledge, values, and preferred treatment) and unresolved decisional conflict (feeling uncertain, uninformed, unclear about values, unsupported); and (c) a follow-up coaching visit with the nurse specialist to discuss the patient's decisional needs and next steps.

In the United States, the DHMC in Lebanon, New Hampshire, provides decision support through several pathways. First, the Center for Shared Decision Making at DHMC provides consultation and relevant PtDAs for patients dealing with a wide range of preference-sensitive medical decisions. Second, the Breast Cancer Program at DHMC has a specialized care pathway for women diagnosed with breast cancer. As part of the Shared Decision Making process, patients view a video-based PtDA, complete an online tool eliciting their post-PtDA decision quality and unresolved decisional conflict, and then see the surgeon, who discusses the options. Quality of care is audited following the surgeon's consultation at the time of actual treatment choice, using a decision quality audit tool that measures knowledge, values, and choice. Third, DHMC has adapted this specialized pathway to the decision support needs in orthopedic services (back, hip, and knee pain). In some cases, the care pathway is slightly altered. Patients see the surgeon first to determine clinical eligibility for options, then review a video-based PtDA and make a decision. As in the other models, quality of care is audited at the time of actual treatment choice.

In Canada, the Ottawa Hospital is beginning to embed decision support into its care pathways. For example, patients on the waiting list to see a surgeon are screened for surgical eligibility by trained GPs or physiotherapists at an orthopedic intake clinic. Surgically eligible candidates use PtDAs and complete a personal decision form. Summarized data on clinical and decisional needs are forwarded to the surgeon if patients prefer surgery and to the referring physician if patients decline surgery.

Benefits and Barriers

PtDAs are adjuncts to counseling that inform, clarify values, and guide in deliberation and communication. They are superior to standard counseling in improving decision quality (making for informed, values-based decisions), eliminating

indecision, and increasing participation in decision making. They play a role in reducing overuse of discretionary options (surgery, PSA testing, menopausal hormones) that informed patients do not value. There are many barriers to widespread implementation that need to be overcome. Large-scale implementation programs are being developed and evaluated.

Annette O'Connor

See also Decision Making in Advanced Disease; Shared Decision Making

Further Readings

Coulter, A., & Ellins, J. (2007). Effectiveness of strategies for informing, educating and involving patients. *British Medical Journal, 335,* 24–27.

Elwyn, G., O'Connor, A., Stacey, D., Volk, R., Edwards, A., Coulter, A., et al. (2006). Developing a quality criteria framework for patient decision aids: Online international Delphi consensus process. *British Medical Journal, 333,* 417.

Gravel, K., Légaré, F., & Graham, I. D. (2006). Barriers and facilitators to implementing shared decision-making in clinical practice: A systematic review of health professionals' perceptions. *Implementation Science, 1,* 16.

International Patient Decision Aid Standards. (2005). *IPDAS 2005: Criteria for judging the quality of patient decision aids*. Retrieved January 8, 2009, from http://ipdas.ohri.ca/IPDAS_checklist.pdf

Kennedy, A. D., Sculpher, M. J., Coulter, A., Dwyer, N., Rees, M., Abrams, K. R., et al. (2002). Effects of decision aids for menorrhagia on treatment choices, health outcomes, and costs: A randomized controlled trial. *Journal of the American Medical Association, 288,* 2701–2708.

National Steering Group for Decision Support Aids in Urology. (2005, October). *Implementing patient decision aids in urology: Final report*. Retrieved October 4, 2007, from http://www.pickereurope.org/Filestore/Research/Urology_steering_group_report.pdf

O'Connor, A. M., Bennett, C. L., Stacey, D., Barry, M. J., Col, N. F., Eden, K. B., et al. (2008). Decision aids for people facing health treatment or screening decisions. *Cochrane Database Systematic Review, 2*.

O'Connor, A. M., Wennberg, J. E., Legare, F., Llewellyn-Thomas, H. A., Moulton, B. W., Sepucha, K. R., et al. (2007). Toward the "tipping point": Decision aids and informed patient choice. *Health Affairs, 26,* 716–725.

Ottawa Health Resources Institute. (2008). *A-Z inventory of decision aids*. Retrieved January 8, 2009, from http://decisionaid.ohri.ca/AZinvent.php

Ottawa Health Resources Institute. (2008). *Cochrane decision aid registry*. Retrieved January 8, 2009, from http://decisionaid.ohri.ca/cochinvent.php

Special issues on shared decision making. (2007). *Medical Decision Making, 27*(5), 516–713.

Special issues on shared decision making in different countries. (2007). *German Journal for Evidence and Quality in Health Care, 101,* 213–258.

Patient Rights

Patient rights are strict claims made by persons seeking and using healthcare resources. Because an actual right obligates someone or some entity to protect or provide something, healthcare professionals and organizations are responsible for acknowledging, honoring, protecting, and supporting patient rights. At the core of the medical decision-making process are significant patient rights—for example, the rights to be treated respectfully, to participate in decisions affecting one's health and future, and to have one's personal medical information kept confidential. During the past 50 years, especially in North America and Western Europe, there has been significant development in identifying what should be included among patient rights.

Historical Development and Context

Multiple cultural and historical events and developments have shaped the patient rights movement during the second half of the 20th century. One development was the Civil Rights Movement in the United States immediately after World War II. This movement aimed to abolish racial discrimination and segregation and to reassert the dignity and equality of all persons regardless of racial, cultural, or socioeconomic background. Based on the U.S. Declaration of Independence's rights of "life, liberty and the pursuit of happiness," the Civil Rights Movement and its core values emphasized dignity and respect owed to all persons and equal opportunities and freedom for each person to live, work, be educated, and participate in society.

In the 1960s and 1970s, a second set of events leading to the patient rights movement were revelations that many persons had been abused as research subjects. Through the Nazi war crimes trials at Nuremberg (1945–1949), the world learned about the atrocities that occurred in Germany in the name of science before and during World War II. However, two decades after the war, as the medical research enterprise continued to expand, there was publication of instances of exploitation of vulnerable persons (e.g., those with mental disabilities, children, orphans, prisoners, African Americans) in the name of scientific advancement. The core principles of the Nuremberg Code for human research, focusing on voluntary consent and on research subjects' sufficient knowledge and comprehension of proposed research projects, were being ignored. The outcry against this exploitation resulted in regulatory protections for research participants and a reemphasis on the rights of research subjects to be engaged in a voluntary, informed consent process prior to research participation. In 1979, the Belmont Report listed and explained three fundamental ethical principles for governing research: respect for persons, beneficence, and justice.

A third development is captured by a set of court cases focusing on the rights of patients to make their own healthcare decisions, including the right to consent to or refuse medical treatment. For example, in 1972, the Circuit Court for the District of Columbia established an objective standard for disclosure in the informed consent process. In *Canterbury v. Spence*, this court recognized "the prudent-patient test" for disclosure of healthcare information. In essence, this patient-centered "test" is what a prudent person in the patient's position would have wanted to know about the significant risks, harms, and potential benefits associated with the proposed procedure or treatment in order to make an informed decision. In *Roe v. Wade*, the U.S. Supreme Court established a woman's personal right to an abortion; the court justified this assertion by appealing to a right to privacy found in the U.S. Constitution. As a final example, the New Jersey Supreme Court, *In re Quinlan*, found that the constitutional right to privacy encompasses a patient's right and decision to refuse medical treatment, especially when the degree of the treatment's bodily invasiveness increases and the patient's prognosis diminishes.

Negative and Positive Rights

Rights can be categorized as either negative or positive. Simplified, a negative right is the right of persons to be left alone or not to have something imposed on them. Consequently, reciprocal obligations corresponding to negative rights are also negative. For example, a negative right guaranteed by the U.S. Constitution and Bill of Rights is the right to exercise one's religion freely. The reciprocal obligation for governmental or other public agencies is not to interfere with this right through procedures or systems that might impose a particular religion. Furthermore, because freedom of religion is a negative right, governmental agencies are not obligated to provide (and persons are not entitled to receive) resources by which they can practice their religion. An example of a positive right is the right to a speedy and public trial in criminal cases. Such a positive right requires the government to provide (and persons are entitled to have made available to them) mechanisms to receive what is promised—that is, a speedy and public trial.

In healthcare settings, patient rights can also be either negative or positive. Significant negative patient rights (i.e., of noninterference) include the right of privacy and the right to refuse treatment. The right to privacy is foundational to the ethical principle of autonomy and autonomous decision making. The right to refuse treatment is connected to the legal concept of battery, by which persons have the negative right not to have their persons intentionally touched without their permission or consent. However, the *informed* element of the informed consent process compels the positive patient right or entitlement to receive adequate and understandable information before making medical decisions. The next section examines specific rights frequently seen when patient rights are listed and illustrates the dominant influence of negative rights.

Specific Patient Rights

Identifying and listing patient rights is now commonplace. Some states (e.g., California, New York) have legislatively established statements or a Patient

Bill of Rights aimed at safeguarding individuals receiving healthcare services. The Medicare program, the World Health Organization, national healthcare associations and advisory commissions, disease-specific advocacy groups, individual hospitals, long-term care facilities, and healthcare organizations and systems have developed, published, and posted listings of patient rights. Based on many such listings, this section identifies and explains a core set of patient rights. As seen in the explanations below, many of these rights contain elements that are interinfluencing and overlapping.

Dignity, Respect, and Nondiscrimination

The right to be treated with dignity and respect is founded on the recognition that patients are vulnerable, have diminished power and authority in healthcare settings, and need to be protected from exploitation and discrimination. Treating patients with dignity and respect includes honoring personal privacy, appreciating and trying to meet spiritual or religious needs, accommodating cultural beliefs and practices, and ameliorating pain. The right not to be discriminated against derives from the ethical principle of distributive justice or fairness, requiring that patients with similar needs be treated similarly. Race, ethnicity, religious beliefs, sexual orientation, and socioeconomic background are rarely, if ever, relevant for the provision of quality treatment and care. However, these patient characteristics are relevant to patients' experiences of care as respectful and dignified. A patient's gender and age can frequently be relevant for determining some diagnoses and treatment plans; nevertheless, these patient characteristics should never be the basis for subquality care or undertreatment.

Information

Patients' right to information facilitates and supports other rights. For example, patients who do not receive relevant, adequate, and current information about their care are not given due respect and subsequently are unable to participate in informed decision making. Information to which patients have a right includes knowing who their healthcare providers are; the nature, purpose, risks, benefits, and alternatives of proposed treatments or diagnostic procedures; and whether the proposed interventions are research or standard of care. Language barriers can significantly hamper patients' abilities to exercise this right; consequently, some listings of patient rights include the right to language interpreters. The right to information also includes disclosure of medical mistakes, especially if mistakes result in significant patient harms. The right to information is not limited to medical information but includes information about the financial aspects of care and how anticipated outcomes of interventions will affect a patient's quality of life.

Decision Making

The right to make one's own healthcare decisions is fundamental to and well established in U.S. healthcare delivery. The ethical principle of patient autonomy or self-determination supports this right. Implied by this right is the belief that a patient has adequate cognitive skills (i.e., competence or decision-making capacity) to participate in the decision-making process. This right also implies not only that patients can and should authorize treatments but also that they can refuse treatments. This right does not allow patients to demand treatments or procedures that are medically unnecessary. Therefore, the expertise of physicians and other healthcare professionals is essential for educating patients about the proposed procedures and making medical recommendations. However, the right recognizes that patients also have expertise relevant to the decision-making process; that is, they are experts as to their own values, preferences, wishes, and life goals. In pediatric care, this right to participate in decision making is extended to children able to understand age-appropriate healthcare information (usually at about age 6 or 7); although these children lack the cognitive skills to consent to treatment, many can provide assent or agreement to treatment.

Privacy and Confidentiality

The right to personal privacy and to have health information held in confidence is supported by the ethical principle of respect for patient autonomy. Privacy not only implies noninterference in personal healthcare decisions but also extends to

aspects of the healthcare environment, such as having one's body appropriately shielded from others' view whenever possible and having discussions about medical conditions conducted in private areas (not, e.g., in public areas such as elevators, corridors, and cafeterias) and only with those professionals who need to know. Some elements of this right were codified in U.S. federal law in the Health Insurance Portability and Accountability Act (HIPAA). Under HIPAA, patients must authorize disclosure of their protected health information; and they have the right to obtain copies of their medical records and submit amendments if the medical records contain information with which patients disagree. HIPAA also requires healthcare institutions to maintain strict confidentiality through multiple procedural safeguards. The American Hospital Association's (AHA's) Patient's Bill of Rights, published in 1972, included similar language regarding patient privacy and confidentiality.

Quality Healthcare

The provision of healthcare is intended to benefit patients by promoting and maintaining health, curing diseases and disorders, and ameliorating pain and suffering. In essence, patients should be able to expect that healthcare professionals, with appropriate expertise in accord with their roles in the healthcare delivery process, will promote patients' best interests and well-being. The ethical principles of beneficence and nonmaleficence support this right and the expectation that healthcare professionals will consistently try to maximize the benefits of medical interventions and minimize harms and risks. To help promote and honor this right, healthcare organizations and professionals have the responsibility of engaging in continuous quality improvement processes so that systems are improved, efficiencies are maximized, patient safety is enhanced, and errors and near misses are reduced.

Access to Emergency Care

In the United States, there is no recognized general right to healthcare, although a few states (e.g., Oregon, Hawaii, Massachusetts) have made efforts to provide access to basic healthcare interventions for all their citizens. However, under the Emergency Medical Treatment and Active Labor Act (EMTALA), persons in the United States have a right of access to at least emergency treatment and care. EMTALA is a federal law requiring hospitals with an emergency department to assess and stabilize all patients who present to their facilities regardless of the patient's ability to pay for services. Patients can be transferred to other facilities if the facilities where the patients initially presented do not have appropriate medical or surgical expertise to meet their needs. Many patient bills of rights include a right of access to emergency care in accord with this federal mandate. Many patients in the United States, primarily because of their indigence or lack of health insurance, have access to healthcare resources only through emergency departments and therefore present themselves to emergency departments even for nonemergency medical conditions.

Published Standards

In the second half of the 20th century, the patient rights movement experienced significant development and crystallization. In the process, the balance of power and authority in medical decision making shifted in the direction of patients and away from the almost exclusive control of healthcare professionals and organizations. A watershed event reflecting this crystallization occurred in 1992, when the Joint Commission for Accrediting Healthcare Organizations (JCAHO) published its first set of standards on patient rights. However, more recently, some healthcare organizations have introduced language that also asserts patient responsibilities. For example, in recent years, JCAHO's Patient Rights chapter and standards have been revised and are now titled "Ethics, Rights, and Responsibilities"; and AHA's Patient's Bill of Rights has been replaced by a publication titled "The Patient Care Partnership: Understanding Expectations, Rights and Responsibilities."

Martin L. Smith and Margot M. Eves

See also Bioethics; Discrimination; Informed Consent; Informed Decision Making; Models of Physician–Patient Relationship

Further Readings

Annas, G. J. (1989). *The rights of patients: The basic ACLU guide to patient rights* (2nd ed.). Carbondale: Southern Illinois University Press.

Canterbury v. Spence, 464 F.2d 772 (D.C. Cir. 1972).

D'Oronzio, J. C. (2001). A human right to healthcare access: Returning to the origins of the patients' rights movement. *Cambridge Quarterly of Healthcare Ethics, 10,* 285–298.

Inlander, C. B., & Pavalon, E. I. (1994). *Your medical rights: How to become an empowered consumer* (2nd ed.). Boston: Little, Brown.

In re Quinlan, 70 N.J. 10, 355 A.2d 647 (1976).

Merz, J. F. (n.d.). *An empirical analysis of the medical informed consent doctrine: Search for a "standard" of disclosure*. Retrieved March 4, 2009, from http://www.piercelaw.edu/risk/vol2/winter/merz.htm

Poland, S. C. (1997). Landmark cases in bioethics. *Kennedy Institute of Ethics Journal, 7,* 191–209.

Roe v. Wade, 410 U.S. 113, 93 S. Ct. 705, 35 L. Ed. 2d 147 (1973).

Patient Satisfaction

Patient satisfaction refers to the extent to which a patient is satisfied with the healthcare he or she receives. Because patient satisfaction is assessed by self-reports from the patient, it is often considered a patient-reported outcome measure. Satisfaction can be assessed with care received in a variety of settings, such as ambulatory care, nursing home, hospital, and home health. The focus can be on care provided by health plans, provider groups, or individual physicians. Patient satisfaction is important because it provides the patient's perspective on the care delivered and is associated with adherence to medical recommendations, allegiance to healthcare providers, and utilization of care.

Measurement

A direct assessment of patient satisfaction with care can be obtained by asking for an overall evaluation, such as "How satisfied are you with the care you have received during the past 6 months?" (response options: *Very satisfied, Somewhat satisfied, Somewhat dissatisfied, Very dissatisfied*). This direct approach is an efficient way to find out the bottom-line perception of a healthcare consumer, but it provides no specific information about the basis for the perception. Similarly, one could use the approach used in the Consumer Assessment of Healthcare Providers and Systems (CAHPS®) surveys of asking patients to rate the care received on a 0-to-10 response scale, where 0 is the worst possible care and 10 represents the best possible care. The CAHPS option elicits evaluations of care rather than satisfaction per se. But satisfaction and global ratings of care are very highly correlated and provide similar information.

Patient satisfaction is driven by the collective experience with aspects of care (e.g., my doctor listens carefully) weighted by the value or perceived importance of those aspects of care (e.g., it is very important that my doctor listen carefully to me). Most work elicits patient reports about care but does not assess value or importance of the domains of care because of response burden and the relative homogeneity in the importance ratings (i.e., domains assessed on standardized surveys tend to all be rated as important by consumers of healthcare). Reports about care are best elicited by asking how often positive and negative aspects of care occurred—for example, "How often did this doctor spend enough time with you?" or "How often did the clerks and receptionists at this doctor's office treat you with courtesy and respect?"

Eliciting reports about whether and how often patients have specific experiences with care requires more items, but reports are less subjective and are easier to interpret and more useful for healthcare providers than satisfaction ratings or evaluations.

Core Reports of Care Domains

Healthcare is multidimensional, and evaluations of care need to capture the relevant dimensions. Patient assessment of technical quality of care should be avoided because patients are typically not a good source of this information. It is more appropriate to assess technical quality of care using expert consensus, such as the RAND-UCLA appropriateness method. Although the important domains of care for which patients are a good source of information can vary depending on the setting, the core elements across settings include access to care, how well providers communicate with patients, and courtesy and respect from office staff.

The core domains of care account for a substantial portion of the variance in satisfaction and global ratings of care, with the access domain explaining a greater proportion of variance in global ratings of a health plan and the communication domain accounting for a relatively larger proportion of variance in ratings of the doctor or other healthcare provider.

Uses of Data

Patient evaluations of care data are used to compare health plans, physician groups, individual physicians, hospitals, nursing homes, and other providers of care with one another. These data have been collected for reporting to patients who are making choices between different options and to providers who are trying to improve the care they deliver. They are also used by sponsors (e.g., health plans, Center for Medicare and Medicaid Services) to monitor the care delivered to the overall target population and subgroups of the population. In addition, these data are used as a part of accreditation (e.g., National Committee for Quality Assurance) or as a part of pay-for-performance initiatives. In addition, these data are increasingly used by providers as a part of quality improvement efforts.

Reliability and Validity of Patient Evaluations of Care

Reliability refers to the extent to which consistent scores are obtained for the target (e.g., individual doctors) being evaluated. Reliability is estimated by partitioning the observed variance into between-target and within-target variance. For example, the variance between physicians is compared with the variance within physicians (i.e., the extent to which different patients rate the same physician in a similar way) to estimate the reliability of patient evaluations of individual physicians. Extensive work has been conducted to ensure that patient evaluations of care meet accepted standards of reliability and validity.

Mode Effects

Patient satisfaction data are collected using self-administered mail surveys, telephone or face-to-face interviewers, and other modes of data collection (e.g., the Internet). Reports and ratings of care tend to be more positive when an interviewer is involved in the data collection. The likely explanation for this result is that the presence of an interviewer creates a socially desirable response bias. Because of potential mode effects, the same mode should be used consistently whenever possible, but in hard-to-reach subgroups such as Medicaid beneficiaries, it is necessary to use a mixed mode (e.g., mail followed by phone) assessment to maximize the participation rate.

Response Tendencies

A recent study by Robert Weech-Maldonado and colleagues found that Hispanics exhibited a greater tendency toward extreme responding to a 0-to-10 rating scale than non-Hispanic whites; in particular, they were more likely than whites in commercial plans to endorse a "10," and often scores of 4 or less, relative to an omitted category of "5" to "8." These findings suggest caution in the use of central tendency measures and the proportions of ratings on a 10-point scale when examining racial/ethnic differences in ratings of care. It is advisable to consider pooling responses at the top end (e.g., 9 and 10) and lower end (e.g., 0–6) of the response scale when making racial/ethnic comparisons.

Case-Mix Adjustment

When comparing different providers of care, it is important to adjust for differences in the kinds of patients they treat that produce differences in reports and ratings of care that are unrelated to the quality of care delivered. Variables that are frequently adjusted for include age, education, and self-rated health because older, less educated, and more healthy respondents tend to rate their care more positively than younger, more educated, and less healthy respondents.

Proxy Reports

Proxy respondents tend to provide less positive evaluations of beneficiary healthcare experiences, especially for global ratings of care. When a proxy assists the target respondent in completing a survey, the differences are similar but about half as

large. M. Elliott and colleagues found that reports from spouse proxy respondents are more positive than those from other proxies and are similar to what would have been reported by the beneficiaries themselves.

Ron D. Hays

See also Models of Physician–Patient Relationship; Regret; Trust in Healthcare

Further Readings

Elliott, M. N., Beckett, M. K., Chong, K., Hambarsoomians, K., & Hays, R. D. (2008). How do proxy responses and proxy-assisted responses differ from what Medicare beneficiaries might have reported about their health care? *Health Services Research*, *43*(3), 833–848.

Elliott, M. N., Swartz, R., Adams, J., Spritzer, K. L., & Hays, R. D. (2001). Case-mix adjustment of the National CAHPS® Benchmarking Data 1.0: A violation of model assumptions? *Health Services Research*, *36*, 555–573.

Fung, C. H., & Hays, R. D. (2008, August 18). Prospects and challenges in using patient-reported outcomes in clinical practice [Electronic version]. *Quality of Life Research, 17*(10), 1297–1302.

Shekelle, P. (2004). The appropriateness method. *Medical Decision Making*, *24,* 228–231.

Weech-Maldonado, R., Elliott, M. N., Oluwole, A., Schiller, K. C., & Hays, R. D. (2008). Survey response style and differential use of CAHPS rating scales by Hispanics. *Medical Care, 49,* 963–968.

Pattern Recognition

Pattern recognition is the identification and proper labeling of particular configurations of data (i.e., data patterns). The concept includes human- and machine-based pattern recognition, but the latter has been receiving increasing attention in medicine due to advances in computer science and information technology.

Extracting Patterns From Data

The process of diagnosing or making prognostic assessments of disease processes has traditionally been performed by clinicians, who abstract common features in patient data and label these patterns appropriately. These data abstractions are based on findings from anamnesis, physical exam, laboratory tests, and other adjunct diagnostic modalities such as imaging and electrophysiology. Each of these findings can be viewed at different abstraction levels, ranging from their instantiation in a single patient (e.g., "creatinine = 2.5 mg/dl") to an abstract concept that can describe features found in certain types of patients (e.g., "high creatinine," "renal failure"). Diseases or pathologic conditions are often defined based on higher-level abstractions. Abstractions allow easier recognition of patterns in the data, as they help the diagnostician ignore small, irrelevant differences that may originate from noise or from the collection of data unrelated to the recognition of patterns of interest. Some findings are themselves high-level abstractions (e.g., "dark urine"), which are often difficult to quantify.

Discovering regularities or configurations in data is often done at a high level of abstraction (e.g., a pattern of pneumonia given the type of opacity seen in a radiograph). While integration of low-level data is easily achieved by humans, computer-aided pattern recognition algorithms may either directly use the most granular, low-level data (e.g., pixel gray-level intensity) or use these data to first obtain higher-level abstractions and subsequently integrate this information into diagnostic or prognostic categories. For a computer algorithm, categorizing low-level data into intermediate or high-level abstractions may be more difficult than integrating high-level abstractions to make a classification. For example, determining that a collection of white pixels constitutes a "condensation in the left lower lobe" of the lung requires the algorithm to perform segmentation of the image and determine the anatomic location of the abnormal finding. Humans can provide this type of information easily, but a model that uses raw data to classify pixel patterns into this type of high-level abstraction may not be very successful. However, a computer algorithm can easily integrate different types of high-level abstractions, such as "condensation in the left lower lobe" and "high temperature," into a model that helps diagnose pneumonia, provided that enough examples are available. For this reason, it is common to find computer-based

applications that integrate high-level abstractions. Computer-based algorithms for pattern recognition can be didactically divided into two types: unsupervised-learning and supervised-learning algorithms.

Unsupervised Learning

There are situations in which patterns in data are either not known or not labeled in advance. Strategies and methods that are used to organize and extract patterns from these data are called unsupervised because there is no way to guide the pattern recognition process so that it classifies the data into known categories. Cluster algorithms are a good example of unsupervised-learning models. They use measures of similarity or dissimilarity in the data to group cases into unlabeled categories or clusters. By inspecting these clusters, it may or may not be possible to label the clusters using known categories. Unsupervised learning is usually associated with stages of research in which there is less knowledge about the process being studied but there is an indication that certain regularities exist. It is a critical process for medical decision making, often performed by human experts. For example, in the initial phases of the AIDS epidemic, there was only recognition of an unusual pattern of immunodeficiency, which was later determined to result from infection by HIV and properly labeled. The initial recognition of the regularity or abnormal pattern was akin to the utilization of an unsupervised-learning method but performed by a human expert. Once the pattern was known and well described, cases could be properly labeled, and supervised models were constructed to categorize patients into diagnostic and prognostic categories. Those models were commonly developed with computer-based supervised-learning algorithms.

Supervised Learning

Supervised learning refers to a class of modeling strategies and methods that characterize known patterns by "learning" or fitting the parameters of a given model to example data. A logistic regression model is a good example of supervised learning. Logistic regression defines a function that best describes or predicts known data patterns given patient-specific information (e.g., patients who are likely to survive when given a certain treatment). Supervised-learning models of this type are called *classifiers*, as the outputs (or dependent variables) consist of well-defined categories. These models recognize patterns in the data by combining input (or independent variable) data in different ways.

The most common supervised-learning algorithms in clinical medicine are those based on statistical regression or classification trees. They are used in a variety of domains and are the basis for popular predictive models for assessing risks for cardiovascular disease, breast cancer, mortality in intensive care unit settings, and so on. While logistic regression models use a simple function, artificial neural networks, classification trees, support vector machines, and several other types of supervised-learning models use more complex functions. Their potential advantage over simple regression algorithms is that the model developer does not need to manually enter interaction terms to model complex problems in which the data are not linearly separable. Classification trees, artificial neural networks (ANNs) with a hidden layer, support vector machines (SVMs) that use nonlinear kernels, and a variety of other machine-learning models can model complex functions and do not have this limitation. However, their main disadvantage is that they have no interpretable coefficients as in logistic regression models, and hence, they are sometimes considered to be "black boxes." Furthermore, the relative paucity of cases compared with the abundance of measurements per case, coupled with the use of models that have too many possible parameters, makes them very prone to the phenomenon of overfitting. In this situation, models can almost perfectly fit the "training" or example data but often do not generalize well to previously unseen cases. However, as medical data sets grow larger in the number of samples relative to the number of variables per case, the potential advantages of more complex models are expected to be realized.

There is no good way to predetermine which pattern recognition method will perform best for a certain problem, so a good practice is to always establish a simple baseline model against which more complex models can be compared. For supervised-learning classification tasks, this baseline model has been the logistic regression model without interaction terms.

Lucila Ohno-Machado

See also Artificial Neural Networks; Logistic Regression; Recursive Partitioning; Support Vector Machines

Further Readings

Bishop, C. M. (2006). *Pattern recognition and machine learning*. New York: Springer.

Duda, R. O., Hart, P. E., & Stork, D. G. (2001). *Pattern classification*. New York: Wiley-Interscience.

Hastie, T., Tibshirani, R., & Friedman, J. (2001). *The elements of statistical learning*. New York: Springer.

Ripley, B. D. (1996). *Pattern recognition and neural networks*. Cambridge, UK: Cambridge University Press.

Schurmann, J. (1996). *Pattern classification*. New York: Wiley-Interscience.

Personality, Choices

Personality specifies a stable set of individual psychological characteristics that influence thinking, motivation, and behavior in various situations. Personality has to be distinguished from individual differences due to, for example, cultural heritage or socioeconomic class. Rather than being solely determined by the psychosocial environment, personality is thought to be the result of an interaction between genetic and environmental factors that lead to an underlying behavioral disposition and thus constitutes the psychological uniqueness of a human being. A current personality concept is the Big Five, a set of five independent traits describing a person's character: (1) openness to experience, (2) conscientiousness, (3) extraversion, (4) agreeableness, and (5) neuroticism. Personality has been proposed as a factor affecting the behavior of patients seeking medical treatment, especially with regard to treatment adherence, participation in healthcare decision making, and treatment choice. The following sections describe the extent to which personality plays a role in these areas of treatment delivery.

Selected Personality Traits

Comprehensive personality concepts such as the Big Five that claim to be able to describe all relevant personality differences have to be distinguished from single, usually theory-based personality traits that describe a specific attribute such as generalized self-efficacy, locus of control, or optimism. Of the Big Five traits, openness to experience describes imaginative, intellectually curious people versus straightforward, conservative people less interested in change. Conscientious people are disciplined, hardworking, orderly, and thorough. Extraversion describes people seeking stimulation and the company of others as well as showing initiative and being action oriented. An agreeable person is someone who is compassionate, cooperative, and interested in the well-being of others but may shy away from conflict. Neuroticism describes a person who often experiences negative feelings and views situations as potentially threatening and difficult, which impedes his or her capacity to deal with emotional difficulties. With regard to single personality traits, generalized self-efficacy describes the ability to generally handle difficult situations well. Internal locus of control describes the belief that one personally can influence situations, whereas external locus of control stands for the conviction that others have a greater control over a given situation than oneself. Optimism describes the tendency of an individual to have a positive outlook on events and the belief that things will take a positive turn.

Personality and Behavior

Personality is only one factor affecting a person's behavior in a given situation. In addition, a person's behavior is also determined by factors such as beliefs, expectations, prior experiences, roles, and situational constraints or incentives. Thus, the importance of personality in determining an individual's behavior in a given situation will be limited. On the other hand, personality does predict general behavioral tendencies such as stress-coping behavior. An active, problem-focused coping style, for instance, is far more likely in individuals high in extraversion and low in neuroticism. Health behaviors such as exercise typically are determined by beliefs regarding the benefits of this particular behavior, the opinion of peers with regard to it, and the extent to which the individual experiences control over the behavior. With the exception of generalized self-efficacy and conscientiousness,

personality generally plays less of a role in predicting health behaviors.

Personality and Treatment Adherence

Treatment adherence describes the extent to which a person follows the prescribed or advised medical treatment. A related term is treatment compliance, though this term may have a more paternalistic connotation. Treatment adherence is considered an important factor in medical treatments as failure to adhere to treatment may render it ineffective. Several factors influence the extent to which a patient adheres to his or her treatment, such as understanding of prescriptions or medical advice, memory, the occurrence and extent of side effects, social support, the patient-physician relationship, and the personality variables conscientiousness and self-efficacy. Studies show that conscientiousness, but no other Big Five personality trait, predicts adherence to medical treatment in renal dialysis patients, HIV-positive patients, and patients with high cholesterol. Self-efficacy beliefs, on the other hand, tend to predict adherence to behavioral treatments and exercise as the individual potentially has to actively overcome obstacles to initiate and maintain the required behaviors. Generally, however, personality variables only play a limited role in the extent of treatment adherence.

Personality and Treatment Choice

Treatment preference is the extent to which a certain treatment is preferred over another therapy; treatment choice describes the patient's selection of a treatment when similarly efficacious treatments are available. Personality differences have been detected between patients opting for complementary and alternative medicine (CAM) and those patients relying on conventional medicine. Patients choosing CAM for the treatment of chronic pain or cancer have been described as less agreeable, more conscientious, and having a greater fighting spirit and a higher level of locus of control. In other areas, personality tends to have only a small influence on the patient's treatment choice. For example, no differences were found between patients choosing radical prostatectomy (i.e., the surgical removal of the prostate gland) and radiation therapy, nor were differences found between women choosing hormone treatment or psychological treatment for hot flashes. Also, the preference of passive therapies such as massages over active therapies such as exercise for the treatment of chronic pain is independent of personality. Obviously, other factors such as disease concerns, the physician's recommendations, or personal experience with a certain treatment are more relevant than personality for the patient's decision-making process.

Personality and Participation in Healthcare Decision Making

Healthcare decision making refers to the choice and planning of treatment. Involvement of the patient in this decision-making process generally will improve treatment adherence and outcome and is therefore held to be advantageous. However, a patient's interest in participating in medical decision making varies. Next to factors such as patient-physician communication, the personality of the patient plays a role in the extent to which he or she will wish to participate. Patients interested in active decision making are more conscientious, more open to experience, less agreeable, and less neurotic and have a higher internal locus of control. On the other hand, shy patients and patients with low self-efficacy and a tendency to believe that others know better will be more comfortable if the medical decisions are made by their physician. Thus, doctors should acknowledge these differences in their communication style and treatment efforts.

Gerhard Blasche

See also Advance Directives and End-of-Life Decision Making; Decision Making and Affect; Decision-Making Competence, Aging and Mental Status; Decision Making in Advanced Disease; Decisions Faced by Patients: Primary Care; Informed Decision Making; Patient Decision Aids

Further Readings

Blasche, G., Melchart, H., Leitner, D., & Marktl, W. (2007). Personality does not predict treatment preference, treatment experience does: A study of four complementary pain treatments. *Forschende Komplementärmedizin, 14*(5), 274–280.

Block, C. A., Erickson, B., Carney-Doebbling, C., Gordon, S., Fallon, B., & Konety, B. R. (2007). Personality, treatment choice and satisfaction in patients with localized prostate cancer. *International Journal of Urology, 14,* 1013–1018.

Davidson, R., Geoghegan, L., McLaughlin, L., & Woodward, R. (2005). Psychological characteristics of cancer patients who use complementary therapies. *Psychooncology, 14,* 187–195.

Flynn, K. E., & Smith, M. A. (2007). Personality and health care decision-making style. *The Journals of Gerontology. Series B, Psychological Sciences and Social Sciences, 62*, P261–P267.

Ryckman, R. M. (2004). *Theories of personality*. Pacific Grove, CA: Brooks/Cole.

Person Trade-Off

The person trade-off (PTO) is a method to elicit social preferences that has been advocated for use in cost-utility analyses instead of elicitations that yield individual utilities for different health states. The PTO is thought to incorporate societal considerations about how treatment benefits are distributed across a population rather than simply maximizing the health benefit of treatments. Allocation decision makers are increasingly acknowledging the need to incorporate social value judgments of distributive effects in allocation decisions and are increasingly turning to the public for input. Thus, the PTO elicitation method is intended to derive social preference values in order to incorporate consideration of the distributive effects of treatment benefits in an allocation setting. It is unique in embracing considerations of distributive justice. The following sections describe the PTO elicitation method, the rationale for the method, its application, and its challenges.

Elicitation Method

A typical PTO elicitation asks respondents to imagine that they are a decision maker faced with having to choose between two equally expensive healthcare treatment programs that improve quality of life or save lives for varying groups of patients. There is only enough money to fund one of the two mutually exclusive programs. Respondents must decide which program they would fund. Fixing the number of patients in one of the programs, respondents are asked how many patients would need to be treated to make them indifferent between the two programs. For example, Program A might extend the life of 100 healthy individuals for 1 year. Program B might cure 100 individuals of a chronic health condition. Many people may choose to fund Program A because of the imperative to save lives. If this is the case, respondents are then asked how many patients must be cured in Program B for it to be equally good as Program A. For example, respondents may give a median value of 1,000. Thus, 1,000 individuals would need to be cured of the chronic health condition to make the program equally good as another program that saves the lives of 100 healthy people.

Computing Preference Weights

Similar to individual utilities, preference weights can be computed from PTO responses on a 0-to-1 scale, where 0 is equal to death and 1 is perfect health. If A_i is the baseline number of individuals treated in Program A and B_i is the number of individuals who must be treated in Program B for it to be equally good as Program A, then a preference weight for Program B (W_B) can be computed as

$$W_B = 1 - \frac{A_i}{B_i}.$$

Thus, in the example above, the preference weight for curing the chronic health condition is 0.9 (1 – 100/1000). The particular equation used to derive a preference weight depends on the elicitation and the baseline used for comparison. For example, some elicitations compare curing one chronic (or acute) health condition versus another. To compute weights directly on a 0-to-1 scale, a comparison must be made with saving a life or preventing the onset of a condition. Weights can also be computed indirectly by "chaining" a series of elicitations. However, this approach has not been tested empirically and may introduce new sources of biases into the estimates.

Mode of Elicitation

PTO elicitations are difficult for people to comprehend; it is difficult to know how many

individuals would have to be treated for a program to be equally good as another program treating another group of individuals under any circumstance. When asked, most respondents agree that the elicitation is hard. Some researchers advocate eliciting values through face-to-face interviews because responses to paper surveys have yielded highly inconsistent responses. The interviewer is able to explain the task more fully, answer questions, and help ensure that the respondent understands the task; the respondent may make more effort to give a thoughtful response; and the interviewer can check the responses and resolve inconsistencies if needed. One study found no difference, however, between in-person interviews and a computerized elicitation program administered over the Internet. The computerized program mimicked an in-person interview with comprehensive instructions, help, and tips and incorporated simple consistency checks.

Little empirical support exists for identifying an optimal search routine to elicit the point of indifference. Techniques include using an open-ended question ("Enter the number of individuals who would need to be treated"), a ping-pong search procedure (numbers of individuals are traded back and forth between high and low values in an iterative search to close in on an indifference point), or titration (values are incrementally increased or decreased until the point of indifference is reached).

Rationale

Allocation decisions based on the results of cost-effectiveness analyses that use quality-adjusted life years (QALYs) derived from individual utilities for treating various health states assume that all QALY distributions are equally valued. A small number of people gaining a large number of QALYs is valued equally as a large number of people gaining a small number of QALYs. However, many people, including bioethicists, are concerned that this approach does not support a just distribution of healthcare treatment benefits because it fails to put more weight on people who are worse off, places more value on saving the lives of healthy individuals than individuals with a debilitating health condition, and does not put a sufficiently high priority on treating individuals in need.

Erik Nord was an early proponent of the PTO elicitation method, proposing that societal preference weights could be used to compute QALYs gained under treatment programs instead of individual utilities. Although the PTO method has intuitive appeal, there is no formal underlying theory to support it. Its choice-based property appeals to economists as a way to reveal preferences for a good not available in the market; however, choices are made in a social context, and thus, economic theory related to consumer choice cannot apply. It does have a hypothetical advantage in that it asks respondents to make trade-offs between groups of people, which mimics actual allocation decisions.

Elicitations From Individual Utilities

Numerous studies have shown striking differences in preferences elicited using the PTO method versus methods eliciting individual utilities. Differences are most striking when lives are at stake. In general, respondents do not want to discriminate between groups of individuals when lives are at stake; they more frequently take a societal (rather than personalized) perspective when responding and often place more value on curing people who are worse off than would be indicated by the gain from treatment.

Challenges

A number of challenges continue to plague the complex task of eliciting values from respondents.

Refusals

Respondents can give one of two types of refusals: (1) equivalence refusals, where the two treatment programs under consideration are equal in value, or (2) off-scale refusals, which occur when respondents give an inordinately large point of indifference. For example, a respondent may give an equivalence refusal by saying that a program to cure foot numbness may be equally good as another that cures paraplegia. Alternatively, the respondent may give an off-scale refusal by saying that 6 billion individuals would need to be cured of foot numbness for the program to be equally good as the one curing paraplegia. The PTO, as

with individual utility elicitation methods, is plagued by refusal responses. Across many studies, 12% to 91% of responses have been identified as equivalence refusals, the most common type of refusal. Off-scale refusals have accounted for 4% to 19% of responses.

In some circumstances, refusals are appropriate. The most common example of this situation is when respondents place equal value on saving the lives of healthy individuals and individuals with a debilitating health condition. On the other end of the scale, off-scale refusals can occur when a program curing a very minor condition is compared with a program for a seriously debilitating condition. In fact, the number of equivalence refusals monotonically increases as severity of the health conditions in the two groups of individuals decreases, and vice versa for off-scale refusals. However, many refusals arise because of other effects that are described in the following sections.

Elicitation Method Effects

Ideally, PTO responses would not be influenced by the method of elicitation. However, ample evidence exists that the way the questions are posed influences responses. Particular attention has been paid to the influence of framing on the proportion of refusals. Equivalence refusals are minimized when the elicitation emphasizes the constraint of money and that choosing one program over another will result in a group of individuals not getting treatment.

Careful consideration must be given to what comparisons respondents are expected to make. For example, if Program A extended the life of 100 healthy individuals by 1 year and Program B extended the life of 100 individuals with a debilitating health condition for 1 year, many respondents would call them equivalent because the decision to save lives should not be based on prior health condition. If the comparison were varied so that Program B would cure 100 individuals with a debilitating health condition who will live another year either way, respondents might place higher value on Program A because lives will be saved. If Program B were varied yet again, where 100 individuals would avoid contracting the debilitating health condition, responses are likely to change yet again. Mathematically, responses in all three elicitation scenarios should be the same. However, different framing evokes different ethical considerations and other unknown psychological phenomena, affecting how people respond to elicitations.

Influence of Individual Attributes

Paradoxically, respondents who say the elicitation is hard are less likely to give equivalence refusals. These respondents are more likely to give off-scale refusals, however. Respondents who are outraged at the idea of having to choose one treatment program over another, have below-median education, or have below-median numeracy are all more likely to give an equivalence refusal.

Internal Consistency

A number of problems have been found with numeric responses to PTO elicitations. The good news is that respondents are generally ordinally consistent—indifferent points are consistent with rankings of health conditions and across varying baseline group sizes. In addition, PTO responses do appear to yield a consistent set of core values for a range of health states that incorporates distributive concerns when compared with measures of individual utility.

However, numeric indifferent points vary based on the size of the baseline group of individuals being considered. In the earlier examples, 100 individuals were in the baseline group. If this number is varied downward or upward, indifference points change. More work is needed to understand why. Perhaps responses are rational and vary because the marginal value of person-years between the two programs changes based on quantity. Questions also remain about other implicit underlying assumptions of PTO preferences. For example, is there diminishing marginal value for years of life gained? Does the marginal rate of substitution of healthy individual person-years for disabled person-years vary? Answers to these questions remain elusive.

Laura J. Damschroder

See also Bias; Cost-Utility Analysis; Disability-Adjusted Life Years (DALYs); Equity; Quality-Adjusted Life Years (QALYs); Utility Assessment Techniques

Further Readings

Damschroder, L. J., Baron, J., Hershey, J. C., Asch, D. A., Jepson, C., & Ubel, P. A. (2004). The validity of person tradeoff measurements: Randomized trial of computer elicitation versus face-to-face interview. *Medical Decision Making, 24*(2), 170–180.

Damschroder, L. J., Roberts, T. R., Goldstein, C. C., Miklosovic, M. E., & Ubel, P. A. (2005). Trading people versus trading time: What is the difference? *Population Health Metrics, 3*(1), 10.

Damschroder, L. J., Roberts, T. R., Zikmund-Fisher, B. J., & Ubel, P. A. (2007). Why people refuse to make tradeoffs in person tradeoff elicitations: A matter of perspective? *Journal of Medical Decision Making, 27*(3), 266–280.

Damschroder, L. J., Zikmund-Fisher, B. J., & Ubel, P. A. (2005). The impact of considering adaptation in health state valuation. *Social Science & Medicine, 61*(2), 267–277.

Green, C. (2001). On the societal value of health care: What do we know about the person trade-off technique? *Health Economics, 10*(3), 233–243.

Mansley, E. C., & Elbasha, E. H. (2003). Preferences and person trade-offs: Forcing consistency or inconsistency in health-related quality of life measures? *Health Economics, 12*(3), 187–198.

Nord, E. (1995). The person-trade-off approach to valuing health care programs. *Medical Decision Making, 15*(3), 201–208.

Nord, E. (1999). *Cost-value analysis in health care: Making sense out of QALYs*. Cambridge, UK: Press Syndicate of the University of Cambridge.

Pinto-Prades, J. L., & Abellan-Perpinan, J. M. (2005). Measuring the health of populations: The veil of ignorance approach. *Health Economics, 14*(1), 69–82.

Salomon, J. A., & Murray, C. J. (2004). A multi-method approach to measuring health-state valuations. *Health Economics, 13*(3), 281–290.

PHARMACOECONOMICS

Pharmacoeconomics is a field of study that simultaneously considers the clinical consequences and costs attributed to the use of pharmaceutical products and services. Today's healthcare professionals encounter challenges determining optimal clinical and humanistic outcomes that minimize economic impacts on both individuals and society. By using pharmacoeconomic principles, analytical methods, and data, decision makers, including healthcare providers, payers, policy makers, and patients, can make informed medical decisions regarding optimal pharmaceutical care as well as allocations of scarce medical and financial resources.

History

Pharmacoeconomics consists of two root words: *pharmaco* and *economic*. The word *pharmaco* comes from the Greek word *pharmakon*, which means poisons or drugs. The word *economic* comes from two Greek words: (1) *oikos*, meaning the household or family estates, and (2) *nomos*, meaning rules, natural laws, or laws. As these root words imply, pharmacoeconomics derives its theoretical framework from principles of economics and social sciences and integrates them with pharmaceutical sciences.

Pharmacoeconomics is sometimes viewed as a subdiscipline of health economics. The basic concepts of economic analyses in pharmacy were introduced in the 1970s, and the foundation of pharmacoeconomics was set in the 1980s. William McGhan, C. R. Rowland, and J. Lyle Bootman at the University of Minnesota introduced cost-benefit analysis and cost-effectiveness analysis to pharmacy students as early as 1976, published the concepts of economic analysis in pharmacy in the *American Journal of Hospital Pharmacy* in 1978, and reported the first cost-benefit analysis related to pharmaceutical care in 1979. The term *pharmacoeconomics* was first introduced in 1987 in an article titled "Postmarketing Drug Research and Development," by Raymond Townsend, who advocated the need for pharmacoeconomic research. Later, Bootman, Townsend, and McGhan edited a book titled *Principles of Pharmacoeconomics*, which is the first textbook in this discipline.

In the 1970s, pharmacoeconomic education and research in the United States were developed for graduate students and researchers. Since the 1990s, more and more pharmacy schools include this discipline as an elective course in the professional pharmacy curriculum. Recent surveys indicate that 80% of pharmacy schools in the United States offer pharmacoeconomics at the professional level and 52% of pharmacy schools outside the United States offer this course at either the graduate or the

professional level or both. Outside the academic field, there has been an increase in clinical applications of pharmacoeconomic analysis to facilitate medical decision making. This indicates that clinicians are getting more involved in pharmacoeconomic research. Overall, pharmacoeconomics has received more attention clinically and globally among researchers and healthcare professionals and is becoming an important part of the pharmacy curriculum.

Pharmacoeconomic Research

Identification and Measurement of Outcomes

Pharmacoeconomic research identifies, measures, and evaluates clinical, humanistic, and economic outcomes simultaneously among competing therapeutic alternatives of interest. Outcomes represent the overall consequences from multiple causes. Therefore, the portions of outcomes attributed by the interested alternatives must be properly and articulately identified. This requires a deep understanding of the topic of interest and sufficient scientific training to identify relevant sources of causes. Often, outcomes cannot be directly measured, so a surrogate will be used or a modeling technique can be performed to reasonably estimate or predict outcomes.

Outcomes are frequently classified as economic, clinical, or humanistic in the discipline of pharmacoeconomics. Economic outcomes are measured as resources utilized, which are then assigned monetary values. Costs are considered differently in terms of perspective (e.g., patients, healthcare providers, third-party payers, or society), type (e.g., direct medical, direct nonmedical, indirect, intangible), services received (e.g., medications, professional services, personnel, facility), sources of measurement (e.g., actual, estimated), and other (e.g., time difference, foreign exchange). Because of the different considerations of costs, the comparability and generalizability of pharmacoeconomic data are often arguable. Therefore, economic outcomes must be compared and interpreted with caution.

Clinical outcomes normally refer to clinical end points of interest. The selection of clinical end points and time period considerations depends on the nature of the conditions being studied, the target population to which study results apply, and clinical judgments where sufficient effects can be captured. For example, it may take decades for development of osteoporosis in the general population but only months for drug-induced osteoporosis to develop; women have higher prevalence and incidence rates of osteoporosis than men; and fracture rates due to osteoporosis may be more appropriate than values of bone mineral density (BMD) as clinical outcomes in pharmacoeconomic analysis. Appropriate measurement of clinical outcomes must be able to translate to medical decision making in clinical practice.

Humanistic outcomes, including health-related quality of life and patient preference and satisfaction, receive much attention nowadays yet are more difficult to be quantified and measured than economic and clinical outcomes. This type of outcome, recognized as a patient-reported outcome, is relatively subjective. The development of instruments involves a very complex process of psychometrics and validation, so that instruments measure a wide spectrum of domains (i.e., physical health/functioning, mental health/functioning, and general health/well-being) and are "responsive" (i.e., capable, sensitive, and specific) to capture clinical changes over time or subtle differences in humanistic outcomes. Instruments that have been designed for measuring general health profiles include the Medical Outcomes Study Short Form (MOS-SF-36), EuroQOL (EQ-5D), Health Assessment Questionnaire (HAQ), Health Utilities Index (HUI), Nottingham Health Profiles, and Sickness Impact Profile (SIP). The general instruments may not be responsive enough to subtle changes in respondents with specific conditions, so disease-specific or population-specific instruments have been developed. Examples of specific instruments are the Minnesota Living With Heart Failure Questionnaire, Quality of Life in Epilepsy (QOLIE), American Urological Association Symptom Index (AUASI), European Organization for Research and Treatment of Cancer (EORTC QLQ-C30), Crohn's Disease Activity Index (CDAI), Arthritis Impact Measurement Scales (AIMS), and West Haven-Yale Multidimensional Pain Inventory (WHYMPI or MPI).

Evaluation and Methodology

It is a unique feature that pharmacoeconomics integrates clinical outcomes measures along with

financial and economic theories and techniques. Historically, the typical techniques used in pharmacoeconomics include cost-minimization analysis (CMA), cost-effectiveness analysis (CEA), cost-benefit analysis (CBA), and cost-utility analysis (CUA). The main difference among these techniques is the unit of outcomes measurement for analysis. For example, CMA evaluates cost differences, CEA evaluates costs per natural unit of outcomes, CBA evaluates net benefits where all outcomes have been assigned monetary values, and CUA evaluates costs per quality of life changed. Additional analyses include cost-of-illness analysis, quality-of-life assessment, and disease state management. These analyses are frequently embedded within decision analysis for medical decision support.

These types of outcomes measures often encounter a degree of uncertainty, and therefore, more information is needed for medical decisions. There are four sources of uncertainty:

1. *Methodological uncertainty* comes from the disagreement among analysts who use different analytical methods in terms of definition, inclusion, measurements, and valuation of outcomes in the analysis, so the study results may not be directly comparable.
2. *Parameter uncertainty* refers to uncertainty of model inputs (parameters). Sampling variation from different inclusion/exclusion criteria and sample characteristics is also classified as parameter uncertainty.
3. *Modeling uncertainty* includes uncertainty due to model structure and the whole modeling process. Basically, model analysts must explicitly describe study methods in detail so that the sources of uncertainty can be identified accordingly.
4. *Generalizability* implies the uncertainty of extrapolating study results to the general target population. Assumptions and limitations should also be assessed when results are interpreted. The use of a "reference case" of core methods is advocated for handling methodological uncertainty and makes possible comparisons of results among studies using different analytical methods.

Uncertainty in pharmacoeconomics research is usually handled by sensitivity analysis. One-way sensitivity analysis varies values of one key variable at a time to evaluate the impact of this particular variable on outcomes of interests and is the most frequently used technique. Two-way sensitivity analysis varies values of two key variables at the same time, and multivariate sensitivity analysis simultaneously varies values of multiple key variables. The larger the number of variables examined, the more complex the calculations involved in the sensitivity analysis. Empirically, multivariate sensitivity analysis is performed by using techniques of modeling and simulations, such as second-order Monte Carlo simulations. Interested readers may review corresponding topics in this encyclopedia for more details on analytical techniques. In any case, pharmacoeconomists must continuously adopt existing techniques and approaches from other disciplines and develop methodology for better research practice.

Interpretations and Presentation

The results of pharmacoeconomic analysis must be comprehensible to readers and decision makers with different levels of backgrounds. The information must be conveyed with transparent manipulations, sufficient details, and a summary in the simplest form so that informed medical decisions can be made accordingly. Transparency in methodology is especially important in modeling studies where assumptions, simplifications, and parameter settings may alter the interpretation and generalizability of study results. Some peer-reviewed journals, such as *PharmacoEconomics* and *British Medical Journal*, have published checklists or good-practice guidelines for published articles in economic studies.

Pharmacoeconomic results can be presented in many ways. The most common presentation in cost-effectiveness analysis is the incremental cost-effectiveness ratio (ICER), which is the additional cost needed or incurred to have a unit change of effectiveness in an alternative compared with the reference or control group, which is usually the standard care for the condition of interest. Another example of presentation in cost-benefit analysis for decision makers is willingness to pay (WTP), which is the ceiling cost or maximal allowance for

the difference between two alternatives. Along with the WTP approach, acceptability curves are figure presentations of estimated possibilities of outcomes given various values of WTP. Additional information on probabilities of outcomes is especially useful to a variety of readers and decision makers evaluating different scenarios.

Assessment of Pharmacoeconomic Research

Several points are recommended for readers to assess the quality of pharmacoeconomic research. The checklist includes (a) the appropriateness of the title indicating key components of the study; (b) significant study questions, clear objectives, and hypotheses; (c) the perspective addressed; (d) specifications of study alternatives matching clinical practice; (e) proper identification of the healthcare resources consumed; (f) justification of the outcomes measured; (g) right type of analysis and appropriate use of techniques; (h) uncertainty and assumptions addressed; (i) reasonable interpretations and comprehensible presentations; (j) conclusion and suggestions based on study findings; (k) limitations addressed; and (l) unbiased, impartial attitude portrayed.

Guidelines

Because of the variety of perspectives, study designs, outcomes measures, and methods, pharmacoeconomic evaluation encounters problems of lack of comparability across studies as well as questionable generalizability to populations in clinical practice. This leads to the need for consensus, guidelines, a standard template of study designs, and reporting of results. An article titled "Pharmacoeconomic Guidelines Around the World," which was published in the *International Society for Pharmacoeconomics and Outcomes Research (ISPOR) Connections* (ISPOR member newsletters) and the ISPOR Web sites in August 2004, compares 28 guidelines from 23 countries with 32 key features. In the United States, guidelines for good research practice regarding cost-effectiveness studies were provided in 1996 by the Panel on Cost-Effectiveness in Health and Medicine, U.S. Public Health Service. The 1996 recommendations introduced the concept of reference case analysis, which enhances comparability across studies. Later, the Academy of Managed Care Pharmacy (AMCP) published formulary submission dossiers in October 2002 and a revision in April 2005, which were based on comments from managed-healthcare systems, pharmacy benefit management companies, and the pharmaceutical industry, in an effort to standardize formulary submissions of clinical and economic data by healthcare systems in the United States.

Applications

Pharmacoeconomic studies are particularly important to the pharmaceutical industry. Pharmacoeconomic evaluations can be implemented in any phase of clinical trials, and the results can be used in strategic decisions of study designs, marketing, pricing, and reimbursements. For example, outcomes measures in Phase II of clinical trials may not be perfect and can be further modified for Phase III of the trials. Additionally, investigators can use pharmacoeconomic results from Phase II to predict possible outcomes in Phase III. As a result, clinical trials in Phase III can be designed to measure the most appropriate outcomes in a cost-effective manner. Furthermore, analysis of return on investment (ROI) helps pricing strategies. Pharmacoeconomic results in postmarket surveillance provide information, facilitating decisions on marketing and reimbursement. Currently, submission of pharmacoeconomic data to the U.S. Food and Drug Administration (FDA) for new-drug approval is not mandated, but it is encouraged.

Pharmacoeconomic analysis is also very useful to clinicians for their daily practice. For example, the pharmacy and therapeutics (P&T) Committee in hospitals, health systems, or managed-care organizations may use pharmacoeconomic results for formulary decisions and management, such as projected impacts of new drugs on formulary, evaluation and comparison of drugs in the same pharmacologic or therapeutic class, and making guidelines for therapeutic interchange. Medication prescribers may compare drug costs and utilization across inpatient units to evaluate the prescribing patterns associated with patient outcomes. Medication safety teams use pharmacoeconomic analysis to evaluate the benefits and effectiveness of medical safety initiatives for patients and hospitals. Pharmacy directors may use pharmacoeconomic

analysis to evaluate the economic impacts of pharmacist interventions on patient outcomes, so that pharmacist services can be evaluated and quantified. From the operational point of view, pharmacoeconomic analysis evaluates outcomes from different pharmacist practice models (e.g., centralized or satellite pharmacies), the time differences between medication order and delivery, and the association between pharmacist satisfaction and workload and improvement of order-processing flows. Clinicians may also share pharmacoeconomic outcomes with patients to decide together on the best therapeutic plan so that patient compliance is increased, which in turn results in better outcomes.

From the perspective of community pharmacy practice, cost-benefit analysis evaluates the benefits of screenings, vaccinations, preventive interventions, weight loss programs, and smoking cessation programs. Budget impact analysis helps in pharmacy inventory management and allocation of healthcare resources. Pharmacoeconomic analysis facilitates decisions regarding pricing and marketing strategies for community pharmacies. From the perspective of other pharmacy practices, pharmacoeconomic analysis evaluates the success of disease management programs in managed care. Pharmacoeconomic analysis can be used in public health by policy makers to make informed decisions. For example, national budgets can be properly allocated by conducting a budget impact analysis for Medicare Part D. Overall, pharmacoeconomics facilitates medical decision making regarding optimal pharmaceutical care with minimal financial impacts.

Future Directions and Challenges

With the increase in the recognition and application of pharmacoeconomic studies in clinical practice, it is expected that more schools and teaching institutions will include pharmacoeconomics in their professional pharmacy curriculum and residency programs. Despite the many guidelines that have been established, some still challenge pharmacoeconomic studies. Country-specific guidelines were developed to fit regional needs, yet international guidelines would be beneficial to the global society. A consensus of transparent methodology, especially for modeling studies, has been reached globally; however, the components of the methodology remain debatable. Other issues include bias and ethical dilemmas; transferability of economic data, including retrospective data; connections between evidence-based medicine and outcomes research; standardization of drug costs; standardization of fellowship and education programs; and instrument development for patient-reported outcomes.

Example of Pharmacoeconomic Research

When a patient takes glucocorticoid medications for more than 3 months, an important side effect is loss of bone mass, which increases the risk of osteoporosis and fractures and also affects the overall quality of life. Several medications have shown promising results for the prevention of osteoporosis-related fractures. However, a relatively low percentage of long-term glucocorticoid users have received these medications from their physicians to prevent glucocorticoid-induced osteoporosis (GIOP) and fractures. Given that development of osteoporosis is slow and considering the economic burden of this illness, physicians and patients are not sure whether it is better to use these medications to prevent GIOP and fractures in patients who take glucocorticoids for a long period of time or to just treat fractures when they occur. A cost-effective analysis was conducted to determine which therapeutic approach is preferable by considering both the costs and the outcomes of therapy. Additionally, the analysis projected long-term estimates from nationally representative survey data by using a technique called Markov modeling, so that the results reflect real-life situations as closely as possible over the long term. A second-order Monte Carlo simulation served as a tool of sensitivity analysis to address uncertainties at the level of all variables simultaneously.

The study results were presented with cost-effectiveness ratios, ICERs, and acceptability curves with WTP against percentages of chances that a specific alternative is cost-effective. Of 1,692 qualified female long-term glucocorticoid users (representing 2.65% of the female noninstitutionalized U.S. population, average age = 49.8 years, average prednisone-equivalent dose = 10.7 mg/day, average duration of therapy = 215 days, percentage of whites = 85.6), 29.9% reported use of any antiresorptive agent; of those, 76.5% used hormone

replacement therapy (HRT) only, 12.1% used bisphosphonates only, 2.0% used calcitonin only, 1.6% used raloxifene only, and 7.8% used more than one antiresorptive agent. Reference case analysis showed that compared with the controls, the estimated 10-year/lifetime ICERs (cost per fracture avoided) were $2,250 to $7,776 for HRT, $10,149 to $28,078 for bisphosphonates, $27,891 to $46,102 for raloxifene, and $60,862 to $61,660 for calcitonin in hypothetical 50-year-old female glucocorticoid users. By using the cost-effectiveness acceptability curve, different decision makers may find the corresponding range of probabilities that remain cost-effective based on personalized WTP.

Some assumptions and limitations include small sample sizes for the calcitonin and raloxifene groups and a likely selection bias in that bisphosphonate users are more likely to report a longer duration of glucocorticoid therapy. Because few guidelines included cost-effectiveness information, consideration of these results may facilitate better management of GIOP. Accordingly, this information helps decision makers determine whether it is better to use medications to prevent GIOP and fractures and if the use of these preventive medications is warranted, to choose the best option for long-term glucocorticoid users.

Jun-Yen Yeh and Morton P. Goldman

See also Acceptability Curves and Confidence Ellipses; Cost-Benefit Analysis; Cost-Effectiveness Analysis; Cost-Minimization Analysis; Cost-Utility Analysis; Decision Analyses, Common Errors Made in Conducting; Willingness to Pay

Further Readings

Bootman, J. L., Townsend, R. J., & McGhan, W. F. (Eds.). (2004). *Principles of pharmacoeconomics* (3rd ed.). Cincinnati, OH: Harvey Whitney Books.

Briggs, A., Claxton, K., & Sculpher, M. (2006). *Decision modelling for health economic evaluation*. New York: Oxford University Press.

Drummond, M. F., & Jefferson, T. O. (1996). BMJ working party on guidelines for authors and peer-reviews of economic submissions to the BMJ. *British Medical Journal, 313,* 275–283.

Drummond, M. F., & McGuire, A. (Eds.). (2001). *Economic evaluation in health care: Merging theory with practice*. New York: Oxford University Press.

Fry, R. N., Avey, S. G., & Sullivan, S. D. (2003). The Academy of Managed Care Pharmacy format for formulary submissions: An evolving standard—a foundation for managed care pharmacy task force report. *Value in Health, 6*(5), 505–521. Retrieved January 13, 2009, from http://www.amcp.org/amcp.ark?c=pr&sc=link

Glick, H. A., Doshi, J. A., Sonnad, S. S., & Polsky, D. (2007). *Economic evaluation in clinical trials*. New York: Oxford University Press.

Gold, M. R., Siegal, J. E., Russell, L. B., & Weinstein, M. C. (Eds.). (1996). *Cost-effectiveness in health and medicine*. New York: Oxford University Press.

Grauer, D., Lee, J., Odom, T., Osterhaus, J., Sanchez, L., & Touchette, D. (2003). *Pharmacoeconomics and outcomes: Applications for patient care* (2nd ed.). Kansas City, MO: American College of Clinic Pharmacy.

Hunink, M. G. M., & Glasziou, P. P. (Eds.). (2001). *Decision making in health and medicine*. Cambridge, UK: Press Syndicate of the University of Cambridge.

International Society for Pharmacoeconomics and Outcomes Research. (2008). *Comparative table and country-specific guidelines*. Retrieved February 28, 2008, from http://www.ispor.org/PEguidelines/index.asp

Rascati, K. L. (2008). *Essentials of pharmacoeconomics*. Philadelphia: Lippincott Williams & Wilkins.

Rascati, K. L., Drummond, M. F., Annemans, L., & Davey, P. G. (2004). Education in pharmacoeconomics: An international multidisciplinary view. *PharmacoEconomics, 22*(3), 139–147.

Robine, J. M., Jagger, C., Mathers, C. D., Crimmins, E. M., & Suzman, R. M. (2003). *Determining health expectancies*. Hoboken, NJ: Wiley.

Spilker, B. (1996). *Quality of life and pharmacoeconomics in clinical trials*. Philadelphia: Lippincott-Raven.

Tam, T. Y. H. (2004, August). Pharmacoeconomic guidelines around the world. *International Society for Pharmacoeconomics and Outcomes Research (ISPOR) Connections, 10*(4). Retrieved January 13, 2009, from http://www.ispor.org/News/IC2004.aspx

Townsend, R. J. (1987). Postmarketing drug research and development. *Drug Intelligence & Clinical Pharmacy, 21*(1), 134–136.

Yeh, J. Y., Lawson, K. A., Novak, S., Rascati, K. L., Barner, J. C., & Johnsrud, M. (2008). Longitudinal estimates and cost-effectiveness analysis of anti-resorptive agents for glucocorticoid-induced osteoporosis and fractures based on US national surveys. *Value in Health, 11*(3), A259.

Physician Estimates of Prognosis

Physicians are routinely asked to make estimates of patient survival. In providing such estimates, physicians undertake two separate tasks: (1) they *formulate a prognosis*, or make a mental calculation of the patient's expected survival, and (2) they *communicate the prognosis* to the inquiring individual, often a patient or the patient's family. The survival estimates, or prognoses, that physicians formulate and then communicate are important to both physicians and patients in all phases of a patient's life because they guide both medical and nonmedical decisions. At the end of life, these prognoses can become critically important, as they may signal a change from primarily curative or life-prolonging care to primarily supportive or palliative care, a change that clearly influences clinical and personal decisions. Ironically, physician prognostication is often inaccurate, both in terms of the prognoses physicians formulate and in terms of the prognoses physicians communicate to patients or their families.

Importance of Prospective Identification of the End of Life

There is wide agreement among patients, their families, and doctors that the "end of life" is an important period to recognize prospectively because, among other things, the type of medical care that patients receive during this period should be different than that which they receive at other points in their life. Specifically, there is agreement that the medical care should be supportive in nature, focused on the control of symptoms such as pain, rather than invasive in nature, and aimed at extending life. Consistent with this approach, most agree that the favored place of death is the home rather than the hospital. Most physicians report that such home-based, symptom-guided care should be initiated at least 3 months prior to patient death for optimal palliative care.

Despite fairly broad agreement that home-based, symptom-guided care is the preferred form of medical care at the end of life, epidemiologic and health services research reveals that the current patterns of medical care for those dying in America are far from this ideal. For example, a study of Medicare claims data (an excellent population-level source of medical treatment and survival data for elderly Americans) shows that about half of all Medicare beneficiaries die in acute-care hospitals rather than in their homes. Furthermore, fewer than 20% receive hospice care, the most common route to home-based, symptom-guided therapy, prior to death. Finally, of the few who receive this idealized form of medical care at the end of life, most receive it for a period far shorter than the idealized 3 months, generally less than 1 month prior to death. The same work reports that fewer than 15% of Medicare beneficiaries enrolled in hospice programs survive longer than the allotted 6 months.

Inaccuracy of the Formulated Prognosis

While physician prognostication is largely an understudied aspect of clinical medicine, there are studies in the palliative care literature and in the clinical oncology literature that suggest physicians are generally inaccurate in estimating patient survival (i.e., prognosis). Specifically, in the palliative care literature, there are several studies specifically designed to determine the quality of physicians' formulated prognoses in patients with advanced illness. These studies report quality in the form of physicians' prognostic accuracy in predicting survival of patients following admission to hospice programs. Investigators in these studies have measured physicians' prognostic accuracy by comparing patients' observed survival with their predicted survival (these predictions are not necessarily those communicated to patients; rather, they are the ones physicians formulate for themselves). Results of these studies show that, in aggregate, physicians' overall survival estimates tend to be incorrect by a factor of approximately 3, always in the optimistic direction. A representative study documents that physicians overestimate patient survival by a factor of 5 and patients, on average, live only 24 days in hospice.

In the clinical oncology literature, there are studies of physicians' prognostic accuracy in ambulatory cancer patients undergoing chemotherapy. In one such study, investigators asked oncologists to first predict patients' likelihood of cure and then to estimate the duration of survival for those

whose likelihood of cure was 0. At the 5-year point, patients who were alive and disease-free were termed "cured"; the dates of death of the incurable patients also were determined. The researchers reported that oncologists were highly accurate in predicting cure. That is, for subgroups of patients (i.e., not individual patients), the ratio of the observed cure rate at 5 years to the predicted cure rate was quite high: .92. However, the same oncologists had difficulty predicting the length of survival of individual incurable patients. They predicted survival "correctly" for only one third of patients, with the errors divided almost equally between optimistic and pessimistic.

Improving the Accuracy of Formulated Prognoses

As noted, prognostication is an understudied aspect of clinical medicine; this fact may explain part of the difficulty physicians have in predicting their patients' survival. The predictive algorithms that are so common and useful in the narrower organ-system-based aspects of clinical medicine (e.g., Goldman criteria, TNM cancer staging system, Glasgow coma score) have few parallels in the broader clinical area of "end-of-life care." In fact, the current Medicare and National Hospice Organization guidelines for hospice eligibility for patients with certain highly prevalent noncancer diagnoses (i.e., dementia or advanced lung, heart, or liver disease) have been shown to be inadequate for discerning which patients with these conditions have less than 6 months to live. At present, no such formal guidelines exist for the terminal illness of cancer.

However, within palliative oncology research, there is a growing literature focused on identifying predictors of survival of advanced-cancer patients that might aid physicians in their prognostic estimates for similar patients. Multiple prospective and retrospective cohort studies have consistently identified three broad classes of survival predictors: (1) patients' performance status, (2) patients' clinical signs and symptoms, and (3) physicians' clinical predictions. Research that integrates these, and other, prognostically relevant domains through survival models to yield easy metrics for clinicians may help attenuate the problem of prognostic inaccuracy in cancer.

Performance Status

Performance status is a global measure of a patient's functional capacity and has consistently been found to predict survival in cancer patients. Given the importance of survival, performance status is frequently used as a selection criterion for patients entering clinical trials and also as an adjustment factor in the subsequent analyses of treatment effect. Several different measures have been developed to quantify performance status; among them, the Karnofsky Performance Status (KPS) is the most often used. The KPS ranges from values of 100, signifying fully normal functional status with no complaints or evidence of disease, to 0, signifying death. Table 1 contains a representation of the complete spectrum of values for the KPS scale.

Multiple studies have reported associations between cancer patients' survival and their performance status. The direction of the association is positive; that is, as a patient's performance status declines, so too does his or her survival. The magnitude of the association is described differently in different studies depending on the statistical methods employed, but several studies report that among patients enrolled in palliative-care programs, a KPS of less than 50% suggests a life expectancy of less than 8 weeks.

Signs and Symptoms

Clinical signs and symptoms have also been shown to be associated with survival in the setting of advanced cancer. Several investigative groups have examined the prognostic importance of patients' symptoms; Antonio Vigano and colleagues have described this importance in their systematic review of prognostic factors in advanced cancer. In examining 136 different variables from 22 studies, they found that, after performance status, specific signs and symptoms were the next best predictors of patient survival. The presence of dyspnea, dyphagia, weight loss, xerostomia, anorexia, and cognitive impairment provided the most compelling evidence for independent association with patient survival in these studies.

Several groups of investigators have evaluated associations between biological markers (i.e., laboratory values) and survival in advanced cancer patients. For example, in their retrospective analysis

Table 1 Karnofsky performance status scale

Value	*Level of Functional Capacity*
100	Normal, no complaints, no evidence of disease
90	Able to carry on normal activity, minor signs or symptoms of disease
80	Normal activity with effort, some signs or symptoms of disease
70	Cares for self, unable to carry on normal activity or do active work
60	Requires occasional assistance but is able to care for most needs
50	Requires considerable assistance and frequent medical care
40	Disabled, requires special care and assistance
30	Severely disabled, hospitalization is indicated although death is not imminent
20	Hospitalization is necessary, very sick, active supportive treatment necessary
10	Moribund, fatal processes progressing rapidly
0	Dead

Source: Zubrod, G. C., Schneiderman, M., Frei, E., Brindley, C., Gold, G. L., Shnider, B., et al. (1960). Appraisal of methods for the study of chemotherapy in man: Comparative therapeutic trial of nitrogen and mustard and triethylene thiophosphoramide. *Journal of Chronic Diseases*, 11, 7–33.

of 339 Phase I chemotherapy patients with advanced cancer at the University of Chicago, Linda Janisch and colleagues found that among routine pretreatment laboratories, only platelet count elevation and serum albumin depression were associated with shorter survivals in a multivariate model that included KPS. Among a sample of 207 consecutive advanced non-small-cell lung patients, M. F. Muers and colleagues found that in addition to performance status and symptoms, lymphocyte count, albumin, sodium, and alkaline phosphatase were all predictive of survival. From these studies, one can conclude that there appear to be negative associations between survival and bone marrow parameters (e.g., platelets, white blood cells) as well as positive associations between survival and synthetic parameters (e.g., serum proteins) in this patient population.

Clinical Predictions

Physicians' clinical predictions about patient survival are the third broad class of predictors of survival in terminal cancer. As noted previously, numerous studies suggest that physicians' predictions regarding patients' survival in palliative-care programs are frequently inaccurate and systematically optimistic. However, the overly optimistic estimates are well correlated with actual survival. While physicians are not well calibrated with respect to survival (i.e., they are systematically optimistic), they nevertheless have discriminatory abilities. That is, they are able to order patients in terms of how sick they are or how long they have to live. This fact suggests that physicians' clinical predictions may be a useful, but not exclusive, source of information regarding patient survival. Thus, integration of clinical predictions with other known prognostic factors may be beneficial in predicting patient survival. For example, William Knaus and colleagues, in their study of SUPPORT patients, found that multivariate regression models that included physicians' prognostic estimates were more accurate than the models without the physician input. Therefore, while it is true that statistical models can be more accurate than human intuition alone, it is also true that physicians provide valuable prognostic information

that, thus far, has not been captured in the objective models. Currently, integrated models hold the greatest promise for improving physicians' predictive accuracy in advanced-cancer patients.

Integrated Tools

Through integrated models of survival, investigators seek to explicitly combine these previously identified clinical predictors to yield easy-to-use clinical tools. The most recent generation of studies describe integrated models that combine these and other prognostic variables into a single prognostic score. For example, Tatsuya Morita and colleagues developed a regression model predicting survival from performance status and certain clinical signs and symptoms. Coefficients from the regression were then transformed into partial scores, and summing the values of each partial score led to the final score, termed the Palliative Prognostic Index (PPI). After developing the PPI in a sample of 150 patients, the investigators then tested the approach on a second sample of 95 patients, finding that the PPI predicted 3-week survival with sensitivity of 83% and a specificity of 85% and 6-week survival with sensitivity of 79% and a specificity of 77%. Table 2 contains a description of the PPI scoring system and Table 3 a summary of the predictive relevance of PPI scores. Several other groups have developed similar scoring systems that rely on integration of all or some of the previously described classes of prognostic indicators for patients with advanced cancer and under palliative care. Such scoring systems need to be sensitive to a variety of methodological concerns. The most recent generation of studies in this area seek to determine if these scoring systems are useful in the clinical care of cancer patients and if they are applicable to patients who are not yet enrolled in palliative-care programs or who are dissimilar from such patients. With respect to the clinical utility of the scoring systems, treating physicians will need to determine if the tools' test characteristics (e.g., sensitivity and specificity) fall above certain minimum thresholds for use in clinical decisions.

Table 2 Components of the Palliative Prognostic Index: A scoring system for survival prediction of terminally ill cancer patients

Prognostic Domains	*Partial Score Value*
Performance status	
10–20	4.0
30–50	2.5
≥60	0
Clinical symptoms	
Oral intake	
Moderately reduced	1.0
Severely reduced	2.5
Normal	0
Edema	1.0
Dyspnea at rest	3.5
Delirium	4.0

Source: Morita, T., Tsunoda, J., Inoue, S., & Chihara, S. (1999). The Palliative Prognostic Index: A scoring system for survival prediction of terminally ill cancer patients. *Supportive Care in Cancer, 7,* 128–133.

Note: Scores from each prognostic domain are summed, and the total is mapped to survival probability.

Other Sources of Prognostic Information

Among other sources of information regarding survival in advanced cancer are studies that include cancer patients who do not undergo anticancer therapy. Natural history studies and randomized therapy trials that include a "best-ssupportive-care" arm describe patients who do not undergo anticancer therapy. Typically, natural history studies are single-institution case series of untreated patients with mortality follow-up. Such reports have been published for a variety of advanced solid tumors. Survival information can also be found by examining the survival of patients on the control or best-supportive-care arms of randomized clinical trials.

Implications

Physicians are usually inaccurate in formulating patient prognoses, and the direction of their prognostic error is largely optimistic, with patients dying much sooner than their physicians anticipate.

Table 3 Median survival of patients according to Palliative Prognostic Index score

Palliative Prognostic Index score	*Median survival (d)*
0.0–2.0	90
2.1–4.0	61
>4.0	12

Source: Morita, T., Tsunoda, J., Inoue, S., & Chihara, S. (1999). The Palliative Prognostic Index: A scoring system for survival prediction of terminally ill cancer patients. *Supportive Care in Cancer, 7,* 128–133.

Note: Median survival value was estimated from survival curve on paper.

As such, initiation of supportive care at the end of patients' lives may be unwittingly delayed in favor of continued "aggressive" medical care. Physicians' prognoses may become more accurate as better clinical prediction tools are developed and disseminated. Ultimately, such improvement would expect to manifest through increasing rates of referral to palliative-care programs and increased survival times after referral to the same programs. More broadly, however, such improvement may provide patients with a better understanding of their expected survival and thereby allow them to make informed medical and social choices regarding their treatment path at the end of life, whether it is curative or palliative.

Elizabeth B. Lamont

See also Advance Directives and End-of-Life Decision Making; Bias; Decision Making in Advanced Disease; Judgment; Life Expectancy; Probability Errors

Further Readings

Janisch, L., Mick, R., Schilsky, R. L., Vogelzang, N. J., O'Brien, S., Kut, M., et al. (1994). Prognostic factors for survival in patients treated in phase I clinical trials. *Cancer, 74,* 1965–1973.

Knaus, W. A., Harrell, F. E., Lynn, J., Goldman, L., Phillips, R. S., Connors, A. F., Jr., et al. (1995). The SUPPORT prognostic model: Objective estimates of survival for seriously ill hospitalized adults. *Annals of Internal Medicine, 122,* 191–203.

Morita, T., Tsunoda, J., Inoue, S., & Chihara, S. (1999). The Palliative Prognostic Index: A scoring system for survival prediction of terminally ill cancer patients. *Supportive Care in Cancer, 7,* 128–133.

Muers, M. F., Shevlin, P., & Brown, J. (1996). Prognosis in lung cancer: Physicians' opinions compared with outcome and a predictive model. *Thorax, 51,* 894–902.

Vigano, A., Dorgan, M., Buckingham, J., Bruera, E., & Suarez-Almazor, M. E. (2000). Survival prediction in terminal cancer patients: A systematic review of the medical literature. *Palliative Medicine, 14,* 363–374.

Zubrod, G. C., Schneiderman, M., Frei, E., Brindley, C., Gold, G. L., Shnider, B., et al. (1960). Appraisal of methods for the study of chemotherapy in man: Comparative therapeutic trial of nitrogen and mustard and triethylene thiophosphoramide. *Journal of Chronic Diseases, 11,* 7–33.

Poisson and Negative Binomial Regression

The response variable in medical data is often in the form of counts. Examples include visits to the doctor, cases of stroke or heart attacks, and number of deaths due to various causes. Two common distributions used to model counts that arise in situations such as these are the Poisson and negative binomial distributions. In this entry, the important aspects of Poisson and negative binomial regression are covered along with an example to illustrate basic inference for these models.

Perhaps the most common realization of Poisson data is that of "rare-event" data, which are events that occur relatively few times in a large population. In this case, the Poisson distribution is seen as a limiting form of the binomial distribution when the sample size, n, grows large and the probability of an event occurring, π, grows small. Generally, this assumption holds reasonably well for $n > 20$ and $\pi < .1$. In medical and epidemiological literature, an example of this would be the number of cancer deaths in an at-risk group.

A second and obviously related realization of Poisson data is that of discrete count events in time or space. Events are typically considered as "arrivals" or discretized points over a continuous domain;

for example, it may be of interest to count the number of doctor visits for an individual or family over a period of time. Furthermore, one may count the number of white blood cells per unit volume of a blood culture.

The probability that a Poisson random variable Y takes the observed value y is expressed in the probability mass function

$$\Pr(Y=y)=\frac{e^{-\lambda}\lambda^{y}}{y!},\quad \lambda>0,\ y=0,1,2,\ldots.$$

In this parameterization, λ is described as an intensity or rate parameter and is often interpreted as the expected number of events in the rare-events paradigm or as the rate of events per unit time/space in the spatiotemporal paradigm. The value e is Euler's constant, and the denominator $y!$ is the factorial function performed on the integer y, where $y! = (y)(y-1)\ldots(2)(1)$. Thus, for gamma-distributed Poisson with $\lambda = 3$, $\Pr(Y=2)=e^{-3}3^2/2!=.22$ Typically, it is of interest to make some inference about the value of the unknown parameter λ. When subjects are followed over varying periods of time, the distribution is often parameterized as $\mu = d\lambda$, where d is the amount of follow-up time, and is often referred to as the offset.

Poisson Regression Model

Poisson regression is one example of a broader class of models known as the generalized linear model. The generalized linear model includes ordinary least squares regression with normal errors, logistic regression, beta regression, and others. For Poisson regression, it is assumed that the value of the mean depends on a function of an observed vector of covariates, $\mathbf{x}_i' = (x_1, x_2, \ldots, x_p)$, and model parameters, $\beta' = (\beta_0, \beta_1, \ldots, \beta_p)$. Since the Poisson rate, λ_i, is strictly nonnegative, the expected number of events is usually modeled as

$$E[y_i \mid \mathbf{x}_i] = \lambda_i = \exp(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip}).$$

If there is an offset, the mean is modeled as

$$E[y_i \mid \mathbf{x}_i] = \mu_i = d_i\lambda_i = \exp(\mathrm{In}(d_i) + \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip}).$$

In the generalized linear model terminology, the exponential function connecting the expected value and the covariates is referred to the log link because if the log of the mean function is taken, a linear combination of the regression parameters results.

Poisson regression coefficients have a significantly different interpretation from the coefficients in linear regression. The derivative of the mean function with respect to the jth covariate, x_j, is

$$\frac{\partial E[y|\mathbf{x}]}{\partial x_j}=\beta_j E[y|\mathbf{x}].$$

Thus, the impact of a one-unit change in x_j, holding all other independent variables constant, results in a multiplicative change in the expectation as opposed to a linear change.

In the case of a single binary predictor, it is interesting to note that

$$\frac{E[y|x=1]}{E[y|x=0]}=\frac{\exp(\beta_0+\beta_1\times 1)}{\exp(\beta_0+\beta_1\times 0)}=\exp(\beta_1).$$

Thus, in the important two-group case, the coefficient has the interpretation that the group with $x = 1$ has an expected value that is $\exp(\beta_1)$ times larger than the group with $x = 0$.

Estimation in Poisson Regression

The primary method for estimation of Poisson regression parameters is maximum likelihood estimation (MLE). For a random sample of n observations, the likelihood function for λ is

$$L(\lambda|\mathbf{y})=\prod_{i=1}^{n}\frac{e^{-\lambda}\lambda^{y_i}}{y_i!}=\frac{e^{-n\lambda}\lambda^{\sum_i^n{}_1 y_i}}{\prod_{i=1}^{n} y_i!}.$$

On substituting $\lambda_i = \exp(\mathbf{x}_i'\beta)$ into the above, the likelihood of β given the observed data is

$$L(\beta|\mathbf{y},\mathbf{x})=\frac{\exp\left(-\exp(\mathbf{x}_i'\beta)+\sum_{i=1}^{n} y_i\mathbf{x}_i'\beta\right)}{\prod_{i=1}^{n} y_i!}.$$

The method of maximum likelihood is the default approach to estimating the Poisson regression parameters in many software packages such as SAS and S-Plus. The maximum likelihood estimator is found by maximizing the log likelihood:

$$\ln L(\beta|\mathbf{y},\mathbf{x}) = \sum_{i=1}^{n} (y_i \mathbf{x}_i'\beta - \exp(\mathbf{x}_i'\beta) - \ln(y_i!). \qquad (1)$$

The problem of determining how the population mean λ_i covaries with each element of $\mathbf{x}$ can be simplified into solving for each element of vector β. The resulting estimate vector, $\hat{\beta}$, does not have a closed-form solution and is typically computed via an iterative algorithm, such as the Newton-Raphson method, in most statistical computing packages. Other approaches for estimating the parameters include Bayesian methods and the generalized method of moments, also known as generalized estimating equations.

The maximum likelihood estimator, $\hat{\beta}$, that results from maximizing Equation 1 has many nice properties. As the sample size increases, $\hat{\beta}$, is unbiased, consistent, efficient, and normally distributed with variance approximated by

$$\left[-E\left(\frac{\partial^2 \ln L(\beta|\mathbf{y},\mathbf{x})}{\partial\beta\partial\beta'} \right) \right]^{1}.$$

Using the properties of maximum likelihood estimators, hypothesis tests can be constructed along with confidence intervals for parameters and prediction intervals for future observations. These computations can be performed with a wide variety of commonly used statistical software, such as SAS, S-Plus, and SPSS. Also, goodness of fit can be assessed through various statistics based on the residuals $y_i - \hat{\mu}_i$ where $\hat{\mu}_i$ is the mean function of y_i evaluated at the MLE.

Negative Binomial Regression

A unique property of the Poisson distribution is that the mean and the variance both equal λ. This property is often referred to as *equidispersion* of the Poisson random variable. What this implies is that

$$\mathrm{Var}[y_i \,|\, \mathbf{x}_i] = \lambda_i = \exp(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip}).$$

Generally, when this assumption fails, the data exhibit *overdispersion*; that is, the variance exceeds the mean. In this case, interestingly, the Poisson regression estimators are still consistent, but standard errors will be underestimated, which has a negative impact on the coverage probabilities of confidence intervals and Type I error of hypothesis tests.

When overdispersion is evident in the data, the negative binomial distribution is a convenient way to model the data. The negative binomial results as a gamma-distributed mixture of Poisson distributions and has the form

$$\Pr(Y = y) = \frac{\Gamma(1/\alpha + y)}{y!\Gamma(1/\alpha)} \left(\frac{1/\alpha}{1/\alpha + \lambda} \right)^{1/\alpha} \left(\frac{\lambda}{1/\alpha + \lambda} \right)^{y},$$
$$\lambda, \alpha > 0,\ y = 0, 1, 2, \ldots.$$

The expected value for this particular parameterization of the negative binomial is λ, but unlike the Poisson, the variance is $\lambda + \alpha\lambda^2$. This form of the negative binomial is sometimes referred to as the NB2 model because of the squared λ term in the variance. In general, the negative binomial can be parameterized such that the variance is of the form $\lambda + \alpha\lambda^k$. Choosing $k = 1$ yields what is known as the NB1 model, which is also frequently used. Thus, the parameter α works as a dispersion parameter. As α decreases, the negative binomial approaches the Poisson. Overdispersion in the data will yield larger values of α. To incorporate covariates, the same log-linear model as for the Poisson is used; specifically,

$$E[y_i \,|\, \mathbf{x}_i] = \lambda_i = \exp(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip}).$$

The coefficients for the negative binomial regression have similar interpretations to the coefficients discussed for the Poisson regression above.

Estimation in Negative Binomial Regression

As in Poisson regression, the most commonly used method to estimate the parameters is maximum likelihood estimation. The log likelihood function is

$$\ln L(\beta, \alpha|\mathbf{y},\mathbf{x}) = \sum_{i=1}^{n} (y_i \mathbf{x}_i'\beta + y_i \ln(\alpha) - \ln(1 + \alpha \exp(\mathbf{x}_i'\beta)) + \ln \Gamma(y_i + 1/\alpha) - \ln \Gamma(1/\alpha) - \ln(y_i!)).$$

Again, an iterative algorithm such as the Newton-Raphson method is used to find the solution of the MLE on differentiating the log likelihood with respect to the parameters and setting the resulting system of equations equal to 0.

The MLEs for the parameters of the negative binomial regression have the same nice properties as those of Poisson regression, which provide ways to form confidence intervals and construct hypothesis tests.

Testing Poisson Versus Negative Binomial

An important first step in a count data regression is to determine if the Poisson model adequately fits the data or if overdispersion is present and the negative binomial is preferred. Since the Poisson model is "nested" within the negative binomial, the likelihood ratio test is a reasonable approach to determine which model fits better. The Poisson is nested within the negative binomial since, as discussed above, if $\alpha = 0$, the negative binomial reduces to the Poisson. To construct the likelihood ratio, both models are fitted from the data, and the log likelihoods are computed. SAS, S-Plus, and other software packages give these values as part of the output. The likelihood ratio test statistic is computed by taking the negative of two times the difference in the log likelihoods, which under the null hypothesis of model equivalence has an approximate chi-square distribution with 1 degree of freedom (*df*). Additionally, most software packages estimate at least one "deviance" statistic that gives a rough measure of dispersion within a Poisson model. When the deviance statistic divided by the number of degrees of freedom for the model is approximately 1.0, the equidispersion assumption holds reasonably well. However, when the deviance divided by the degrees of freedom exceeds 1.0, the data are likely overdispersed, whereas if the statistic is less than 1.0, the data are likely underdispersed.

Example

An example is now presented, originally published by LaVange et al., using a subset of the original variables analyzed. The data are from a study that seeks to determine the relationship between the number of annual lower-respiratory infections among infants and the covariates of exposure to secondhand smoking (Smoke) and socioeconomic status (SES; 0 = *High*, 1 = *Middle*, 2 = *Low*). Furthermore, the participants are followed for varying lengths of time, and thus the time at risk for each individual, d_i, must be considered as the offset. It is of interest to determine which covariates predict significant changes in disease incidence. To compute the estimates for this example, SAS Version 9.1.3 was used. For more information on this data set, see *Categorical Data Analysis Using the SAS System* by Maura Stokes et al.

First, the Poisson regression model is fit. Although the Poisson case is a simplified version of the negative binomial distribution, the Poisson model is first fit since it is the most parsimonious model, and then it is determined whether overdispersion exists. On fitting the model, the parameter estimates displayed in Table 1 obtain.

These estimates remain in a log-transformed scale, and thus the parameter estimates are exponentiated to obtain the estimated impact of the covariates. For example, the coefficient of .4357 for the variable "Smoke" indicates that holding other variables the same, infants who have been exposed to secondhand smoke have exp(.4357) = 1.55 times the incidence of annual lower-respiratory infections as do infants who have not been exposed. Likewise, negative coefficient estimates indicate a protective effect with regard to the number of respiratory infections. Infants from families with high socioeconomic status (SES = 0) have exp(−.6137) = .54 times, or approximately half, the annual incidence of lower-respiratory infection when compared with the low-SES group. If interest lies in the expected number of events for certain values of the covariates, simply plug in the covariates to the regression equation. For example, the predicted mean number of lower-respiratory infections for an infant with SES = 1 who has not been exposed to passive smoking (Smoke = 0) is

$$\begin{aligned}\hat{\mu}_i &= \exp\left(\hat{\beta}_0 + x_{\text{smoke},i}\hat{\beta}_{\text{smoke}} + x_{\text{SES},i}\hat{\beta}_{\text{SES}}\right)\\ &= \exp(.127 + 0 - .105) = 1.02\end{aligned}$$

infections per year.

The computer output includes statistics measuring the properties of the Poisson model, including deviance and log likelihood. These are presented in Table 1.

The value of the log likelihood is included for later use. The value of the deviance divided by the degrees of freedom exceeds 1.0 by a reasonably large value. Although this is not a formal test,

Table 1 Model parameter estimates

		Poisson Model		*Negative Binomial Model*	
Parameter		*Estimate*	*Standard Error*	*Estimate*	*Standard Error*
Intercept		.1265	.2058	.1876	.2846
Smoke		.4357	.1644	.4588	.2178
SES[a]	0	–.6137	.2036	–.6388	.2837
SES	1	–.1053	.1954	–.1599	.2847
SES	2	.0000	.0000	.0000	.0000
Dispersion		—	—	1.0933	.2758

a. SES, socioeconomic status.

overdispersion likely exists within the data, and this situation possibly warrants adopting a more flexible model, such as assuming negative binomially distributed responses.

The data are next fit using a negative binomial model that includes an estimate of the dispersion parameter α from the NB2 parameterization, labeled as "Dispersion" in Table 1. The parameter estimates for the negative binomial model are quite similar to those of the Poisson model. However, the standard errors of the parameter estimates are uniformly larger than those of the Poisson model. This suggests that the Poisson model may be a poor fit, as it appears that the Poisson model underestimates standard errors when compared with the more general case of the negative binomial distribution.

A dispersion parameter estimate near 0 indicates that using a negative binomial model is unnecessary given the current data. However, the observed estimate is 1.09 with a small standard error estimate, which indicates a significant contribution by the dispersion parameter. More formally, one may test the Poisson model versus the negative binomial model via the likelihood ratio test statistic $\chi_1^2 = -2((-267.51) - (246.55)) = 41.90$. Note that the value of the rejection region for a chi-square statistic with a single degree of freedom is $\chi_1^2 > 3.84$ for a typical Type I error rate of .05, and thus it is necessary to reject the simpler Poisson model in favor of the negative binomial model in order to account for the overdispersion of the data.

James D. Stamey and Daniel Beavers

See also Distributions: Overview; Logistic Regression; Maximum Likelihood Estimation Methods; Ordinary Least Squares Regression

Further Readings

Allison, P. (1999). *Logistic regression using the SAS system*. Cary, NC: SAS Publishing.

Cameron, A. C., & Travedi, P. K. (1998). *Regression analysis of count data*. Cambridge, UK: Cambridge University Press.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2004). *Bayesian data analysis* (2nd ed.). Boca Raton, FL: Chapman & Hall.

Hardin, J. W., & Hilbe, J. M. (2007). *Generalized linear models and extensions*. College Station, TX: STATA Press.

LaVange, L. M., Keyes, L. L., Koch, G. G., & Margolis, P. E. (1994). Application of sample survey methods for modeling ratios to incidence densities. *Statistics in Medicine, 13,* 343–355.

Stokes, M. E., Davis, C. S., & Koch, G. G. (2000). *Categorical data analysis using the SAS system* (2nd ed.). Cary, NC: SAS Institute.

Positivity Criterion and Cutoff Values

Although medical professionals use the terms *positive* and *negative* to describe the results of diagnostic tests, many tests produce results along a continuum (e.g., millimeters of ST-segment depression in an exercise stress test, brain natriuretic peptide level for making the diagnosis of decompensated congestive heart failure). For such tests, a criterion must be established for defining a result as being either positive or negative. This cut-point, or *cutoff value*, is called a *positivity criterion*.

Selection

Although for many tests, positivity criteria have been selected based on the variation observed in a population of apparently normal individuals (e.g., mean ±2 standard deviations) encompassing 95% of the population, such a definition may not be optimal for clinical purposes. Ideally, the choice of a positivity criterion should consider the following: (a) the medical consequences of false-positive and false-negative test results, (b) the prevalence of disease in the population being tested, and (c) the distribution of test results in patients with and without disease.

A Clinical Example

A clinical example that illustrates the above principles is the tuberculin skin test, used to determine whether individuals have been exposed to tuberculosis (TB) and developed so-called latent disease, which would necessitate medical treatment. Tuberculin skin testing is performed by injecting a small amount of purified tuberculin extract under the skin. The test is read 48 hours later by seeing if redness and swelling (also called induration) develops and, if so, how large an area of induration. Figure 1 shows a hypothetical distribution of results in which the horizontal axis represents the amount of induration in millimeters (mm). The top distribution describes test results in a population without latent TB, while the bottom distribution describes results for patients with latent TB. Each vertical line represents a different potential positivity criterion. For any criterion, all the patients to the left of the line are deemed to have a negative test result, and those to the right of the line are deemed to have a positive result. The line representing each cutoff divides the distributions into four quadrants. In the top distribution describing patients without disease, those to the left of the cutoff who have a negative test result are the true negatives (TN), while those to the right who have a positive result are false positives (FP). In the bottom distribution describing those with disease, those to the right who have a positive result are true positives (TP), while those to the left are false negatives (FN). As the criterion moves to the right (from A to B to C), the proportion of patients with true-negative test results increases, while the proportion with true-positive results decreases. Since the area under each distribution is unity, the true-negative area corresponds to the *specificity*, while the true-positive area corresponds to the *sensitivity*. Thus, moving to the right in

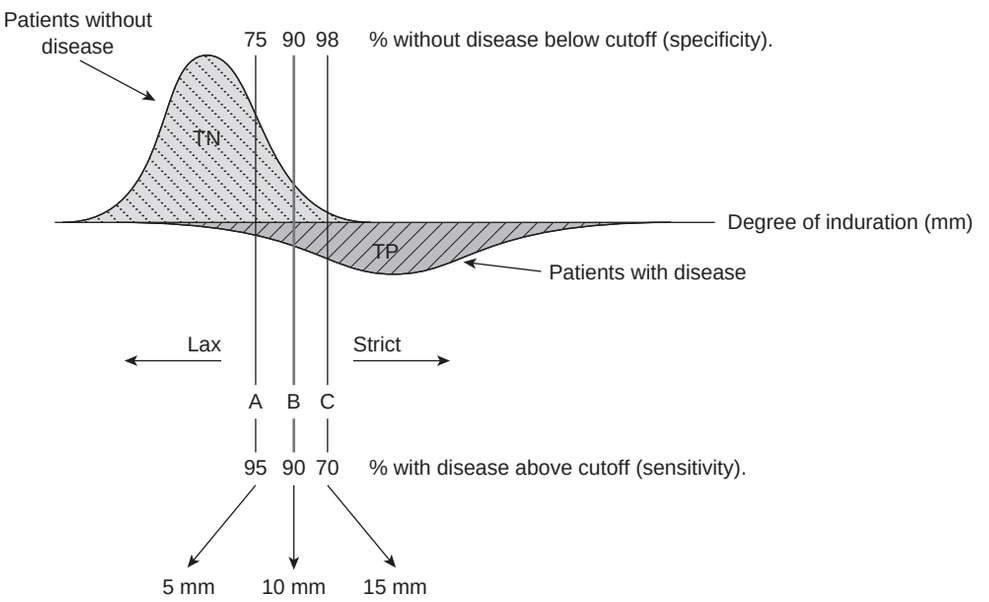


Figure 1 Distribution of test results in those with and without disease

the figure, specificity increases while sensitivity decreases. The American Thoracic Society has updated the guidelines for the interpretation of TB skin tests to account for the principles described above. The old guidelines simply used 10 mm of induration as the cutoff for all patients. The new guidelines are as follows:

1. In high-risk, high-prevalence populations (e.g., HIV infection, immunosuppressed, recent close contact with active TB), a cutoff of 5 mm (Line A) is recommended.
2. In low-prevalence, low-risk populations, a cutoff of 15 mm (Line C) is recommended.
3. In populations at intermediate risk and prevalence (e.g., patient born in a foreign country with increased prevalence, nursing home resident), a cutoff of 10 mm (Line B) is recommended.

Receiver Operating Characteristic (ROC) Curves

This relationship between sensitivity and specificity is described by the *receiver operating characteristic* or *ROC curve* (see Figure 2), which plots the true-positive rate, or sensitivity, on the vertical axis against the false-positive rate, or (1 – specificity), on the horizontal axis. Because the distributions of test results in those with and without disease overlap to some degree, any change in the positivity criterion that improves sensitivity must invariably make specificity worse (i.e., increase the false-positive rate). Also, note that each positivity criterion in Figure 1 at Points A, B, and C corresponds to a point on the ROC curve in Figure 2 called the *operating point*. The operating point on the ROC curve therefore denotes a set of operating characteristics for the test (i.e., a unique combination of sensitivity and specificity). As we move the cutoff to the right and select a stricter interpretation of the test result (Figure 1), we correspondingly move along the ROC curve (Figure 2) toward the lower left. In a similar manner, as we move the cutoff to the left and select a more lax interpretation of the test result, we move along the ROC curve toward the upper right. In this hypothetical example, the optimal operating point, shown at Point A, is based on the results of the decision model described as follows.

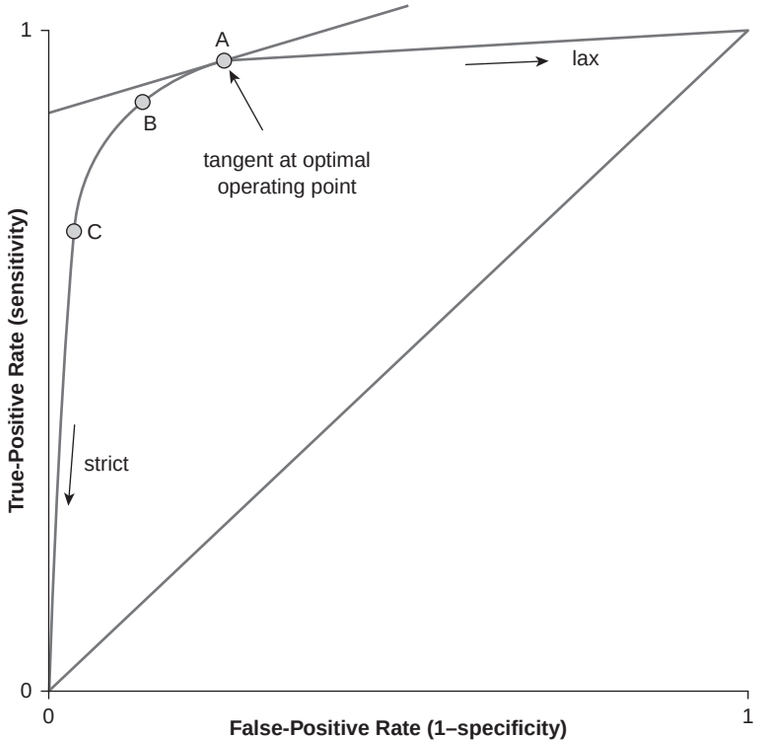


Figure 2 Corresponding receiver operating characteristic (ROC) curve

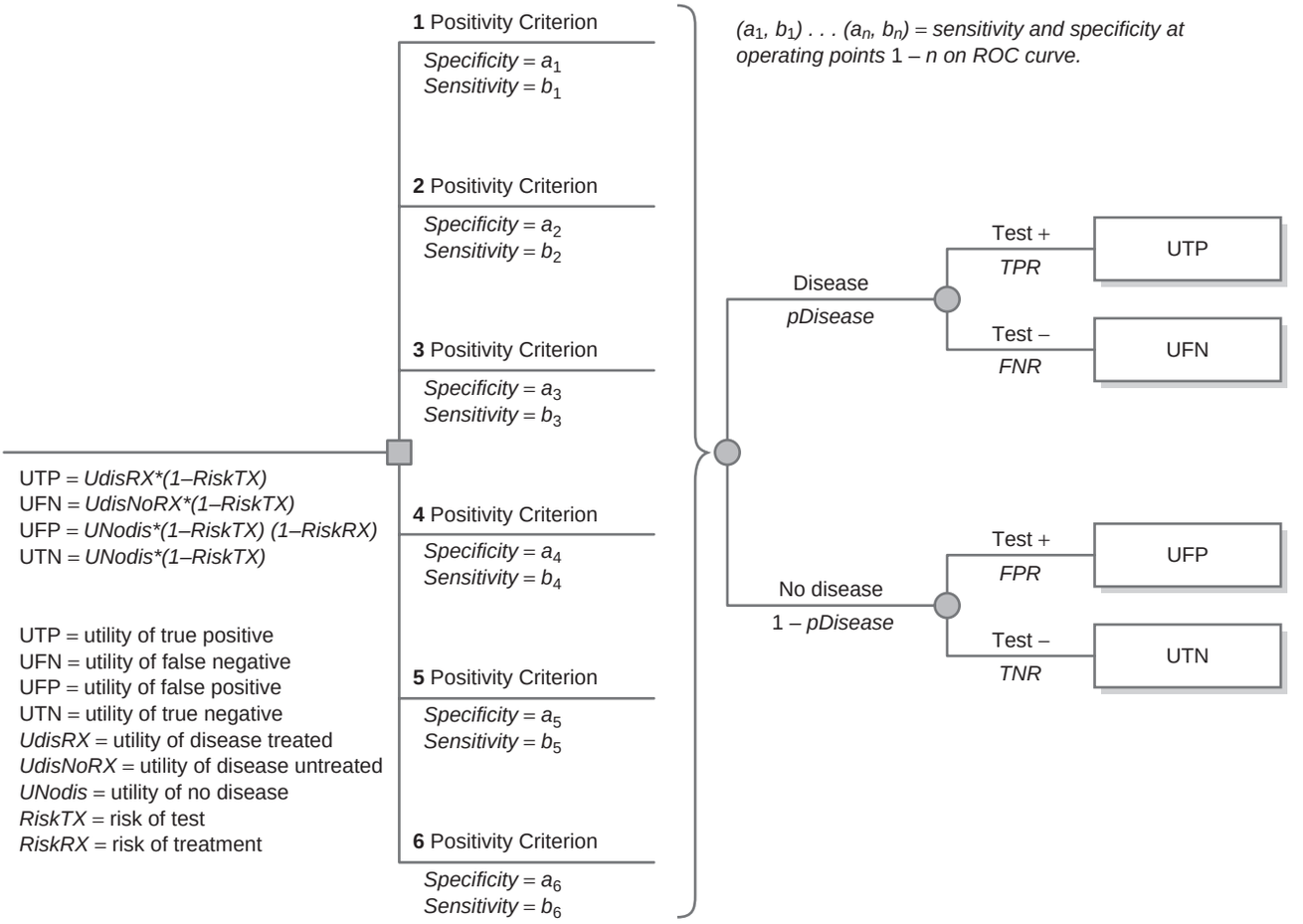


Figure 3 Decision fan to determine the optimal operating point on the ROC curve

Decision Analysis

The following section assumes a basic familiarity with decision analysis and decision trees. The selection of an appropriate positivity criterion along the continuum of results for a test has as its corollary in ROC space the selection of the *optimal operating point* on the ROC curve. Therefore, the optimal operating point depends on the costs or consequences of incorrect diagnoses as well as the prevalence of disease and can be considered formally through the process of decision analysis. Figure 3 shows a decision fan that has as each branch at the decision node a different positivity criterion and operating point. In other words, the only difference between each of these decisions is the combination of sensitivity and specificity from points along the ROC curve. Each of the four outcomes of a dichotomous test has an assigned value or utility (e.g., utility of a true negative, or UTN). The optimal operating point would correspond to the branch with the greatest expected utility. This also can be represented algebraically such that the optimal criterion corresponds to the point where the instantaneous slope of the ROC curve equals

$$\frac{\text{Probability of disease being absent}}{\text{Probability of disease being absent}} \times \frac{\text{Cost of false positives}}{\text{Cost of false negatives}},$$

where (Probability of disease being absent/Probability of disease being present) is the odds of disease being absent and where the cost of a false

positive (in Figure 3) is UTN – UFP (utility of a false positive) and the cost of a false negative (in Figure 3) is UTP – UFN.

In summary, tests are used to discriminate between patients with and without disease. If results are interpreted in a dichotomous manner, then a cutoff must be selected to differentiate between a positive and a negative test result. The selection of this cutoff should depend on the balance of the consequences between false-positive and false-negative results, along with the prevalence of disease, and is also represented by the optimal operating point on the ROC curve.

Mark H. Eckman

See also Decision Trees, Construction; Decision Trees, Evaluation; Diagnostic Tests; Receiver Operating Characteristic (ROC) Curve

Further Readings

Griner, P. F., Mayewski, R. J., Mushlin, A. I., & Greenland, P. (1981). Selection and interpretation of diagnostic tests and procedures. *Annals of Internal Medicine, 94,* 553.

McNeil, B. J., Keller, E., & Adelstein, S. J. (1975). Primer on certain elements of medical decision making. *New England Journal of Medicine, 293*(5), 211–215.

Prediction Rules and Modeling

In this entry, the role of prediction models for medical decision making is discussed. Decision rules can be based on prediction models and are important for more individualized decision making. Prediction models have potential applications in both medical practice and research. Prediction models are ideally derived from large-volume, high-quality empirical data to quantify the relationship between a set of predictors and a diagnostic or prognostic outcome. Model development needs to be followed by model validation and an analysis of the model's impact on decision making and outcomes of individual subjects.

Prediction Models

Clinical prediction models may provide the evidence-based input for shared decision making, by providing estimates of the individual probabilities of risks and benefits. Clinical prediction models are sometimes also referred to as prognostic models or nomograms. Clinical prediction models combine a number of characteristics (e.g., related to the patient, the disease, or treatment) to predict a diagnostic or prognostic outcome. Typically, between 2 and 20 predictors are considered. The number of publications with clinical prediction models has increased steeply in recent years in various medical fields.

Applications of Prediction Models

Prediction models are valuable for medical practice and for research purposes. In public health, models may help target preventive interventions to subjects at relatively high risk of having or developing a disease. In clinical practice, prediction models may inform patients and their treating physicians on the probability of a diagnosis or a prognostic outcome. Prognostic estimates may, for example, be useful to assist in planning of the remaining lifetime in terminal disease or give hope for recovery if a good prognosis is expected after an acute event such as a stroke. Classification of a patient according to his or her risk may also be useful for communication among physicians, for research purposes, and for benchmarking.

Prediction models may also assist medical decision making, for example, as part of a decision support system. In the diagnostic workup, predictions can be useful to estimate the probability that a disease is present. When the probability is relatively high, treatment is indicated; if the probability is very low, no treatment is indicated. For intermediate probabilities of disease, further diagnostic testing is necessary. In therapeutic decision making, treatment should only be given to those most likely to benefit from the treatment. Prognostic predictions may support the weighing of harms versus individual benefits. If risks of a poor outcome are relatively low, the maximum benefit will also be relatively low. Any harm, such as a side effect of treatment, may then readily outweigh any benefits. The claim of prediction models is that

better decisions can be made with a model than without and that their predictions are sometimes better than those made by physicians.

In research, prediction models may assist in the design and analysis of randomized trials. Adjustment for baseline risk in the analysis of a trial results in higher statistical power for the detection of a treatment effect. Models are also useful to control for confounding variables in observational research, either in traditional regression analysis or with modern approaches such as propensity scores.

From Prediction Models to Decision Rules

Prediction models provide diagnostic or prognostic probabilities. They may assist medical decision making without telling clinicians what to do precisely. One motivation for providing probabilities only is that decision thresholds may differ from patient to patient. Some argue, however, that prediction models will more likely have an impact on clinical practice when clear actions are defined in relation to the predictions—that is, in the form of a decision rule (or prediction rule). Prediction models may hereto require simplification to provide clear advice on actions with high and low predictions. A decision threshold has to be defined, chosen either informally or by formal decision analysis. When applying some diagnostic rules, clinicians may not want to miss any patient with the outcome of interest (e.g., Ottawa ankle rules). This implies that clinicians aim for a sensitivity of 100% and hope for reasonable specificity. They accept false-positive classifications, since 100% sensitivity implies an infinite cost of false-negative classifications. Decision rules can often be presented in a simpler format than detailed prediction models that provide individualized predictions.

Modeling

Statistical models for prediction try to relate a set of predictor variables (X) to an outcome (Y). The most common method in medical research is regression analysis. The resulting predictive regression models can be entered relatively easily in decision analytic models, including decision trees and Markov models.

Regression models make a number of assumptions on the relationships between predictors and the outcome, such as additivity of effects (which can be tested by adding interaction terms) and linearity of effects for continuous predictors (which can be tested by adding nonlinear terms). The specific type of regression model to use is guided by the type of outcome. The linear regression model is the default for continuous outcomes. In the context of medical decision making, the outcome is commonly dichotomous (e.g., presence vs. absence of a target diagnosis or occurrence of a prognostic outcome, such as mortality). Logistic regression analysis is the most commonly used statistical technique to predict such dichotomous outcomes. Alternative methods include recursive partitioning, or Classification and Regression Tree (CART) methods, as well as neural networks. These alternative prediction methods usually make less stringent assumptions than regression models but require larger sample sizes. CART methods assume interactions between predictors, and neural networks usually allow for nonadditive and nonlinear effects of predictors. Details are found in many excellent statistical textbooks.

If outcomes are not observed for all subjects in a study ("censoring"), statistical survival models should be used. The Cox regression model is commonly used to predict the probability of occurrence of an outcome by a certain time point. For decision modeling, parametric survival models, such as the Weibull model, may, however, have some advantages, including more stable predictions at the end of follow-up and possibilities of extrapolation beyond observed follow-up time.

Study Design

Prognostic studies are inherently longitudinal in nature, most often performed in cohorts of patients, who are followed over time for an outcome to occur. The cohort is defined by the presence of one or more particular characteristics—for example, having a certain disease, living in a certain place, having a certain age or simply being born alive. For example, researchers may follow a cohort of patients with an acute myocardial infarction for long-term mortality.

Diagnostic studies are most often designed as a cross-sectional study, where predictive patient characteristics are related to an underlying diagnosis. The study group is defined by the presence of a particular symptom or sign suggesting that the

subject may have a particular (target) disease. Typically, subjects undergo the test of interest and, subsequently, a reference test to establish the "true" presence or absence of the target disease over a short time span. For example, clinicians may aim to diagnose those with an acute myocardial infarction among patients presenting at an emergency department.

Predictors and Outcome

Strength of Predictors

For a well-performing prediction model, strong predictors have to be present. Strength is a function of the association of the predictor with the outcome, and the distribution of the predictor. For example, a dichotomous predictor with an odds ratio of 2.0 and 50% prevalence is more relevant for a prediction model than a dichotomous predictor with an odds ratio of 2.5 with 1% prevalence. Also, continuous predictors with a wider range are more relevant for prediction.

When some characteristics are considered as key predictors, these have to be registered carefully, with clear definitions and preferably no missing values. This is usually best possible in a prospective study, with a protocol and prespecified data collection forms.

Reliability of Predictors

Ideally, predictors are well defined and reliably measurable by any observer. In practice, observer variability is a problem for many measurements. In addition, some measurements are prone to biological variability. A well-known example is blood pressure, where a single measurement is quite unreliable. Usually, at least two measurements are obtained, and preferably more, with some spread in time. Most prediction models include predictors that are quite readily available, are not too costly to obtain, and can be measured with reasonable precision.

Choice of Outcome

The outcome of a prediction model should be relevant, either from an applied medical perspective or from a research perspective. From a medical perspective, "hard" end points are generally preferred. Especially, mortality is often used as an end point in prognostic research. Mortality risks are relevant for many acute and chronic conditions and for many treatments, such as surgery. In other diseases, other outcomes may be preferred, including nonfatal events (e.g., disease recurrence), patient-centered outcomes (e.g., scores on quality-of-life questionnaires), or wider indicators of burden of disease (e.g., absence from work). Statistical power may also direct the choice of outcome. The infrequency of an outcome may make an outcome less appropriate for statistical analysis. Ideally, a dichotomous outcome has a 50:50 distribution. Continuous outcomes generally provide more statistical power than categorized or dichotomized outcomes.

The prognostic outcome should be measured as reliably as possible. Prediction models may be developed with pragmatic definitions of predictors, since this may resemble the future use of a model. But the outcome should be determined with similar rigor as in an etiologic study or randomized clinical trial. In the future, decisions are to be based on the predictions from the model. Hence, predictions need to be based on robust statistical associations with an accurately determined outcome.

Steps in Model Development

There are seven logically distinct steps in the development of valid prediction models with regression analysis that researchers may consider. These steps are briefly addressed below, with more detail provided elsewhere.

1. *Problem definition and data inspection:* A preliminary step is to carefully consider the prediction problem: What are the research questions? What is already known about the predictors? The next step is to consider the data under study: How are the predictors defined? What is the outcome of interest? An important issue is that missing values will occur in at least some of the predictors under study. Various statistical approaches are available for dealing with missing values, with multiple imputation being used in the more recent prediction models.

2. *Coding of predictors:* When researchers start on building a prediction model, the first issue is the coding of predictors for a model; several choices need to be considered on categorical variables and

continuous variables. Dichotomization of a continuous predictor has many disadvantages and should be discouraged.

3. *Model specification:* The most thorny issue in prediction modeling is how to specify the model. What predictors should be included, considering what is known about the predictors already, and what is observed in the data under study? Stepwise selection methods are widely used but have many disadvantages, such as instability of the selection, bias in estimated regression coefficients, and underestimation of uncertainty in the selected model. Using subject knowledge for model specification is a better approach—for example, considering previous studies on predictors and prediction models or having discussions with clinical experts. Another issue is how researchers should deal with assumptions in regression models, such as additivity and linearity of predictor effects. Iterative cycles of testing of assumptions and adaptation may lead to a model that provides predictions that do not generalize to new subjects outside the data set under study ("overfitting"). A simple, robust model that may not fit the data perfectly should be preferred to an overly fine-tuned model for the specific data under study.

4. *Model estimation:* Once a model is specified, model parameters need to be estimated. For regression models with dichotomous outcomes, researchers estimate coefficients for each predictor with maximum likelihood methods. Some modern techniques have been developed that aim to limit overfitting of a model to the available data, such as statistical shrinkage techniques, penalized maximum likelihood estimation, and the least absolute shrinkage and selection operator (LASSO).

5. *Model performance:* For a proposed model, researchers need to determine the quality. Several statistical performance measures are commonly used, including measures for model calibration and discrimination. Calibration refers to the reliability of predictions: If a researcher predicts 10%, on average 10% of the subjects with this prediction need to experience the outcome. Discrimination refers to the ability of a prediction model to separate subjects with and without the outcome and can, for example, be quantified by the area under the receiver operating characteristic (ROC) curve. Most relevant to clinical practice is whether the model is useful—that is, whether better decisions are made with the model than without. Usefulness is difficult to quantify, but novel performance measures such as decision curves may provide relevant insights.

6. *Model validity:* Since overfitting is a central problem in prediction modeling, researchers need to consider the validity of their model for new subjects rather than for those in the data set used for model development. Several statistical techniques are available to evaluate the internal validity of a model—that is, for the underlying population that the data set was sampled from. Internal validation may address statistical problems in the specification and estimation of a model ("reproducibility"). Common methods are cross-validation and bootstrap resampling procedures.

7. *Model presentation:* The final step to consider is the presentation of a prediction model. Regression formulas can be used, but many alternatives are possible for easier applicability of a model, including score charts, nomograms, and Web-based calculators.

From Model Development to Impact Analysis

A prediction model rises to the level of a decision rule if clinicians use its predictions to help make decisions for patients. The first phase is the valid development of a prediction model. Overfitting and measures to prevent overoptimistic expectations of model performance are especially important to consider at each of the seven steps of model development.

Phase 2 is related to external validation of the model, which is essential before application of a model can be recommended. Validation in multiple settings is required to gain confidence in the applicability of a model for yet another setting. Researchers may also use forthcoming data from validation studies to dynamically make changes to a model ("updating").

Finally, researchers need to consider an impact analysis, where a prediction model is used as a decision rule and any improvement in physicians' decisions is determined (quality or cost-effectiveness of patient care). Decision rules generally improve physicians' specificity more than sensitivity;

physicians ascribe greater value to true-positive decisions (provide care to patients who need it) than to true-negative decisions (withhold care from patients who do not need it). Sensitivity and specificity of a decision rule in clinical practice is influenced not only by the quality of the prediction model but also by the adherence of clinicians to the rule. Validation of a prediction model may indicate the efficacy of a rule (the maximum that can be attained with 100% adherence), but impact analysis will indicate its effectiveness in practice. Clinicians may choose to overrule the decision rule, which may improve sensitivity or specificity, but overruling may also dilute the effects of the rule.

There may be various barriers and facilitators to the clinical use of decision rules. Barriers include issues of attitude such as skepticism about guidelines (in general and with respect to the specific rule), questions on the clinical sensibility of the rule, too high confidence in clinical judgment, fear of medicolegal risks, concern that important factors are not addressed by the decision rule, and concern about patient safety. Furthermore, practical issues are important, such as availability of the rule at the time of decision making and ease of use.

Current and Future Successes

Prediction models and decision rules are important for more individualized medical decision making. The impact of a model may vary from setting to setting. A successful example is given by the Ottawa ankle rules, which started with model development, followed by validation and impact assessment. Many more successes are to be foreseen, for example, by the incorporation of stronger predictors such as biomarkers in prediction models, the development and validation of decision rules in close collaboration with clinicians, and the ongoing automatization in healthcare.

Ewout W. Steyerberg

See also Artificial Neural Networks; Biases in Human Prediction; Calibration; Cox Proportional Hazards Regression; Decision Rules; Decision Trees, Construction; Diagnostic Process, Making a Diagnosis; Logistic Regression; Markov Models; Maximum Likelihood Estimation Methods; Nomograms; Odds and Odds Ratio, Risk Ratio; Ordinary Least Squares Regression; Parametric Survival Analysis; Physician Estimates of Prognosis; Probability; Randomized Clinical Trials; Receiver Operating Characteristic (ROC) Curve; Recursive Partitioning; Risk Adjustment of Outcomes; Shared Decision Making; Support Vector Machines

Further Readings

Chun, F. K., Karakiewicz, P. I., Briganti, A., Gallina, A., Kattan, M. W., Montorsi, F., et al. (2006, November). Prostate cancer nomograms: An update. *European Urology, 50*(5), 914–926 (Discussion p. 926).

Harrell, F. E. (2001). *Regression modeling strategies: With applications to linear models, logistic regression, and survival analysis*. New York: Springer.

Harrell, F. E., Jr., Lee, K. L., & Mark, D. B. (1996). Multivariable prognostic models: Issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in Medicine, 15*(4), 361–387.

Hastie, T., Tibshirani, R., & Friedman, J. H. (2001). *The elements of statistical learning: Data mining, inference, and prediction*. New York: Springer.

Perry, J. J., & Stiell, I. G. (2006). Impact of clinical decision rules on clinical care of traumatic injuries to the foot and ankle, knee, cervical spine, and head. *Injury, 37*(12), 1157–1165.

Reilly, B. M., & Evans, A. T. (2006). Translating clinical research into clinical practice: Impact of using prediction rules to make decisions. *Annals of Internal Medicine, 144*(3), 201–209.

Steyerberg, E. W. (2009). *Clinical prediction models: A practical approach to development, validation, and updating*. New York: Springer.

Steyerberg, E. W., Eijkemans, M. J., Harrell, F. E., Jr., & Habbema, J. D. (2001). Prognostic modeling with logistic regression analysis: In search of a sensible strategy in small data sets. *Medical Decision Making, 21*(1), 45–56.

Steyerberg, E. W., Kallewaard, M., van der Graaf, Y., van Herwerden, L. A., & Habbema, J. D. (2000). Decision analyses for prophylactic replacement of the Bjork-Shiley convexo-concave heart valve: An evaluation of assumptions and estimates. *Medical Decision Making, 20*(1), 20–32.

Vittinghoff, E. (2005). *Regression methods in biostatistics: Linear, logistic, survival, and repeated measures models*. New York: Springer.

Preference Reversals

A preference reversal occurs when the same risky or uncertain choices are offered in different forms and the preferred choice changes. As such, preference reversals belong to a collection of phenomena, known as *choice anomalies*, that suggest that decision making is not always rational. The basic laboratory form of the preference reversal as a choice between gambles is described along with examples of preference reversals in medical decision making and other domains. Theories of preference reversal and their implications are discussed.

Choice Between Gambles

Expected utility theories of decision making assume that if individuals have sufficient information and are given sufficient time to process that information, they will make choices that maximize their interests. Since the choices people make reflect their best interests, those choices should be stable irrespective of how the preferences are elicited. For example, if given a choice between Options A and B, the agent prefers A, the agent should, when asked how much he or she would be willing to pay for the two options, pay more for Option A than for Option B. However, a great many studies have shown that the method of eliciting a preference can affect the choices people make. In the classic form, experimental participants are asked to choose between pairs of gambles. One choice has a high probability (P Bet) of winning a modest sum of money; the other has a low probability of winning a large sum of money. In the example below, the expected value (the probability of the outcome multiplied by the value of the outcome) of each choice is the same:

P Bet: 80% chance of winning $20.50,

$ Bet: 20% chance of winning $82.00.

When asked to *choose* between the two gambles most people prefer the P Bet. However, when asked to state the lowest amount of money they would sell the gambles for, or how much the gambles are worth to them, people tend to assign the $ Bet a higher monetary value. Thus, preferences elicited as a choice are reversed when those preferences are elicited in another form and violate expected utility theory's axiom of invariance.

Examples

The effect is not restricted to monetary decisions and occurs when the same individuals make the same choice twice and when the choices made by different individuals are elicited in different ways. For example, in a study of personnel selection, participants were asked to imagine that they were company executives choosing between two candidates, who had been scored by a selection committee along two dimensions (technical knowledge and human relations), for a position as a production engineer. One group of participants made their choices by choosing directly between the two candidates or by a matching technique. A second group received the same information, except that one of the four personnel scores was missing and had to be filled in to make the two candidates equivalent. The choices made directly tended to be reversed in the group whose choices were inferred by their evaluation of the missing attribute.

Preference reversals are not limited to situations in which preferences are elicited using direct choices and evaluations. They occur also when evaluations are made along different dimensions. For example, college students were asked to evaluate various hypothetical consumer- and health-related scenarios in terms of monetary value and life expectancy. The participants were asked to state how much they would be willing to sacrifice financially and in terms of life expectancy in order to gain an AIDS vaccine, immunity to tooth decay, a treatment that gives 20/20 vision, and one that provides immunity to cancer. Consumer items included airline tickets, theater tickets, vacations, and a date with a favorite celebrity. Although it might seem odd to compare different kinds of decisions along these two dimensions, the preferences elicited by them should be rank ordered in the same way on both scales, since any valuation scale reflects subjective utility. That is, the rank order of preferences on the monetary scale should be the same as the rank order of items along the life expectancy scale. The results, however, showed that health items were ranked higher than consumer items in terms of life expectancy value but

the consumer items were ranked higher on the monetary scale than the health items. This result indicates that the features of a decision (e.g., health benefits) are given more weight when they are meaningfully related to the scale on which they are evaluated (e.g., life expectancy).

A more commonly used valuation scale in medical-decision-making research is the time trade-off (TTO) method, in which participants estimate how many years in good health they consider equivalent to a particular outcome. Dutch college students would prefer to live 10 years in a state of constant migraine to living 20 years in the same state. The same students, when asked to state the number of years in perfect health that is equivalent to 10 years with a constant migraine and the number of years in perfect health equivalent to 20 years with a migraine, tended to assign a higher value for the 20 years of migraine than for the 10 years, thus reversing their preferences.

Such examples of irrationality are not restricted to the hypothetical patients but may be common too in medical practitioners. Family practitioners were presented with the case of an elderly patient with chronic hip pain and a diagnosis of osteoarthritis, for whom various nonsteroidal anti-inflammatory medications have proved ineffective. The patient agrees to be referred to an orthopedic surgeon in order to be assessed for possible hip replacement surgery. The case history was then subtly modified for two separate groups of family practitioners. One group was told that on inspection of the case notes, there was one nonsteroidal anti-inflammatory medication (ibuprofen) that had not yet been tried and were asked to choose between the following two alternatives:

A: Refer to an orthopedic surgeon and also start ibuprofen.

B: Refer to an orthopedic surgeon, but do not start any new medication.

The second group was told that two medications had not yet been tried (ibuprofen and piroxicam) and were asked to choose between three alternatives:

C: Refer to an orthopedic surgeon and also start ibuprofen.

D: Refer to an orthopedic surgeon and also start piroxicam.

E: Refer to an orthopedic surgeon, but do not start any new medication.

Just over half (53%) of the practitioners preferred not to start any new medication when given the option of either ibuprofen or no new medication (A and B). However, 72% of the group for whom two medications were available as alternatives preferred not to start either medication. Thus, 19% of the practitioners who would have attempted further medication when there was only one available would not do so when there were two available.

In a similar study, internal medical residents reviewed three hypothetical case histories: depression, sinusitis, and vaginitus. One group of medical residents were asked to choose between a relatively ineffective medicine with infrequent side effects and a relatively more effective medicine with frequent side effects for each of the three patients. A second were asked to make decisions about the same patients and the same medications, and a third, about moderately effective medication with occasional side effects. The results demonstrate that when three options are available, preference for the medication without side effects increases relative to when there are just two alternatives, despite the fact that the third option is inferior to the other options.

Theories

A number of different hypotheses have been proposed to explain preference reversals.

Compatibility Hypothesis

Different methods of eliciting a preference emphasize different pieces of information. When making a direct choice between two alternatives, people are primarily concerned with the attribute that is most important to the success of the outcome—namely, the probability. But when people are asked to evaluate the options with regard to a selling price, greater emphasis is placed on the monetary value of the options. Moreover, a related phenomenon may influence such evaluations: When stating a selling or purchase price, people

tend to *anchor and adjust* their price on the dollar amount stated in the gamble and consequently assign a higher value than it is worth objectively.

Evaluability Hypothesis

Some dimensions along which preferences are elicited may be more difficult to judge in isolation than others. Specifically, when choices involve a trade-off between an easy-to-evaluate dimension (e.g., immediate efficacy) and a hard-to-evaluate dimension (e.g., possible side effects), the harder dimension will receive less attention when the two dimensions are considered together.

Attraction or Dominance Effect

The most likely explanation for the increase in preference for one of the options when there are three rather than two alternatives is that the third, least preferred option dominates the second choice by highlighting its relative ineffectiveness. This kind of preference reversal is particularly relevant in medical decision making with regard to treatment decisions because there are often numerous medications available for a particular condition and the options available increase as new medications and treatments are developed.

Reasoning About Probability

People's perception and ability to reason about objective stated probabilities are often imperfect. Low probabilities tend to be overweighted, and high probabilities tend to be underweighted. The incidence of preference reversals is reduced, but not eliminated entirely, when the options are presented in the form of frequencies (e.g., 10 out of 100) rather than probabilities (e.g., 10% or $p = .10$).

Implications

While the existence of preference reversals and other choice anomalies are a challenge to normative theories of decision making, they perhaps pose a more serious challenge to the individual faced with making difficult and often risky decisions and to those who must give people these options and interpret their choices, because they demonstrate that the choices that people make may not always indicate their true preference. This raises the question, "How is it possible to infer what an individual's true preference is if that preference changes as a result of how the choice is made?" The true preferences of individuals making single decisions, such as a patient deciding among treatment options, can best be elicited using the same principles used to improve the communication of risk (i.e., probability) in any domain. With regard to people making numerous decisions over time, the outlook is somewhat better in the sense that the incidence of preference reversals tends to decline with experience.

Richard J. Tunney

See also Choice Theories; Expected Utility Theory; Gain/Loss Framing Effects; Procedural Invariance and Its Violations; Subjective Expected Utility Theory; Violations of Probability Theory; Willingness to Pay

Further Readings

Chapman, G. B., & Johnson, E. J. (1995). Preference reversals in monetary and life expectancy evaluations. *Organizational Behavior and Human Decision Processes, 62,* 300–317.

Lichtenstein, S., & Slovic, P. (1971). Reversals of preferences between bids and choices in gambling decisions. *Journal of Experimental Psychology, 89,* 46–55.

Redelmeier, D. A., & Shafir, E. (1995). Medical decision-making in situations that offer multiple alternatives. *Journal of the American Medical Association, 273,* 302–305.

Schwartz, J. A., & Chapman, G. B. (1999). Are more options always better? The attraction effect in physicians' decisions about medications. *Medical Decision Making, 19,* 315–323.

Torrance, G. W., Thomas, W. H., & Sackett, D. L. (1972). A utility maximization model for evaluation of health care programs. *Health Services Research, 7,* 118–133.

Tunney, R. J. (2006). Preference reversals are diminished when gambles are presented as relative frequencies. *Quarterly Journal of Experimental Psychology, 59,* 1516–1523.

Tversky, A., Sattath, S., & Slovic, P. (1988). Contingent weighting in judgment and choice. *Psychological Review, 95,* 371–384.

PROBABILITY

The concept of probability was introduced as a way of representing our uncertainty about the world. Mathematically, a probability distribution is a mapping from the values of a variable to nonnegative real numbers. Semantically, a probability is the chance or likelihood that an event occurs, currently or in the future.

Variables and Their Values

A *variable* is a mathematical object that takes on values from a certain set, called its *domain*. For instance, the domain of the variable Sex can be {male, female}. Each variable represents a property of the real world.

Values of a Variable

By definition, the values that a variable can take must be both exclusive and exhaustive. *Exclusive* means that two values cannot be true simultaneously. *Exhaustive* means that the values must cover all possible cases. For example, the values male and female are mutually exclusive because a person cannot be a man and a woman; they are exhaustive because there exists no other possibility.

Types of Variables

A variable is said to be *discrete* if its domain has a finite number of values. In the above example, Sex takes on only two values. A variable is said to be *continuous* if its domain is a numerical interval, such as [0,1] or [–∞,∞]. For instance, age, weight, height, temperature, red cell count, end-diastolic area of a valve, and so on are all continuous variables.

A continuous variable can be discretized by partitioning its domain into a finite number of subintervals. For instance, when modeling a medical problem, we could define three intervals for the variable Age: young = from 0 to 25, adult = from 26 to 70, and elderly = over 70. In another situation, it might be more appropriate to define the intervals differently, for example, young = from 0 to 15, adult = from 16 to 65, and elderly = over 65, or even to define more intervals: from 0 to 5, from 6 to 10, from 11 to 15, from 16 to 20, and so on.

Individual Probability

Probability of a Discrete Variable

As a first approach, we can define the probability of a discrete variable X as a function that assigns to each value x a number between 0 and 1 (both inclusive) such that the sum of them must be 1:

$$0 \leq P(x) \leq 1,$$

$$\sum_{x} P(x) = 1.$$

For example, for the variable Age mentioned above, we could have the following assignment of probability: $P(\text{young}) = .35$, $P(\text{adult}) = .46$, and $P(\text{elderly}) = .19$. Each of these probabilities is between 0 and 1, and the sum of all is 1: $P(\text{young}) + P(\text{adult}) + P(\text{elderly}) = 1$.

Probability of a Continuous Variable

The definition of the probability of a continuous variable is much more complex than in the discrete case. Let us assume that X is a continuous variable taking on real values. The basis for the definition of a continuous probability distribution is a function $F(x)$, called a *cumulative distribution function*, which, by definition, must satisfy the following properties.

1. $F(x)$ is monotonically nondecreasing; that is, if $b > a$, then $F(b) \geq F(a)$.
2. $\lim_{x \to -\infty} F(x) = 0.$
3. $\lim_{x \to +\infty} F(x) = 1.$

Roughly speaking, the first expression means that $F(x)$ increases—or at least does not decrease—when x increases, and the latter two mean, respectively, that the smaller the value of x, the closer is $F(x)$ to 0 and the greater the value of x, the closer is $F(x)$ to 1 (see Figure 1).

Then, the probability that X lies in the interval $[a,b]$ is given by

$$P(a \leq X \leq b) = F(b) - F(a).$$

There exists in general a function $f(x)$, called the *probability density function*, defined as $f(x) = dF(x)/dx$, which leads to

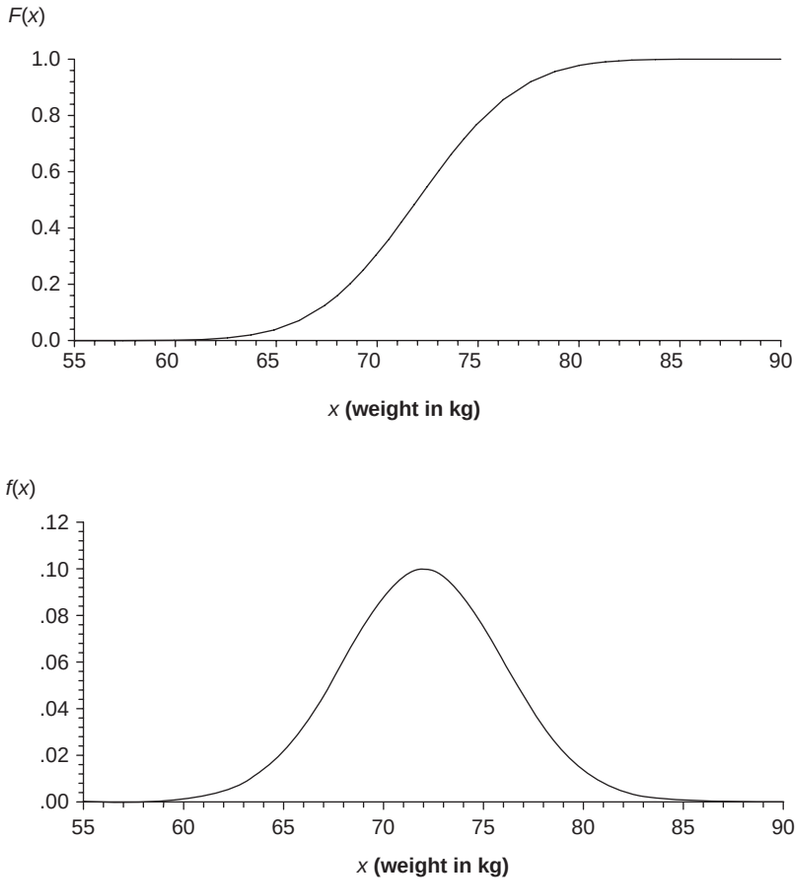


Figure 1 Weight of a population modeled by a Gaussian distribution of mean $\mu = 72$ kg and standard deviation $\sigma = 4$ kg. *Top:* Cumulative distribution function, $F(x)$. *Bottom:* Probability density function, $f(x)$.

$$P(a \le X \le b) = \int_a^b f(x)dx.$$

However, in practice, the probability P is never defined from the cumulative function, F, but from the density function, f.

For example, the weight of individuals in a certain population might be modeled by assuming that its probability density function is given by a Gaussian distribution of mean $\mu = 72$ kg and standard deviation $\sigma = 4$ kg (see Figure 1):

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2} = \frac{1}{4\sqrt{2\pi}} e^{-(x-72)^2/(2\times 4^2)}.$$

The probability that the weight of an individual randomly taken from that population lies between 71 and 75 kg is 37.2%:

$$P(71 \le X \le 75) = \frac{1}{4\sqrt{2\pi}} \int_{71}^{75} e^{-(x-72)^2/(2\times 4^2)} = .372.$$

Families of Probability Distributions

The set of all Gaussian distributions, also called normal distributions, form a family, given by an expression that depends on two parameters, μ and σ. Each assignment of parameters (e.g., $\{\mu = 72$ kg, $\sigma = 4$ kg$\}$) leads to a particular probability distribution function. Other families

of continuous distributions often used in statistics are χ^2 (chi-square), Γ (gamma), β (beta), Student's *S*, Fisher's *F*, and so on. The mathematical expressions for these functions can be found in any textbook on statistics or on the Internet.

Joint and Marginal Probabilities

Joint Probability of Several Discrete Variables

The definition of individual probability (of a variable) can be generalized to that of *joint probability* (of several variables). When having two discrete variables *X* and *Y*, the joint probability $P(x, y)$ is any function that fulfills these properties:

$$0 \leq P(x,y) \leq 1,$$

$$\sum_x \sum_y P(x,y) = 1.$$

The joint probability of three or more variables is defined similarly. For example, let *X* be the variable Sex, which takes on the values male and female, and *Y* the variable Age, which takes the values young, adult, and elderly. In a certain population, the joint probability for these two variables is as follows:

P(male, young) = .183	*P*(female, young) = .167
P(male, adult) = .210	*P*(female, adult) = .222
P(male, elderly) = .093	*P*(female, elderly) = .125

Clearly, all the probabilities are between 0 and 1, and their sum is 1.

Marginal Probability of Discrete Variables

From the joint probability of several variables, we can obtain the *marginal probability* for a subset of them by summing over the rest of the variables. For example, given $P(x, y)$, the marginal probability $P(x)$ is obtained summing over the variable we want to "eliminate," namely, *Y*:

$$P(x) = \sum_y P(x,y).$$

In the same way, the marginal probability $P(y)$ is obtained summing over *X*:

$$P(y) = \sum_x P(x,y).$$

We can represent the joint and marginal probabilities of this example as given in Table 1.

These results can be generalized for more than two variables. For instance, from the joint probability $P(x, y, z)$, we can derive six marginal probabilities:

$$P(x,y) = \sum_z P(x,y,z)$$
$$P(x,y) = \sum_z P(x,y,z)$$
$$P(x,y) = \sum_z P(x,y,z)$$
$$P(x) = \sum_y \sum_z P(x,y,z)$$
$$P(y) = \sum_x \sum_z P(x,y,z)$$
$$P(z) = \sum_x \sum_y P(x,y,z)$$

Joint and Marginal Probability Distributions of Continuous Variables

Axiomatically, the definition of the joint probability of several continuous variables is based on a *cumulative distribution function*, $F(x, y, z, \ldots)$, whose properties are analogous to the case of a single variable, and the *probability density function* derives from it. However, in practice, the probability *P* is never defined from a cumulative function, *F*, but from a density function, *f*, assuming that *f* belongs to one of the families of multivariate continuous probability distributions, such as the multivariate Gaussian, also called multivariate normal. Other families of continuous probability density distributions can be found in statistical textbooks and on the Internet.

The derivation of the marginal density function from a joint probability density is analogous to the case of discrete variables, just replacing the sum with an integral. For example, in the case of a joint

Table 1 Joint and marginal probabilities for a hypothetical population

P(x, y)	*male*	*female*	*P(y)*
young	.183	.167	.350
adult	.210	.222	.432
elderly	.093	.125	.218
P(x)	.486	.514	1.000

density $f(x, y)$ defined over two variables X and Y, the marginal density distributions are

$$f(x) = \int_{\infty}^{+\infty} f(x,y)dy,$$

$$f(y) = \int_{\infty}^{+\infty} f(x,y)dx.$$

The generalization of these equations for a higher number of variables is obvious.

Francisco J. Díez

See also Bayes's Theorem; Conditional Independence; Conditional Probability; Diagnostic Tests; Odds and Odds Ratio, Risk Ratio; Probability, Verbal Expressions of; Subjective Probability; Violations of Probability Theory

Further Readings

de Finetti, B. (1974, 1975). *Theory of probability* (2 vols.). New York: Wiley.

Jeffrey, R. C. (1992). *Probability and the art of judgment*. New York: Cambridge University Press.

Kolmogorov, A. N. (1950). *Foundations of the theory of probability*. New York: Chelsea.

Probability, Verbal Expressions of

Verbal expressions of probability are those used for communicating degrees of uncertainty, such as *likely, very likely, possible*, and *uncertain*, in contrast with numerical expressions, such as "The probability of X is 70%" or "The odds are 3 to 1."

Correspondence Between Verbal and Numerical Expressions

Since the 1960s, several authors have studied the equivalence between verbal and numerical expressions of probability. The typical experiment consists in selecting a set of linguistic expressions, asking a group of subjects to translate each one into a percentage or a number between 0 and 1, and building a table or a graph that summarizes the results. Most of such studies include expressions that do not represent probabilities in a strict sense.

Variability of Numerical Assignments

Several experiments have indicated low within-subject variability, that is, estimates given by the same subject on different occasions are very similar, while all the studies have shown a high degree of between-subject variability. A related empirical finding is that in general, people underestimate how much individuals vary in their interpretation of probability terms.

The studies have also shown a reasonable degree of between-experiment consistency. In some cases, the numerical values obtained in a study differed from those obtained in others, but the ranking of expressions was essentially the same.

Several experiments that compared the numerical values assigned by different groups of subjects have shown that between-subject variability is smaller among people with similar backgrounds. For instance, a study by Nakao and Axelrod showed that consensus was significantly higher among physicians than among laymen for around half of the expressions of frequency examined and also among native-English-speaking physicians than among those with other native languages, but it was not higher among board-certified physicians than among the others.

The Role of Modifiers

The meaning of a verbal expression of probability can be modified by the use of adverbs (*very likely*), affixes (*im-probable*, *un-likely*), or lexical negations (*not likely*). Empirical studies have led to the following ranking of adverbs: *very* > *quite* > no modifier > *rather* > *fairly* > *somewhat*, which means that *very* is the adverb that shifts most the meaning of a probability expression toward extreme values. *Very likely* denotes a higher probability than *quite likely*, which in turn denotes a higher probability than *likely* (no modifier). On the contrary, *rather*, *fairly*, and *somewhat* shift the meaning of the expression toward .5.

The Influence of Context

While most of the experiments asked the subjects to translate isolated linguistic expressions, other researchers have studied those expressions in context.

A surprising finding is that in most of the cases, between-subject variability is higher when probabilistic expressions are given in context. Another finding, not surprisingly, is the *base-rate effect*, which means that, in general, the higher the prior probability of an event, the higher the numerical values assigned. For instance, the term *likely* in "It is likely that it will snow in December" is assigned higher values than in "It is likely that it will snow in October."

Qualitative expressions associated with more severe outcomes (e.g., "likely death" vs. "likely injury") tend to be assigned lower numbers. This might also be explained by the base-rate effect, because in general more severe outcomes have lower base rates. There is also empirical evidence that expressions associated with positive outcomes tend to be assigned higher numbers than those associated with negative outcomes.

More interestingly, Mazur and Merz proved that personal characteristics, such as age, healthcare experience, and perceived health status, influence patients' interpretations of verbal probability terms.

Preferences for Numerical or Verbal Probabilities

Reasons for Preferring Verbal Probabilities

One of the reasons for using verbal probabilities is that they are more natural than numbers: Spontaneously, people express probabilities with linguistic terms, whereas it requires an additional cognitive effort to give numeric estimates.

Additionally, verbal probabilities can reflect the speaker's lack of knowledge: Very often expressions of probability do not stem from systematic data, but they are estimates made by human beings based on the cases stored in their memory, on what they have read or heard, and so on. In this case, people do not dare convey a numerical probability, for two reasons. The first is that an empirical study might later prove the assertion to be wrong. For instance, if an expert says in a book that the prevalence of a disease is 1%—or even "around 1%," which is an imprecise probability—and a posterior study shows that it is .4% or 3%, the expert's reputation will be compromised. On the contrary, if he said that "the prevalence is relatively high," he does not commit to a particular figure, and so his assertion cannot be refuted. The second reason is that subjective estimates expressed as numerical probabilities may mislead the listener to believe that the speaker knows the true probability with precision.

In addition to cases in which the probability has an objective value but is unknown, there are other cases in which it does not make sense to assume that there exists a measurable probability. For instance, a doctor may feel unable to answer with a precise value a question about the probability of a patient's survival, because a question such as this, referring to a *single-event probability*, does not have an objective meaning. In this case, it is much easier to respond with a linguistic probability.

Another reason for using verbal expressions is that in addition to conveying a probability estimate, they can also express *directionality*. A phrase having positive direction, such as "*X* is possible," implicitly points at the reasons for the occurrence of *X*, while a phrase of negative direction, such as "*X* is uncertain" or "*X* is doubtful," implicitly underlines the causes that may prevent *X*. Therefore, verbal expressions may be preferred when the speaker, in addition to conveying a vague probability, wishes to make the listener pay more attention to the reasons in favor of or against the occurrence of an event.

Empirical Evidence

Many experiments have been carried out to study human preferences about probability expressions. The most consistent finding is that while more people prefer to receive information about probabilities numerically, they prefer to express such information verbally. This is called the *preference paradox*.

In addition to the direction of communication (giving vs. receiving information), other factors have been shown to influence human preferences. One of them is the *nature of the event*: When expressing the probability of repeated events with aleatory uncertainty, most individuals prefer to use numerical estimates, which allow them to distinguish between levels of uncertainty with higher precision, but the same individuals tend to use more imprecise methods when communicating single-event probabilities. Another factor is the *strength of the available evidence*: People tend to use more precise expressions of probability when the information is firmer and more reliable.

In some of the studies, the people giving information were doctors and those receiving it were

patients. Other studies have set a scenario in which subjects were randomly assigned to the group of advisers or to the group of decision makers, whose choice is based on the information received from the advisers.

A different problem, related to the construction of decision support systems, is the elicitation of the parameters of a probabilistic model, such as a Bayesian network or an influence diagram. An empirical study carried out by Witteman and colleagues, in which general practitioners had to assess several conditional probabilities, concluded that the less experienced doctors preferred a purely verbal scale, the most experienced preferred a purely numerical scale, while the groups in between preferred a combined verbal-numerical scale.

Impact on Medical Decision Making

The use of verbal expressions of probability poses a serious problem as a potential source of errors, particularly in the case of informed consent. The first problem is the risk of misunderstanding. Let us imagine a patient suffering from a disease that will cause his death. His doctor offers him a treatment that may save his life but may have side effects. The decision of accepting the treatment depends on the probability of survival and on the probability and severity of side effects. In this context, verbal expressions of probability entail an obvious danger of misunderstanding: The doctor's estimate that there is a 60% probability of survival, conveyed as "It is likely that you will get cured," might be interpreted by the patient as a 90% probability, and the assertion that "sometimes the treatment causes severe adverse effects" may be interpreted as having a probability lower than 2% or higher than 15%. The danger is even higher in the case of extreme probabilities, because an expression such as *very unlikely* may mean .1 as well as .00001 probability. However, there is ample evidence that people, including investigators, underestimate the variability of subjective estimations.

The second issue is how patients process the information that they receive from their doctors. On this point, the empirical evidence is contradictory. Some studies seem to demonstrate that subjects are more effective at reasoning with verbal expressions than with numerical expressions, even if the tasks performed rely on frequency information. However, other studies arrived at the opposite conclusion, and others have found no significant difference.

Disadvantages and Advantages

Verbal expressions of probability are often used in medical communications—in fact, much more often than numerical expressions. Experts in the field defend contradictory opinions about their usefulness and their peril.

The main drawback of verbal probabilities is the risk of misunderstanding, because the value interpreted by the listener can be very different from that intended by the speaker. Some researchers have proposed using a very limited number of linguistic probabilities, whose meaning should be explicitly determined beforehand. However, this proposal does not solve two of the main problems: that the interpretation of such expressions varies with the context (the aforementioned base-rate effect) and that verbal probabilities are not able to discriminate extreme values, such as .1 and .0001. On the contrary, this lack of precision of linguistic expressions turns into an advantage in the usual case of imprecise knowledge: In that case it may be very difficult for the speaker to utter a numerical probability, and, even worse, that precise probability may mislead the recipient of the information.

On the other hand, a disadvantage of verbal expressions of uncertainty is the lack of a normative calculus, in contrast with the well-defined principles and techniques of probability theory, which play an essential role in medical diagnosis and decision making. Additionally, some experiments have shown that numerical probabilities lead to better judgments and to better decisions. However, other studies have arrived at the opposite conclusion, or at least a tie.

The directionality of verbal expressions, which carries additional information, has been put forward as one of their advantages. In contrast, numerical expressions, because of their neutrality, should be chosen when the speaker does not wish to bias the listener.

As an attempt to combine the advantages of both, some experts advocate using them together, by appending to each linguistic expression its intended meaning—for instance, "It is very likely (80%–90%) that. . . ."

The movement of evidence-based medicine and the use of computerized decision support systems will give an increasingly prominent role to numerical probabilities, to the detriment of verbal expressions, but because of the above arguments and the strong human preferences, it is clear that the use of linguistic probabilities will never disappear from medical communications, either oral or written.

Francisco J. Díez and Marek J. Druzdzel

See also Human Cognitive Systems; Probability; Risk Communication; Subjective Probability

Further Readings

Brun, W., & Teigen, K. H. (1988). Verbal probabilities: Ambiguous, context-dependent, or both? *Organizational Behavior and Human Decision Processes, 41,* 307–314.

Budescu, D. V., & Wallsten, T. S. (1995). Processing linguistic probabilities: General principles and empirical evidence. In J. R. Busemeyer, R. Hastie, & D. Medin (Eds.), *The psychology of learning and motivation: Decision making from the perspective of cognitive psychology* (pp. 275–318). New York: Academic Press.

Clark, D. A. (1990). Verbal uncertainty expressions: A critical review of two decades of research. *Current Psychology, 9,* 203–235.

Lipkus, I. M. (2007). Numeric, verbal, and visual formats of conveying health risks: Suggested best practices and future recommendations. *Medical Decision Making, 27,* 696–713.

Mazur, D. J., & Merz, J. F. (1994). Patients' interpretations of verbal expressions of probability: Implications for securing informed consent to medical interventions. *Behavioral Sciences and the Law, 12,* 417–426.

Nakao, M. A., & Axelrod, S. (1983). Numbers are better than words: Verbal specifications of frequency have no place in medicine. *American Journal of Medicine, 74,* 1061–1065.

Teigen, K. H., & Brun, W. (1999). The directionality of verbal probability expressions: Effects on decisions, predictions, and probabilistic reasoning. *Organizational Behavior and Human Decision Processes, 80,* 155–190.

Teigen, K. H., & Brun, W. (2003). Verbal expressions of uncertainty and probability. In D. Hardman & L. Macchi (Eds.), *Thinking: Psychological perspectives on reasoning, judgement and decision making* (chap. 7, pp. 125–145). Chichester, UK: Wiley.

Wallsten, T. S., & Budescu, D. V. (1995). A review of human linguistic probability processing: General principles and empirical evidence. *Knowledge Engineering Review, 10,* 43–62.

Witteman, C. L. M., Renooij, S., & Koele, P. (2007). Medicine in words and numbers: A cross-sectional survey comparing probability assessment scales. *BMC Medical Informatics and Decision Making, 7,* 13.

Probability Errors

Physicians and patients deciding on treatment options often need to estimate the probability of various outcomes (e.g., death, partial recovery, full recovery) in order to make effective decisions. However, human probability estimation is often fraught with errors that may interfere with making correct decisions. There are a number of probability errors that physicians ought to be aware of.

Human probability judgments often exhibit a rather basic bias: We overestimate low probabilities and underestimate high probabilities. For instance, people overestimate the likelihood of dying from rare diseases such as smallpox or botulism while underestimating the likelihood of dying from more common afflictions such as strokes or heart disease. One explanation for these biases might be that our probability estimates tend to regress toward the mean. That is, estimates of low probabilities have nowhere to go but higher while estimates of high probabilities can more easily go lower.

Another reason for these biases has to do with the *availability heuristic*, which states that we base our estimates of an event's likelihood on how easy it is to imagine that event occurring. In other words, we overweight information that comes easily to mind. Because rare events get a large amount of graphic news coverage (e.g., plane crashes, the contracting of rare diseases), these events are easier to imagine than mundane and less publicized events (e.g., car accidents, asthma attacks). Given this, it is perhaps not surprising that surgeons from high-mortality specialties give higher estimates of dying in the hospital than do surgeons from low-mortality specialties.

Relatedly, the way in which we think about information can also affect the ease with which we envision certain events and our subsequent estimates of these events' probabilities. If composite events are

"unpacked" into their components, then these components become easier to envision, and people estimate their probabilities to be higher. For instance, the risk of dying from "natural causes" can be unpacked into heart attacks, cancers, and other natural causes. People who were asked about the risk of dying from natural causes responded that it was 58%. However, people who were asked about the probability of the unpacked events responded that the risk of dying from heart attacks was 22%; from cancer, 18%; and from other natural causes, 33%. This sums to 73%, considerably higher than the 58% estimate. Probability estimates increase when individual components are considered. When this occurs, such estimates are considered to be *subadditive*, as the judgment for the composite event is less than the sum of the judgments for its parts.

Decision makers also overweight surface similarities between events when judging probabilities, a strategy known as the *representativeness heuristic*. For instance, given a description of a woman who is bright, outspoken, and concerned with social justice, Amos Tversky and Daniel Kahneman found that people are more likely to state that this woman is a feminist bank teller than just a bank teller. However, as all feminist bank tellers are also bank tellers, one cannot be more likely to be a feminist bank teller than a bank teller. This bias occurs because the description of the woman resembles the category of feminist bank tellers more than the more general category of bank tellers.

Another prominent probability error is *base-rate neglect*, whereby decision makers ignore the statistical properties of an outcome, such as its frequency, and attend more to the specifics of a given circumstance. Consider the following probabilities that were given to participants in a study: 80% of women with cancer will receive a positive mammogram, 9.6% of women who do not have cancer will receive a positive mammogram, and 1% of all women who get a mammogram will have cancer. Participants who were asked to judge the likelihood that a woman with a positive mammogram has cancer often gave high probability estimates (above 70%). However, statistics dictate that the probability is much lower, 7.8%. Without exploring the statistical calculations, the actual probability is lower largely because only 1% of women have cancer. Thus, participants appear to ignore the low base rate, focusing instead on the mammogram results.

Probability errors can also arise through miscommunication. For example, the probability of experiencing side effects from medications is often described in semantic terms (e.g., "Some people may experience *X*"). After hearing semantic descriptors, patients often interpret the probabilities of side effects as being as high as 25%, even though these phrases typically are meant to imply probabilities of less than 5%. In a similar vein, patients told that they were at "low risk" for diseases such as aspergillosis believed that they were over 10,000 times more likely to get the disease than they actually were. The ambiguity inherent in the verbal description leads people to overestimate the actual probabilities of the event.

Although these probability biases pose significant problems for medical decision making, some steps can be taken to mitigate their effects. For instance, describing the frequencies of outcomes (e.g., 10 out of 100) instead of their probabilities (e.g., 10%) might mitigate base-rate neglect. Other means of effectively communicating probabilities include risk ladders, community risk scales, magnifier scales, and diagrams/graphics. Additionally, considering the causal relationships between different events can lead to more accurate probability assessments. For instance, when people are given a cause for false-positive mammograms (rather than simply being told that there is a certain percentage of false positives), their diagnoses in the mammogram scenario from above become much more accurate.

Anuj K. Shah and Daniel M. Oppenheimer

See also Biases in Human Prediction; Heuristics

Further Readings

Berry, D. C. (2004). Interpreting information about medication side effects: Differences in risk perception and intention to comply when medicines are prescribed for adults or young children. *Psychology, Health & Medicine, 9*(2), 227–234.

Calman, K. C., & Royston, G. H. (1997). Risk language and dialects. *British Medical Journal*, *315,* 939–942.

Dawson, N. V., & Arkes, H. R. (1987). Systematic errors in medical decision making: Judgment limitations. *Journal of General Medicine, 2,* 183–187.

Eddy, D. M. (1982). Probabilistic reasoning in clinical medicine: Problems and opportunities. In D. Kahneman,

P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 249–267). Cambridge, UK: Cambridge University Press.
Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica, 47,* 263–291.
Krynski, T. R., & Tenenbaum, J. B. (2007). The role of causality in judgment under uncertainty. *Journal of Experimental Psychology: General, 436,* 130–150.
Lichtenstein, S., Slovic, P., Fischhoff, B., Layman, M., & Combs, B. (1978). Judged frequency of lethal events. *Journal of Experimental Psychology: Human Learning and Memory, 4,* 551–578.
Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science, 185,* 1124–1131.
Tversky, A., & Kahneman, D. (1983). Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological Review, 90,* 293–315.
Woloshin, S., Schwartz, L. M., Byram, S., Fischoff, B., & Welch, H. G. (2000). A new scale for assessing perceptions of chance: A validation study. *Medical Decision Making, 20,* 298–307.
Young, S. D., & Oppenheimer, D. M. (2006). Percentages matter: Framing risk information can affect fear of side effects and medication compliance. *Clinical Therapeutics, 28*(1), 129–139.

Problem Solving

Clinically active physicians possess large amounts of declarative and procedural knowledge. By one estimate, experienced physicians have more than 2 million discrete pieces of information at their immediate disposal. But this knowledge offers little benefit to patients except in how its use relates to problem solving. People seek out physicians not for the large number of facts that the physicians can recall but for help with medical problems. Although there is a correlation between knowledge and problem-solving success, successful problem solving requires more than a large amount of knowledge. Success also requires appropriate organization of the knowledge, application of the knowledge, and monitoring of the results of the application of knowledge so that adjustments can be made.

Problem solving is the process of answering or solving the question being considered. Problem solving requires a person to ascertain the current state of a situation, identify a specific goal state, possess knowledge about how to move from a starting state to a goal state, and monitor the progress of moving toward the goal state.

Historical Roots

Problem solving and decision making are the products of different research efforts. Problem solving has been studied by psychologists, who have concentrated on identifying the problem-solving strategies of experts in a field, with the aim of facilitating the acquisition of these strategies by learners. Problem-solving research has thus focused on the wisdom of practice. Decision research has typically contrasted human performance with a normative statistical model of reasoning under uncertainty. Maximizing subjective utility is often the theoretical model used for the normative model. Decision research has emphasized errors in reasoning about uncertainty, typically demonstrating that even experts in a domain are not immune from these errors and thus raising the case for some type of external support.

Herbert Simon, a Nobel laureate in economics, argued that problem solving involves choosing the issues that require attention, setting goals, and designing the courses of action, while decision making involves evaluating and choosing among alternative possible courses of action. David Jonassen has suggested that decision making involves the process of identifying the benefits and limitations of different alternatives, selecting the best one, and being able to justify the choice. But according to Jonassen, a problem might be solved using only decision making (e.g., a problem involving selecting the goal state), or decision making might be only one of several interventions needed to solve a problem. Thus, the terms *decision making* and *problem solving* may be used interchangeably in the knowledge that they have simply arisen out of different research traditions. The two terms can be used to focus on different aspects of the cognitive process of finding the best solution to a problem, or the two terms might be used with *problem solving* being more global in description than *decision making*.

Problem Structure and Complexity

Problems can be grouped into typologies that are related to their structuredness. A well-structured

problem is one where all the elements of the problem are present; one needs only a limited number of rules, concepts, and principles to solve the problem; the problem has a correct, convergent answer; there is a preferred, prescribed solution process; and the relationship between decision choices and all problem states is known. A puzzle is an example of a well-structured problem, because the initial and goal states are clearly defined as are the rules about the operations that are allowed to be undertaken for closing the gap between the initial and goal states. In contrast to a well-structured problem is an ill-defined problem. Ill-defined problems are challenging because there are no single optimal solutions. In addition, one or more of the problem elements are unknown or not known with any degree of confidence, there is no explicit means for determining appropriate action, and people can be required to express personal beliefs about the problem and to make judgments about the problem and defend them.

An example of an ill-defined problem is the design problem routinely faced by architects when asked by a client to produce a design for a new house. The client may have notions of how he or she wants to use the space but not the expertise to know what form will result in the desired function. The client will also have many constraints that the architect will need to deal with, including financial resources, limitations of the building site, and inexact notions of what the client wants. Due to the complexity of these types of problems, the architect is not able to offer a pregenerated solution that will automatically meet the client's goals while conforming to the constraints. Many medical problems are also design problems—the challenge faced by the physician is in designing a reasonable solution to a complex problem that results in a desired goal state and involves acceptable trade-offs.

Another characteristic of ill-defined problems is that it is often not clear when the problem is satisfactorily solved. Allen Newell and Simon posited that most problem/decision tasks are successfully solved through satisficing rather than optimizing. A better solution, given more time or resources, is almost always possible. Effective problem solvers find "good enough" solutions instead of pursuing perfect ones.

Between these polar extremes of problem structure, there are many different types of problems. A typology of problems proposed by Jonassen includes puzzles, algorithms, story problems, rule-using problems, decision making, troubleshooting, diagnosis-solution problems, strategic performance, systems analysis, design problems, and dilemmas. He has argued that each type of problem requires different problem-specific skills. For example, puzzle problems require that the problem solver clearly understand the allowable rules of the puzzle and demonstrate a logical approach in moving to the goal state and the ability to monitor different attempts at solving the puzzle so that ineffective attempts are not repeated. To help learners be successful with troubleshooting problems, novices need to acquire specific conceptual, procedural, and strategic knowledge related to the machine they are trying to troubleshoot. In general, theoretical knowledge is not very important for troubleshooting success. Of course, novice troubleshooters need to be guided in their efforts to troubleshoot by an expert so that they integrate their knowledge and gain experience. For novices to master the solution of a defined design problem, they need to learn how to formulate a well-defined expression of the typically ill-defined task and to develop a coherent argument for their solution choices.

In addition to structuredness, problems also vary in terms of their complexity—the number of variables involved in the problem and the interrelationship between these variables. Ill-structured problems tend to be more complex than well-structured problems, but even a well-structured problem can be very complex. A Rubik's Cube is an example of a well-structured puzzle problem that is very complex. While the initial and goal states are very clear and the permitted operations for the problem solver are clearly defined, there are 43,252,003,274,489,000 different possible configurations. One, and only one, of these possible configurations is the "solved" Rubik's Cube—squares of a single color on each of its six sides. Complexity increases the difficulty of a problem because the problem solver needs to take into account a greater number of parts, factors, and possible solutions. Novices need to learn how to use memory aids to help them keep track of their different approaches to solving a problem so that they do not repeat ineffective approaches.

Problem-Solving Strategies

There are clearly differences among individuals in their ability to solve problems. Familiarity with a problem is a strong predictor of a problem solver's ability to solve a specific problem. The problem

solver who is familiar with a problem might quickly provide the solution because he or she is recalling it from memory. Another strong predictor of problem-solving success is the solver's level of domain knowledge. An individual's skill in monitoring his or her own performance is also related to problem-solving success, as is an individual's epistemological beliefs. Despite these well-documented differences in ability among problem solvers, it has been only recently that researchers have systematically searched for better ways to teach learners how to solve problems.

It is now clear that guided instruction in problem solving is superior to simply asking learners to tackle the problem on their own. A widely known general problem-solving strategy is the one proposed by the mathematician George Polya in the 1950s. Although he primarily focused on solving mathematical problems, his approach may be applicable to other types of problems. Polya encouraged the problem solver to adopt a systematic set of steps. The first step is to understand the problem: Problem solvers need to clarify the goal and make an inventory of the information they have at their disposal. Polya felt that it was helpful for the problem solver to diagram the problem so that the problem is visualized and better understood. Once the problem is understood, Polya suggested, the problem solver is ready to devise a plan for solving the problem. He suggested that there are many reasonable ways to solve problems and that skill in choosing an appropriate strategy is best learned by solving many problems. This step involves metacognition: The problem solver needs to search his or her memory and assess whether he or she already knows the solution to the problem, or something close to it. If not, the next task is to determine whether the problem solver is able to restate the problem in a way that makes it possible to use a previously learned solution. Another approach is to break the problem down into subproblems, each of which is easier to solve. Once the plan is devised, the problem solver needs to carry out the plan and check back on whether the solution makes sense. Researchers have reported moderate to strong associations between use of Polya's steps by mathematical problem solvers and their success at problem solving.

Polya's second step, devising a plan for solving a problem, is the crux of the challenge of problem solving, and Polya only provided a general description about how the plan could be generated. One approach frequently observed by researchers is means-end analysis, in which the problem solver iteratively attempts to close the gap between the initial conditions and the goal state through a series of steps. A specific example is the process a person would use after turning on a television set to watch a program but finding only snow on the display. The starting state is a television displaying only snow. The goal state is a television displaying the desired program. The problem solver might first set a subgoal—that of confirming that a signal is being delivered to the television. The problem solver might first check to make sure that the cable converter box is attached. If this does not correct the picture, then the problem solver may check if the cable converter is turned on. If it is turned on, the problem solver might check if the cable box is delivering a signal by attaching the unit to a different television set. If the second television also displays snow, the problem solver may check if the cable converter is set on the correct channel. While this means-end strategy is used by the novice, a more problem-specific approach is likely to be more successful, and the experienced problem solver tends to take a forward problem-solving approach. The expert might check the channel setting as a first step since he or she has learned that this is the most common problem.

Although the means-end analysis is not the approach of the expert with routine problems, this approach is observed when the expert encounters a problem outside his or her realm of expertise. Because means-end analysis is used by people facing a novel problem, it might be argued that the means-end analysis is an effective learning approach for novices. Unfortunately, when the novice uses this approach, it frequently does not result in knowledge that allows the problem to be solved in a more forward-oriented manner the next time the problem is encountered. When the unguided novice uses means-end analysis, the problem solver can be so focused on reducing the gap between the initial conditions and the goal state that only these small iterative steps are reflected on instead of the entire problem as a whole. Thus, the problem solver may not acquire the domain-specific strategies used by a more proficient problem solver. This problem highlights the importance of teaching novices specific approaches to solving problems; leaving the novice to discover the solutions is not a very efficient approach.

Domain Specificity

Problems are located within domains; for example, a problem may be within the domain of medicine, engineering, mathematics, or physics. Problem-solving expertise is usually domain specific; most people who are expert in physics problems are not going to be expert with medical problems. Even within one of these domains, problems can be further categorized. Thus, physics problems can be identified as problems involving momentum, acceleration, energy, inertia, or another subdomain. Medical problems can be identified by the discipline the problem focuses on, such as cardiology, neurology, gynecology, or some other discipline. The finding that problem solvers may be highly successful in solving problems in one domain or subdomain but not outside this area is called case or context specificity.

E. L. Thorndike and R. S. Woodworth reported case specificity in 1901, when they found that successfully teaching subjects how to estimate the area of rectangles did not cause the subjects to perform better at estimating the areas of triangles or circles. This finding has been replicated in many other problem domains, including medicine. Thus, teaching a learner how to solve one problem does not automatically help the learner solve a second problem, even if the two problems are conceptually similar. Although transfer between problems is limited, it can be enhanced by helping the learner gain an understanding of how to solve a specific problem instead of having the learner simply mimic a series of steps by rote. Having learners represent problems and their solutions at appropriate levels of abstraction also helps with transfer. Last, through emphasizing metacognition during the learning process, students are more likely to demonstrate transfer between problems.

Although case specificity is evidenced by the limited transfer of problem-solving skills from one problem to another, this limited transfer does not mean that helping a learner master a problem does not help the learner when faced with a new and different problem. While Thorndike and Woodworth demonstrated that mastering a problem in one area does not result in better problem solving in another area, they vigorously argued that once a person has learned to solve one problem, the person might demonstrate quicker mastery of a new and different problem. More recently, researchers have confirmed this effect, and it has been called "transferring in" or "preparation for future learning" because the knowledge gained from learning to solve one problem can make the learner more efficient in learning to solve other problems.

George Bergus and Alan Schwartz

See also Cognitive Psychology and Processes; Decision Psychology; Diagnostic Process, Making a Diagnosis; Errors in Clinical Reasoning; Hypothesis Testing; Judgment; Learning and Memory in Medical Training; Pattern Recognition; Teaching Diagnostic Clinical Reasoning

Further Readings

Bedard, J., & Chi, M. T. (1992). Expertise. *Current Directions in Psychological Science, 1*(14), 135–139.

Chi, M. T. H., & Glaser, R. (1985). Problem solving ability. In R. S. Sternberg (Ed.), *Human abilities: Information processing approach* (pp. 227–257). New York: Freeman.

Chi, M. T. H., Glaser, R., & Rees, E. (1982). Expertise in problem solving. In R. Sternberg (Ed.), *Advances in the psychology of human intelligence* (Vol. 1). Hillsdale, NJ: Lawrence Erlbaum.

Davidson, J. E., & Sternberg, R. J. (Eds.). (2003). *The psychology of problem solving*. Cambridge, UK: Cambridge University Press.

Elstein, A. S., Shulman, L. S., & Sprafka, S. A. (1978). *Medical problem solving: An analysis of clinical reasoning*. Cambridge, MA: Harvard University Press.

Jonassen, D. H. (2000). Toward a design theory of problem solving. *Educational Technology Research and Development, 48*(4), 63–85.

Jonassen, D. H. (2004). *Learning to solve problems: An instructional design guide*. San Francisco: Jossey-Bass.

Kirschner, P. A., Sweller, J., & Clark, R. E. (2006). Why minimal guidance during instruction does not work: An analysis of the failure of constructivist, discovery, problem-based, experiential, and inquiry-based teaching. *Educational Psychologist, 41,* 75–86.

Klahr, D., Chen, Z., & Toth, E. E. (2001). Cognitive development and science education: Ships that pass in the night or beacons of mutual illumination. In S. M. Carver & D. Klahr (Eds.), *Cognition and instruction: Twenty-five years of progress* (pp. 75–119). Hillsdale, NJ: Lawrence Erlbaum.

Newell, A., & Simon, H. A. (1972). *Human problem solving*. Englewood Cliffs, NJ: Prentice Hall.

Schoenfeld, A. (1987). *Cognitive science and mathematics education*. Hillsdale, NJ: Lawrence Erlbaum.

Schwartz, D., Bransford, J., & Sears, D. (2005). Efficiency and innovation in transfer. In J. Mestre (Ed.), *Transfer of learning from a modern multidisciplinary perspective* (pp. 1–52). Charlotte, NC: Information Age.

Procedural Invariance and Its Violations

Procedural invariance states that preferences over prospects (i.e., gambles, or any other risky states that can be described as a probability p of getting outcome/payoff x) are independent of the method used to elicit them. In other words, procedural invariance, an important pillar of rational choice, demands that strategically equivalent methods of elicitation will give rise to the same preference order. Satisfaction of procedural invariance is implied by the orderability axiom of von Neumann-Morgenstern utility theory. In medical decision making, procedural invariance is the condition that a person's preference ranking of two health states should not depend on the elicitation procedure. For example, suppose one uses both the visual analog scale and the standard gamble to assign quality weights to two health states. Procedural invariance requires that the visual analog scale preference ranking of these two health states be the same as the standard gamble ranking. If the rankings are different, then it would be impossible to infer a single unique ranking, and, therefore, to assign a unique quality weight to each health state. Note that procedural invariance is different from description invariance, which states that preferences over prospects are purely a function of the probability distributions of consequences implied by the prospects and do not depend on how those given distributions are described. While procedural invariance is an assumption implicit in any conventional theory of choice, which seems natural to most economists and decision theorists and is rarely even discussed when stating formal theories of choice, in practice, this assumption fails.

Economic Violations of Procedural Invariance

One well-known phenomenon, often interpreted as a failure of procedure invariance, is the preference reversal. Reversals of preferences are observed when a so-called \$-bet (offering a high money prize with low probability) is assigned a higher selling price than a P-bet (offering a lower money prize but with a higher probability) but is subsequently not chosen in a direct choice between the two. A violation of procedure invariance is currently the prevailing explanation of this pattern of behavior. For instance, when assessing the monetary values of gambles (or delayed payments), people base their actions on a particular value system. When choosing between gambles, however, they base their actions on another value system. Therefore, preferences from choosing are different from those yielded by monetary valuation; this fact constitutes a violation of procedure invariance. The explanations invoking a violation of procedure invariance are mostly based on weighted additive models. These models assume that, both for choosing and for valuation, a person evaluates a gamble-delayed payment ($\$x$, p), where x is the expected monetary reward and p is the probability of obtaining this payoff, by using $av(x) + bw(p)$, where v and w are values for the separate attributes and $a > 0$ and $b > 0$ sum to 1 and are importance weights. It is generally assumed that the value functions v and w are the same for choosing and valuation and that only the weights a and b vary. That is, for choosing, particular weights, such as a_c and b_c, are adopted, and for valuation, other weights, such as a_v and b_v, are adopted. Note that because importance weights are positive, choosing and matching yield the same orderings over single attributes.

Preference reversal is also often attributed to response mode effects, one feature of which is scale compatibility. The compatibility hypothesis states that money is the salient attribute (or dimension of judgment) of lotteries in money valuation tasks (the two are compatible), and this renders the high prize in the \$-bet particularly influential in driving the valuation. This engenders a higher money valuation for the \$-bet than for the P-bet. In other words, in the valuation task, the participant's attention is primarily directed toward the attribute for which a matching value is to be provided—that is, the monetary dimension in our example. The participant pays less attention to the other dimension (probability), which therefore receives a lower importance weight. Therefore, the preference reversal suggests that choice and valuation tasks may depend on which attribute is made salient by the context, which invokes a different dimension of judgment to be used. Thus, the preference reversal phenomenon is a violation of procedure invariance.

Medical Violations of Procedural Invariance

In studies of health preferences, utilities for hypothetical health states cannot always be successfully measured, because of violation of procedural invariance (i.e., when the ranking of two health states varies across assessment procedures). Thus, previous research has shown that measurements of preferences can be highly sensitive to the procedures used for elicitation, including the search method for indifference points, standard gamble, visual analog scales, and numerous other factors. This is problematic for medical decision making, because using preference values based on such unsuccessful measurements may result in misinterpretation of patients' attitudes about health.

One study presented preference reversals of a more extreme nature than traditional preference reversals, because procedure invariance was systematically violated even when the preference ordering was over a single attribute (unidimensional) in both the choice and the valuation procedures. In this case, this attribute was time (life years). In an empirical study of preference assessment methods for health states, for poor health states such as metastasized breast cancer or continuous migraines, most people preferred a shorter to a longer life duration. Nevertheless, when asked to state the equivalent number of healthy years, participants as a rule demanded more healthy life years for the longer life duration than for the shorter one. Therefore, the subjects assigned the longest number of healthy years to the nonpreferred outcome. This phenomenon has serious implications for medical decision making in the context of quality-adjusted life years (QALYs), which are very important in contemporary evaluation of health policies. QALYs are based on subjective values of health states and life durations, and one of the most widespread methods of such preference assessment is the time trade-off (TTO) method. This method asks people to trade off shorter life duration in perfect health (X) against longer life years (Y) in a particular state of worse health (Q). The researcher attempts to find the indifference point between X and (Y, Q) (i.e., when the two states are equivalent). Then, according to this method, the quotient X/Y is taken as a measure, $W(Q)$, of the utility of health state Q, and (Y, Q) is estimated by a multiplicative form, $Y \times W(Q)$. The multiplication by $W(Q)$ constitutes the quality adjustment of the worse-health life duration, Y. This measurement method is based on classical rational models for respondents' answers to TTO questions. The violations of unidimensional procedure invariance provide evidence against these classical models.

While the procedural violations discussed here imply that there is no universal standard for preference measurement, one possible indicator of validity is whether an individual maintains the same rank order of his or her preferences for states across different assessment procedures. One study that compared two scaling methods, mean standard gamble and visual analog scale scores, demonstrated that individuals who satisfy procedural invariance have preferences systematically different from people who violate procedural invariance. The violators in this study appeared to have difficulty discriminating between the states and provided more random responses (note that violation of procedural invariance was not associated with age, education level, race, or gender). This result questions the validity of elicitations obtained from violators. Therefore, in addition to reporting the overall results of studies of preferences for health states, investigators should examine the results for violations of procedural invariance and report separately the results for violators and satisfiers.

Reasons for Violations

Violations of procedural invariance in medical decision making can occur for a variety of reasons. One such reason is that the experimental methods may be specifically designed to induce violations; for example, making one attribute, then another, more prominent in descriptions can trigger apparent reversals of preference. The second reason could be random error (i.e., noise in the perceptual/cognitive/motor system): Some authors suggest that elicitations that fail procedural invariance result from combinations of effects of the elicitation procedures, which result in mean values and random error. The third reason for violation of procedural invariance might be insufficiently strong preferences (i.e., when violators have weak or ill-formed preferences for hypothetical states), which would result in smaller differences in values for health states and poor correlation among ratings for health states. For example, the person may not develop strong preferences if a state is unfamiliar or poorly described. Thus, violations might be rooted in problems with health state

descriptions or with assessment procedures that interfere with the process of forming preferences. The fourth reason for violations of procedural invariance is that the individual provides, more or less, biased ratings to satisfy (conform to) the researcher's expectations (or to satisfy what he or she presumes the researcher expects to find out). Medical contexts provide the fifth reason for violations of procedural invariance because patients' cognitive abilities may be impaired: Disorders such as depression, schizophrenia, and Alzheimer's disease may impair cognition or insight such that a patient cannot perform (meaningful) health valuation tasks.

Implications

Procedure invariance and its violations appear, on the face of it, to challenge the idea that choices can, in general, be represented by any single preference function. Thus, a violation of procedural invariance does raise serious doubts about the meaningfulness of the measured utilities in terms of the individual. In this context, does a preference measurement reflect the individual's true utility, or is it merely an artifact of the assessment process? The answer to this question is still not clear, and more research is needed to address it. What is clear is that the current evidence suggests that seeing descriptions is not enough to express one's genuine preferences, and interaction with the choice options is also an essential part of the preference expression process. For example, when people are asked to decide about immediate actions (e.g., how much they need to exercise now or what tasty but unhealthy foods they need to give up) and they are presented with the long-term benefits of this costly (unpleasant) action (e.g., in terms of a longer/healthier life), then they may prefer less exercise/dieting because their focus/weight would be on their current state and the future will be discounted. Conversely, when they are asked to decide in terms of their future benefits (e.g., how much longer than normally expected they wish to live) and they only observe the necessary (implied) immediate choice of action (e.g., exercise), then a longer/healthier life, and hence more exercise/dieting, may be preferred as the attentional focus will be on the future benefits and away from the current costs. Thus, violation of procedure invariance implies that simply informing people of the future outcomes of their immediate choices is not enough to alter their behavior. The evidence presented here implies that directly manipulating a decision dimension (attribute) significantly alters decision making—most likely because the direct manipulation of a decision dimension increases the attention, or weight, given to it during judgment, and hence attention is moved away from other dimensions that are not directly manipulated (e.g., the only observed/described attributes as in the exercise/dieting vs. longevity example here). Many such violations of procedure invariance in the literature are a demonstration of the preference reversal phenomenon in real-world domains with practical importance.

Some scholars argue that adherence to the axioms of expected utility theory is not necessary for the results of utility assessment to be useful (e.g., decision models may be useful to generate insights for patients whose values and attitudes are not normative and may still help generate a better understanding of trade-offs). Nowadays it is also possible to create computer software that recognizes errors in preference measurements and helps participants resolve those errors.

Ivo Vlaev

See also Axioms; Computer-Assisted Decision Making; Errors in Clinical Reasoning; Health Status Measurement, Reliability and Internal Consistency; Patient Decision Aids; Preference Reversals; Quality-Adjusted Life Years (QALYs); Treatment Choices; Utility Assessment Techniques

Further Readings

Hornberger, J., & Lenert, L. A. (1996). Variation among quality-of-life surveys: Theory and practice. *Medical Care, 34,* 23–33.

Lenert, L. A., Cher, D. J., Goldstein, M. K., Bergen, M. R., & Garber, A. M. (1998). Effects of search procedures on utility elicitations. *Medical Decision Making, 18,* 76–83.

Lenert, L. A., & Treadwell, J. R. (1999). Effects on preferences of violations of procedural invariance. *Medical Decision Making, 19,* 473–481.

Nease, R., Jr. (1996). Do violations of the axioms of expected utility theory threaten decision analysis? *Medical Decision Making, 16,* 399–403.

Stalmeier, P. F. M., Wakker, P. P., & Bezembinder, T. G. G. (1997). Preference reversals: Violations of unidimensional procedure invariance. *Journal of Experimental Psychology, Human Perception and Performance, 23,* 1196–1205.

Starmer, C. (2000). Developments in non-expected utility theory: The hunt for a descriptive theory of choice under risk. *Journal of Economic Literature, 38,* 332–383.

Viscusi, W. K., & Evans, W. N. (1990). Utility functions that depend on health status: Estimates and economic implications. *American Economic Review, 80,* 353–374.

Propensity Scores

Propensity score analysis is a statistical method used in nonrandomized studies to compare treatments after removing selection bias due to observed baseline covariables. To the extent that all true confounders are observed and selection bias removed, unbiased estimation of the average treatment effect can be made.

Propensity score analysis is a two-stage process. First, a logistic regression model predicting treatment assignment from available baseline potential confounders is used to assign each patient a propensity score representing the probability that he or she would receive treatment (vs. control, no treatment), regardless of whether treatment is actually received. Second, treated and nontreated patients are compared on the outcome of interest after conditioning on the propensity scores—typically through stratification or matching.

Propensity score analysis is often preferred over traditional multivariable modeling to adjust for confounding because it makes less model assumptions, allows visualization of covariable balance after matching on the propensity score, and has stronger theoretical properties for estimation of causal effects. A key limitation is that all true confounding variables are usually not available or known. Therefore, caution must be used in making a causal inference.

Methods

Motivation for Propensity Score Analysis

In a *randomized study*, a simple comparison of the randomized groups on the outcome gives an unbiased estimate of the average causal effect of treatment versus control because there is no selection bias—patients are randomly assigned to receive either treatment or control. Randomized groups are well-balanced on both observed and unobserved potential confounders, so that the observed difference in means or proportions between groups estimates the average causal effect, or average treatment effect, for individuals.

In a *nonrandomized study*, treatments are not randomly assigned, and so confounding through selection bias is a major impediment to assessing the true treatment effect. Groups to be compared will differ systematically on confounding variables, which by definition distort the relationship of interest because they are related to both the treatment and the outcome. Unadjusted comparisons of nonrandomized groups will give biased estimates of average treatment effect.

Traditionally, adjustment for confounding in nonrandomized studies was done through multivariable regression of an outcome on an exposure while adjusting for potential confounders in the model and in a case-control study by comparing cases and controls matched on important confounders. Major limitations of multivariable regression are the difficult-to-assess assumptions of the model, inability to adjust for enough variables and interactions, and inability to visualize how well selection bias has been removed. Case-control matching on numerous variables becomes logistically impossible, limiting the amount of confounding that can be removed.

Propensity score analysis is a more efficient way to remove selection bias, and thus confounding. It is preferred when the number of outcomes is relatively small, when there are a host of potential confounders to adjust for, or when there is insufficient overlap between the treated and nontreated on some confounders. Causal inference is theoretically possible with either propensity score or multivariable analysis, but it is easier with propensity score analysis.

The propensity score is a balancing score, such that treatment assignment is independent of the included potential confounders at any score value. For any particular propensity score, the treated and nontreated should *on average* have very similar distributions of each covariable used to create the score. The goal of propensity score analysis is thus to estimate the average treatment effect by

comparing the treated and nontreated on outcome at similar propensity score values.

Creating Propensity Scores

Propensity score analysis is a two-stage procedure: First, the propensity scores are estimated, ignoring the outcome (this section). Second, the scores are used to create balanced groups and compare groups on outcome (next section).

Creating the Propensity Score Model

A propensity score is the predicted probability of receiving treatment (vs. no treatment) given a set of covariables X, regardless of whether or not treatment was actually received. Propensity scores are calculated by first fitting a logistic regression model predicting treatment versus nontreatment from available baseline covariables. Patient "logit" scores are calculated by plugging each patient's covariable values into the regression equation. Logit scores are transformed into propensity scores with range 0 to 1 as follows: Propensity score for the ith patient = $\exp^{\text{logit}(x_i)}/(1+\exp^{\text{logit}(x_i)})$.

Form of Propensity Score Model

Since the goal of creating the propensity score model is to create individual scores that are used to balance the treated and nontreated on confounders, the propensity model is evaluated by assessing how well this balance is achieved. The goal is *not* to create a parsimonious model (say, $p < .05$ only) but rather to achieve the best possible balance between the treated and nontreated on all confounders. The model should have liberal entry criteria and consider interactions among the covariables, with little regard to traditional model checking.

Overlap

Side-by-side histograms of the estimated propensity scores for the treated and nontreated can help decide if all patients should be included in assessment of the treatment effect. For example, if there is no overlap between groups at extremes of the propensity score distribution, the corresponding patients should be excluded. This limits generalizability but also indicates an advantage over multivariable regression. Overlap can be increased by focusing on confounders strongly related to outcome rather than to treatment.

Using the Propensity Score in Assessing Treatment Effect

After creating the propensity score model, the scores are used to create a balance between the treated and nontreated on the potentially confounding variables, and then the groups are compared on outcome. Balance is attained through stratification on the propensity score, matching, or direct adjustment.

Assessing the Propensity Score Model

A prerequisite to making an inference on average treatment effect using the methods described below is good balance between treated and nontreated on the available confounders, overall and within propensity score levels (say, quintiles). A good metric to assess balance is the absolute standardized difference, the absolute value of the difference of means between treated and nontreated divided by the pooled standardized deviation. Absolute standardized difference is preferred to the p value because it is not affected by differential sample size between the unmatched and matched. If clinically important imbalance on covariables is detected, the propensity score model should be refit, adding more covariables, interactions, or polynomial terms until uniform balance is achieved. This is done before assessing the treatment effect.

Stratification, Matching, and Direct Adjustment

In stratification (or subclassification), patients are first placed in strata defined by, typically, quintiles of the propensity score. Average treatment effect is assessed by comparing the treated and nontreated on outcome within quintile and aggregating the results in a stratified analysis. Theoretical results indicate that comparison of the treated and nontreated via stratification on propensity score quintiles removes 90% of the confounding due to each variable.

In matching, the treated and nontreated are matched on the propensity score or logit score to within a prespecified distance criterion (e.g., .05 propensity score units). Matching on the propensity score should result in treated and nontreated groups with very similar means or proportions for

each covariable used in creating the scores, even though individual matched patients may differ on many covariables.

Greedy matching sequentially matches each treated patient to the closest matching control(s). Pairs are set aside as they are matched, so the method does not consider all possible matches. Greedy matching usually results in a substantially reduced data set because all treated and nontreated may not have matches; or perhaps there are many more nontreated than treated, but only 1-1 or 1-2 matching is desired. Matched patients are then compared on outcome using an appropriate test.

Optimal matching is theoretically superior to greedy matching because it minimizes the sum of all distances between treated and nontreated; it is optimal because it finds the best possible matches among the treated and nontreated. However, if all available treated and nontreated patients are used, as opposed to an equal number of each, the resulting matching may not be much superior to greedy matching. Average treatment effect is assessed by comparing groups on outcome using conditional or stratified analysis on the matched clusters.

In direct adjustment, the propensity score is simply included as a covariable in the traditional multivariable model of all patients assessing the relationship between treatment and outcome. It is less reliable in removing selection bias than is stratification or matching.

Residual Confounding

Residual confounding is expected in all the above methods because the observed propensity score is only an estimate of a patient's true pro-

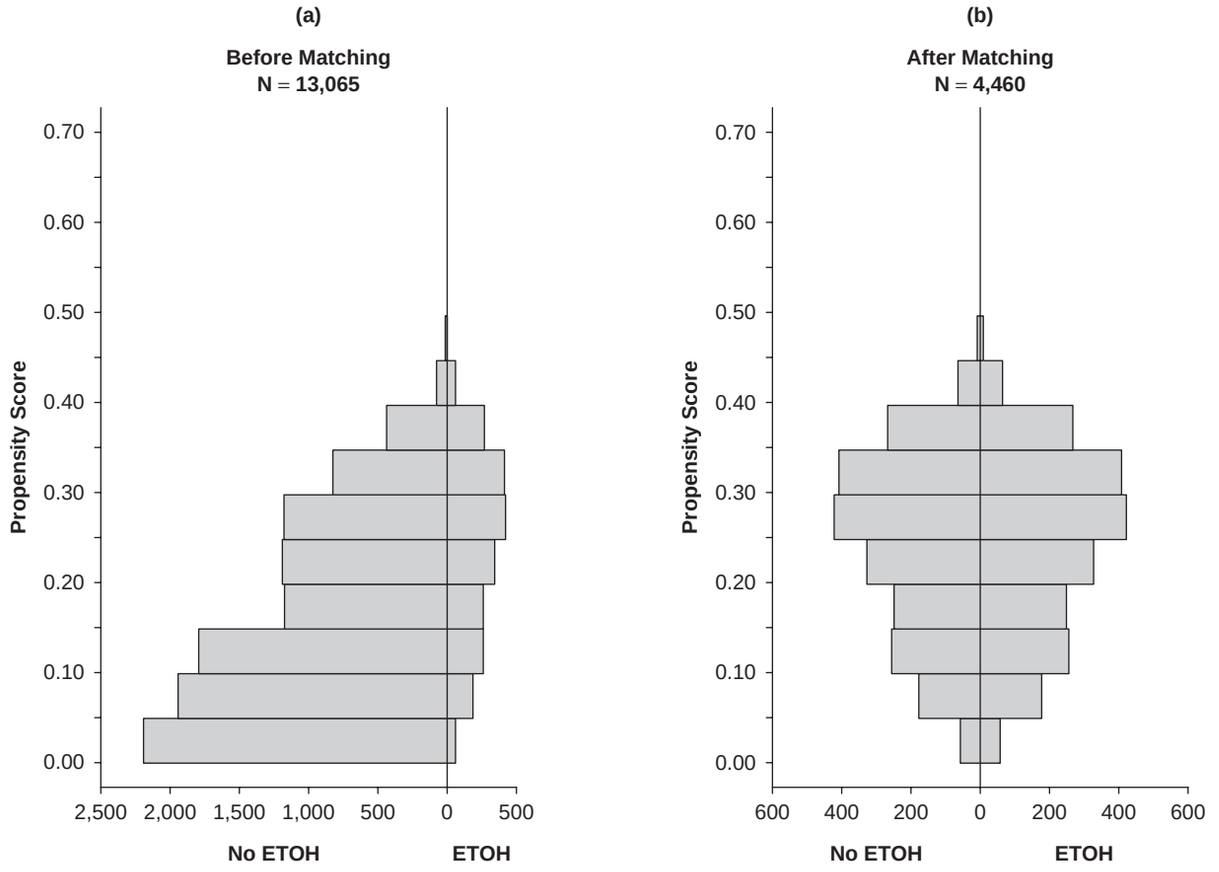


Figure 1 Propensity score distribution before and after matching: (a) Propensity scores before matching; overlapping propensity score distributions before matching and (b) Propensity scores after matching; nearly identical propensity score distribution after matching

Source: Maheshwari et al. (2008). Used with permission.

pensity score. Therefore, the observed propensity score only balances observed covariables between groups *in expectation*—that is, across repeated sampling. Also, exact matches will not likely be found. For these reasons, any covariables that are still unbalanced between the matched groups or are significantly related to outcome in the presence of the treatment group should be added to any model assessing the average treatment effect.

Table 1 Covariable balance before and after matching: Selected factors

	All Patients N = 13,065			*Matched Patients N = 4,460*		
	Drinks/Week			*Drinks/Week*		
# Factor	*0–2 (N = 10,830)*	*3+ (N = 2,235)*	*P Value*[+]	*0–2 (N = 2,230)*	*3+ (N = 2,230)*	*P Value*[+]
Categorical factors: Data are *N* (%)						
1 Female gender[a]	3,435 (32)	201 (9)	<0.001	199 (9)	200 (9)	0.96
2 Smoker[a]	6,219 (57)	1,842 (82)	<0.001	1,809 (81)	1,837 (82)	0.28
3 Caucasian race[a]	9,163 (85)	1,954 (87)	<0.001	1,962 (88)	1,949 (87)	0.55
7 COPD/Asthma[b]	1,074 (10)	210 (9)	0.45	206 (9)	210 (9)	0.84
16 Ventricular tachycardia[a]	250 (2)	80 (4)	<0.001	70 (3)	79 (4)	0.45
32 Preop Diabetes Agent[a]	1,781 (17)	251 (11)	<0.001	246 (11)	251 (11)	0.81
33 Preop Statins[a]	3,563 (33)	563 (25)	<0.001	536 (24)	562 (25)	0.37
Continuous factors: Data are mean (*SD*) or median [quartiles]						
34 Age (yr)[a]	67 (11)	65 (11)	<0.001	65 (11)	65 (11)	0.27
36 Weight (lb)[a]	183 (40)	192 (36)	<0.001	191 (36)	192 (36)	0.89
37 BMI (kg/m^2)[a]	29 (6)	29 (5)	0.04	29 (5)	29 (5)	0.90
38 Hematocrit (%)[c]	40 (5)	41 (5)	<0.001	41 (5)	41 (5)	0.29
39 Creatinine[c]	1.1 (0.5)	1.1 (0.4)	0.02	1.1 (0.4)	1.1 (0.4)	0.56
40 CPB Time (m)[b]	91 (50)	91 (48)	0.99	91 (49)	91 (48)	0.95
42 BUN (mg/dL)[a,b]	18 [15, 24]	17 [14, 21]	<0.001	17 [14, 22]	17 [14, 21]	0.73
46 Operative RBC units[b]	0 [0, 1]	0 [0, 0]	<0.001	0 [0, 0]	0 [0, 0]	0.84

Source: Maheshwari et al. (2008). Used with permission.

Design variable number.

[+]*P* values from Pearson's chi-squared test for categorical factors and *t* test or Wilcoxon's Rank Sum Test for continuous.

a. Used in propensity score matching.

b. Included in outcome model.

c. Included in neither propensity score model nor outcome model.

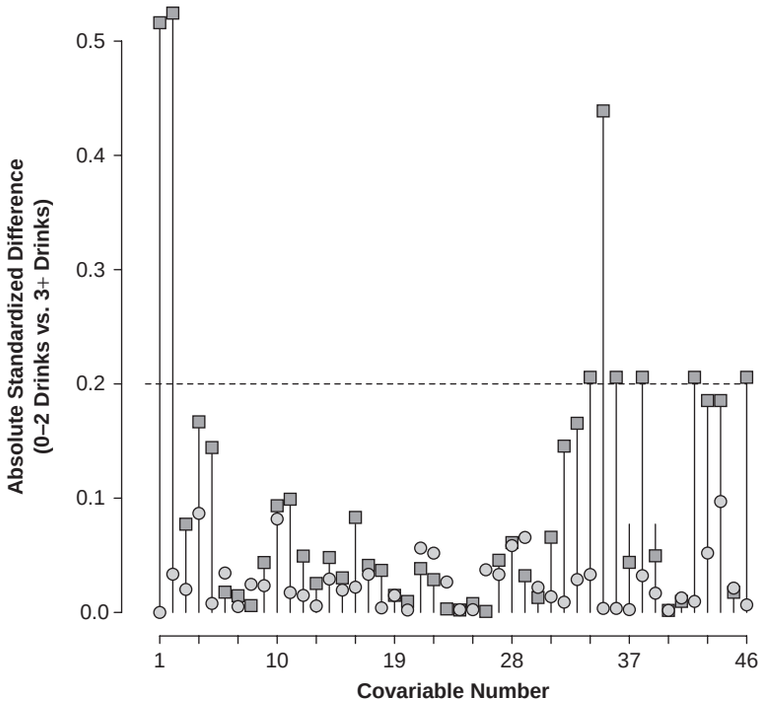


Figure 2 Absolute standardized difference before and after matching

Source: Maheshwari et al. (2008). Used with permission.

Strongly Ignorable Treatment Assignment and Sensitivity Analysis

Propensity score analysis can only adjust for available covariables. However, to validly make causal inference using propensity score analysis, one must assume strongly ignorable treatment assignment (SITA). SITA means that the propensity score analysis has removed all selection bias due to both observed and unobserved variables, making treatment assignment z at any covariate vector value x independent of the underlying potential outcomes r_1 and r_0, as in a randomized study: $(r_1, r_0) \perp\!\!\!\perp z|x$. The SITA assumption should be assessed by a sensitivity analysis of how strongly a hypothesized unobserved covariable would have to be related to both treatment and outcome to substantially alter the estimated average treatment effect or change the study conclusions.

Data Application

In a database review of 13,065 Cleveland Clinic patients undergoing elective coronary artery bypass graft (CABG) surgery, 2,235 were self-described alcohol drinkers (ETOH [ethyl alcohol], ≥ 3 drinks/week), and 10,830 were mild drinkers/non-drinkers (non-ETOH, < 3 drinks/week). Researchers were interested in the effect of ETOH versus non-ETOH on a binary (yes/no) postoperative complication outcome. Propensity score was used to assess the association since confounding was a large concern. First, logistic regression was used to create propensity scores for each patient with ETOH (1 = *ETOH*, 0 = *non-ETOH*) as outcome and independent variables chosen from 46 preoperative and operative variables plus their interactions and polynomial terms (inclusion criterion $p < .35$). The propensity score model thus included 14 covariables and 5 pairwise interactions.

Greedy matching on the propensity score resulted in 2,230 ETOH matched to 2,230 non-ETOH using a propensity score distance criterion of .05. Figure 1a shows the substantial overlap between the ETOH and non-ETOH patients on propensity score distribution before matching, while Figure 1b shows the nearly identical propensity score distributions after matching.

Table 1 shows balance before and after greedy matching on 15 representative covariables from the 46 considered for the propensity score model. Variables are marked as being included in the propensity score model, the outcomes model (see below), or neither. Note that a particular variable not being included in the propensity score model does not mean that imbalance will occur for that factor; balance may result if that variable is correlated with another that was included.

Figure 2 shows the reduction in absolute standardized difference from before to after matching in nearly all variables.

Logistic regression was used to compare the matched ETOH and non-ETOH patients on complications (1 = *Any*, 0 = *None*) as the outcome. Baseline confounders that were strong independent predictors of outcome (six variables, $p < .05$) or insufficiently matched (two variables: diabetes mellitus, BUN [blood urea nitrogen]) were included in the model. The estimated odds ratio (95% CI [confidence interval]) of complications for ETOH versus non-ETOH was 1.13 (.95, 1.35). In a stratified analysis on propensity score quintiles using all 13,065 patients, results were very similar to the greedy-matching approach, with a Mantel-Haenszel odds ratio of 1.11 (.90, 1.37). Both approaches differed markedly from a crude analysis of the data ignoring confounders, with odds ratio of .80 (.66, .97).

Ed Mascha

See also Causal Inference and Diagrams; Causal Inference in Medical Decision Making; Confounding and Effect Modulation; Randomized Clinical Trials

Further Readings

Blackstone, E. (2002). Comparing apples and oranges. *Journal of Thoracic and Cardiovascular Surgery, 123,* 8–15.

Glynn, R., Schneeweiss, S., & Sturmer, T. (2006). Indications for propensity scores and review of their use in pharmacoepidemiology. *Basic & Clinical Pharmacology & Toxicology, 98*(3), 253–259.

Maheshwari, A., Dalton, J., Mascha, E., Bakri, M., Yared, J. P., Kurz, A., et al. (2008). *The association between alcohol consumption and morbidity and mortality in patients having coronary artery bypass surgery*. Manuscript submitted for publication.

Rosenbaum, P. (1991). A characterization of optimal designs for observational studies. *Journal of the Royal Statistical Society, Series B (Methodological), 53*(3), 597–610.

Rosenbaum, P. R. (2002). *Observational studies* (2nd ed.). New York: Springer Verlag.

Rosenbaum, P., & Rubin, D. (1983). Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcome. *Journal of the Royal Statistical Society, Series B (Methodological), 45*(2), 212–218.

Rosenbaum, P., & Rubin, D. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika, 70*(1), 41–55.

Rosenbaum, P., & Rubin, D. (1984). Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association, 79,* 387, 516–524.

Prospect Theory

Prospect theory is a descriptive theory of risky choice proposed by the psychologists Daniel Kahneman and Amos Tversky in 1979 and revised in 1992. Like subjective expected utility theory (SEU), prospect theory is a probability-weighted additive utility model: Risky options are evaluated by adding up the utilities of the possible outcomes after weighting each by a factor that reflects the decision maker's belief about how likely each outcome is to occur. Like SEU, prospect theory can be mathematically derived from a simpler set of axioms about preference judgments and can provide predictions about choice behavior. Prospect theory, however, has several key differences from SEU that enable it to better account for observed systematic violations of axioms and predictions of SEU that were widely documented by researchers in the 1960s and 1970s (e.g., people's willingness to simultaneously pay too much to insure against small probabilities of large losses—a risk-averse choice—and pay too much to purchase lottery tickets that offer small probabilities of large prizes—a risk-seeking choice).

Key Features

First, prospect theory posits that decision makers initially edit the decision stimulus in some fashion and then evaluate the edited stimulus rather than the original stimulus. For example, decision

makers may combine or separate decision outcomes, reframe outcomes in psychologically different terms or contexts, or cancel out common outcomes in a choice set.

Second, prospect theory assumes that decision makers evaluate the values of outcomes relative to a reference point. That is, rather than evaluating outcomes in absolute terms (e.g., a final health state that will result from this decision), outcomes are evaluated in relative terms, as gains or losses (e.g., how much better or worse the final health state will be than the current health state). Prospect theory is thus a weighted average over the potential gains and losses from the decision, relative to a reference point, often the status quo or current state. An implication, therefore, is that decision makers evaluating the same prospects (with the same absolute outcomes) with regard to different reference points may have quite different evaluations.

Third, prospect theory specifies a mathematical form for the value function that maps gains and losses to values, as illustrated in the right panel of Figure 1. The prospect theory value function is concave for gains and convex for losses and predicts diminishing marginal value for additional gains and losses: The gain (loss) of \$100 is less valuable (painful) to a person who simultaneously gains (loses) \$1,000 than to a person who simultaneously gains (loses) \$1. The value function is also steeper for losses than for gains, reflecting the psychological phenomenon of loss aversion: It is more painful to lose \$100 than pleasurable to gain \$100. Kahneman and Tversky proposed that the value functions for gains and losses are power functions of the form x^{α} for gains and $-\lambda x^{\beta}$ for losses and found median values of $\alpha = .88$ and $\beta = .88$ (implying moderate curvature in both gain and loss value functions) and $\lambda = 2.25$ (implying that losses have more than twice the impact of gains).

Fourth, prospect theory introduces a transformation of the outcome probabilities into decision weights. The decision-weighting function, shown in the left panel of Figure 1, displays several notable features. In general, small probabilities are psychologically overweighted, and large probabilities are psychologically underweighted. Because the decision weights associated with impossibility and certainty are fixed (at 0 and 1, respectively), the function is discontinuous near these extremes. This probability distortion predicts phenomena such as certainty effects (larger subjective differences between 99% and 100% than between 98% and 99%).

Cumulative Prospect Theory

The 1992 cumulative prospect theory formulation differs from the original 1979 prospect theory by specifically positing separate rank-dependent probability weighting functions for gains and losses. This insight also forms the basis for R. Duncan Luce and Peter C. Fishburn's rank- and sign-dependent utility theory.

As an illustration, consider a gamble with four outcomes: (1) a 10% chance of losing \$50, (2) a 70% chance of gaining \$5, (3) a 10% chance of gaining \$10, and (4) a 10% chance of gaining \$20. Under original prospect theory, each of the probabilities would be transformed to a decision weight using a function like that depicted in the left panel of Figure 1; the decision weights for the \$50 loss and the \$10 gain would be the same. Under cumulative prospect theory, however, the decision weights are computed separately for gains and losses, and within each domain, the weight for each outcome depends on the rank of the outcome within the set. For example, the weight for the \$10 gain and the \$20 gain will differ because the \$20 gain is the highest-ranked gain and the \$10 gain is a less highly ranked gain.

For each gain outcome, the decision weight is the difference between a transformation of the cumulative gain probabilities of the outcome and all smaller gain outcomes and a transformation of the cumulative gain probabilities of only the smaller gain outcomes. A similar process is applied to loss outcomes. Mathematically, the decision weight π_i^+ for an outcome G_i with probability p_i in the domain of gains is

$$\pi_i^+ = w^+(p_i + p_{i+1} + \ldots + p_n) - w^+(p_{i+1} + \ldots + p_n),$$

where $p_0, \ldots, p_i, \ldots, p_n$ are the probabilities of outcomes $G_0, \ldots, G_i, \ldots, G_n$ and $0 \le G_0 \le \ldots \le G_i \le \ldots G_n$. For example, in a prospect with gain outcomes \$5, \$10, and \$20, with probabilities .7, .1, and .1, respectively, the decision weight associated with the \$5 outcome is $w^+(.7 + .1 + .1) - w^+(.1 + .1)$, and the decision weight associated with the \$10 outcome is

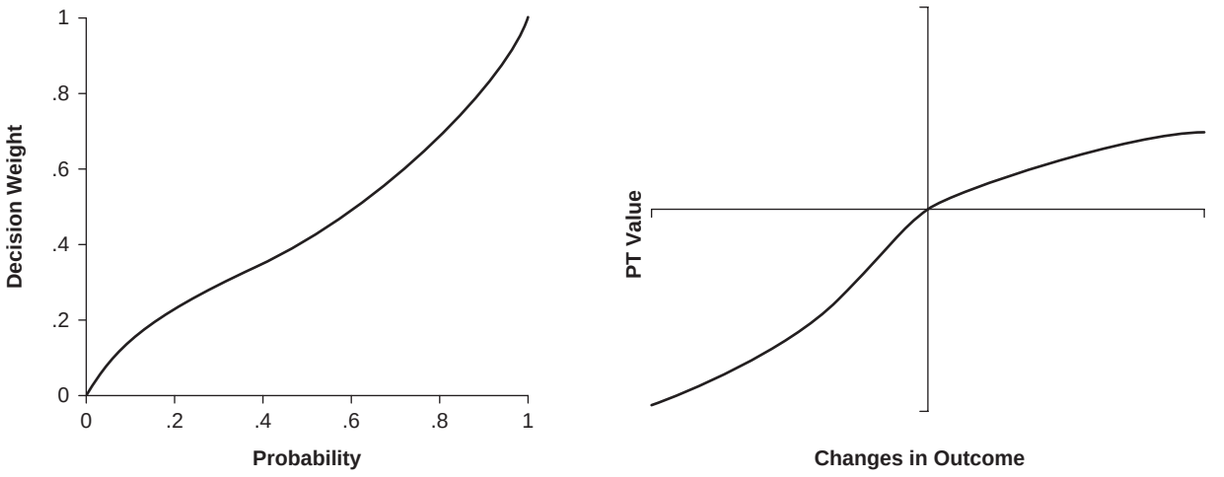


Figure 1 Prospect theory (PT) value function and decision-weighting function

$w^+(.1 + .1) - w^+(.1)$. The decision weight associated with the largest gain, G_n, is simply $w^+(p_n)$; in the example, the decision weight for the \$20 outcome is $w^+(.1)$.

Similarly, for losses, the decision weight π_i^- for a loss outcome L_i with probability p_i is

$$\pi_i^- = w^-(p_0 + p_1 + \ldots + p_i) - w^-(p_0 + p_1 + \ldots + p_{i-1}),$$

where p_0, p_1,…, p_i are the probabilities of outcomes L_0, L_1,…, L_i and $L_0 \leq L_1 \leq \ldots \leq L_i \leq 0$. Again, the decision weight associated with the greatest loss, L_0, is simply $w^-(p_0)$.

Empirical observations suggest that the form of each of the gain and loss decision-weighting functions (w^+, w^-) is well characterized by the following transformations:

$$w^+(p) = p^\gamma/[p^\gamma + (1 - p)^\gamma]^{1/\gamma},$$

$$w^-(p) = p^\delta/[p^\delta + (1 - p)^\delta]^{1/\delta}.$$

Kahneman and Tversky reported median values of $\gamma = .61$ and $\delta = .69$.

Criticisms

Although prospect theory has proven a robust descriptive theory of risky choice, some criticisms have been identified in laboratory settings. Little research has established the conditions under which particular editing operations are applicable; as a result, it can be difficult to form unique predictions for particular choice sets. Configural weight utility models may account for patterns of violation and satisfaction of independence axioms better than weighted additive utility models such as prospect theory. Tests of generic utility theory, of which prospect theory and cumulative prospect theory are special cases, have demonstrated violations of predicted invariance across contexts, which limits the applicability of such theories.

Impact

The impact of prospect theory has been substantial and broad. In addition to its position as the leading psychological descriptive utility theory, it has been a key impetus to the development of behavior economics. Indeed, Kahneman was recognized with the Nobel Prize in Economics in 2002, in part specifically for the contribution of prospect theory to the field of economics. (Tversky predeceased the awarding of the prize but was also mentioned by the Nobel Foundation in this context.)

In medical decision making, prospect theory has been studied as a descriptive utility theory and proposed as a correction to utility assessment procedures. Jonathan Treadwell and Leslie Lenert reviewed the empirical evidence for prospect theory in health decisions, which was largely but not

entirely supportive. They also discussed its implications for medical cost-effectiveness analyses, pointing out that when prescriptive analyses use behavioral measures of utility based on risky choice (e.g., standard gamble utility assessments), the choices made by respondents are subject to the value and probability weightings described by prospect theory. Accordingly, it may be desirable to correct assessed utilities to account for these cognitive processes before applying them to prescriptive analyses.

Alan Schwartz

See also Decision Weights; Editing, Segregation of Prospects; Expected Utility Theory; Rank-Dependent Utility Theory; Subjective Expected Utility Theory; Value Functions in Domains of Gains and Losses

Further Readings

Birnbaum, M. H. (2005). Three new tests of independence that differentiate models of risky decision making. *Management Science, 51*(9), 1346.

Birnbaum, M. H., Patton, J. N., & Lott, M. K. (1999). Evidence against rank-dependent utility theories: Tests of cumulative independence, interval independence, stochastic dominance, and transitivity. *Organizational Behavior and Human Decision Processes, 77*(1), 44–83.

Bleichrodt, H., & Pinto, J. L. (2000). A parameter-free elicitation of the probability weighting function in medical decision analysis. *Management Science, 46*(11), 1485–1496.

Bleichrodt, H., Pinto, J. L., & Wakker, P. P. (2001). Making descriptive use of prospect theory to improve the prescriptive use of expected utility. *Management Science, 47*(11), 1498–1514.

Chechile, R., & Cooke, A. (1997). An experimental test of a general class of utility models: Evidence for context dependency. *Journal of Risk and Uncertainty, 14*(1), 75–93.

Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica, 47,* 263–291.

Luce, R. D., & Fishburn, P. C. (1991). Rank- and sign-dependent linear utility models for finite first-order gambles. *Journal of Risk and Uncertainty, 4,* 29–59.

Treadwell, J. R., & Lenert, L. A. (1999). Health values and prospect theory. *Medical Decision Making, 19*(3), 344.

Tversky, A., & Kahneman, D. (1992). Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and Uncertainty, 5*(4), 297–323.

Protected Values

Protected values or other very similar notions, such as sacred values, taboo values, and moral mandates, are values people preclude from trade-offs with other values, particularly secular values. Protected values refer to any concrete or abstract entity (e.g., human beings, animals, dignity, health, love, honor, honesty, human rights) an individual or a community considers as infinitely significant, not substitutable and inviolable, and therefore as nontradable and noncompensatory. For instance, many people think that it is morally wrong to sacrifice human or animal lives in return for monetary benefits or to have a free market of transplant organs. Trading them off would offend deeply held beliefs and undercut people's self-image as moral beings or threaten a way of life.

An important foundation for research on this issue has been laid by Philip Tetlock and colleagues. They have examined how people respond to and cope with violations of protected values or sacred values. Other research efforts, mainly influenced by Ilana Ritov and Jonathan Baron, have focused more on how such values affect decision-making processes.

Relevance for Medical Decision Making

The phenomenon of protected values is interesting for theoretical and practical reasons. From a theoretical perspective, protected values create a problem for utilitarian theories. Such models presuppose that any value can be traded off for any other value. When protected values are involved, however, trading off such values is precluded. Researchers have recognized trade-off reluctance as a common problem when using contingent valuation methods to quantify values or public goods. For example, people often reject decisions and refuse "putting a price on life" by trading off life with monetary expenses. It seems that if goods or services tap ethical issues or reflect a protected

value, placing a monetary value on this specific value is difficult yet unacceptable.

From a practical perspective, protected values are highly relevant for medical decision making. Ethical reservations or strong protected values are very likely involved in fields such as euthanasia, prolongation of life, abortion, prenatal diagnosis, vaccination, organ transplantation, genetic therapy, and cloning. Given that such decisions tap into strong beliefs and moral commitments, they are highly emotion laden. Research suggests that people can respond with strong outrage to threats to protected values and socially distance themselves from potential violators of taboo trade-offs. Furthermore, interactions between physicians and patients can become exceedingly difficult and conflicting when protected values are involved on one or both sides.

Research and Current Findings

Omission and Action Tendencies

Previous research suggests that protected values are often linked with *deontological principles*. Deontology refers to duty and is usually contrasted with consequentialism. The distinctive idea of deontological reasoning is that the focus is on the inherent rightness or wrongness of an act per se rather than on the magnitude of the consequences (the consequentialist perspective). Deontological principles reflect morally mandated actions or omissions, such as the duty to keep promises or the duty not to lie. Such duties can be religious in nature (e.g., the Ten Commandments), socially contracted (e.g., the Hippocratic Oath, human rights), or intuitive (e.g., to do no harm).

Indeed, some researchers have shown that protected values sometimes reflect prohibition rules (do-no-harm rules), which elicit an *omission bias*—that is, the tendency to favor omissions over otherwise equivalent, or even better, actions. One typical paradigm used to explore such issues has provided people with a choice between a harmful act and a harmful omission. Usually, studies also ask participants for a threshold at which the act becomes more desirable than the omission. For example, in a scenario often used, participants were faced with a flu epidemic that is expected to kill 10 out of 10,000 children and a vaccine that can cause death due to side effects. When asked about the highest amount of harm from the vaccine at which they would prefer to take action (i.e., to vaccinate), most subjects demanded that the vaccine risk be smaller than the disease risk. There were also people who would not accept any death from vaccination. In general, trade-off reluctance and omission bias were greater for people endorsing protected values than for people without protected values (as assessed independently). Apparently, actively killing someone with an act (i.e., vaccinating) was worse than allowing someone to die (i.e., not vaccinating). Such patterns suggest a violation of consequentialist thinking.

Other studies have emphasized that protected values are likely to shape a tendency to prefer action to inaction. They suggest that protected values may be an essential source of action motivation that is mobilizing people. The phenomenon of activism may be seen as providing strong examples of action tendencies, reflecting strong commitments to values such as human rights, nature, or human life and a duty to act. Decisions involving such issues are driven not by anticipated consequences but rather by a sense that the act is "the right thing to do." Moreover, the moral intuition that doing nothing appears to be morally condemnable is common in everyday life. The medical domain provides further examples. For example, a physician facing a patient suffering cannot just stand by and do nothing, though the physician may be aware that any further treatment would be unsuccessful or even harmful.

Resistance to Situational Influences

Application of deontological rules suggests that decision making is not based on the consequences of the alternatives involved but on applying rules about what is in principle right or wrong. Consistent with this, previous research found some support that people holding protected values are less sensitive to the magnitude of consequences than people without protected values (killing 1 child is as bad as killing 100 children; 1 abortion is as bad as 10 abortions).

In addition, and also in line with the deontological focus, research has found that protected values result in stronger resistance to situational influences, such as framing or monetary incentives.

Over the past decades, framing effects have been reported that refer to the finding that different but otherwise equivalent descriptions of choice problems give rise to preference reversals. The classic example of framing effects by Amos Tversky and Daniel Kahneman (1981) involves the Asian disease problem, in which participants are told that an outbreak of disease threatens to kill 600 people. People are asked to choose between a certain option and a risky option having the same expected value. In the positive frame, the outcomes are described in terms of the number of lives saved and in the negative frame in terms of lives lost. The common finding is that participants tend to choose the certain option when the problem is framed in terms of gains but choose the risky option when the problem is framed in terms of losses. Such a finding is consistent with a consequentialist perspective, which suggests that framing causes people to view the outcomes as gains or losses. A deontological perspective, however, implies that consequences matter less. We, therefore, would expect that people with a deontological focus should be more insensitive to framing. That is, they should show no difference in preferences for risky or certain options under different framing conditions.

Given that protected values seem to derive from deontological thinking, recent research using similar life-and-death scenarios tested the idea that protected values increase attention to acts versus omissions and decrease attention to outcomes. The results have suggested that people who indicated protected values for the issue described in the decision task were more likely to prefer acts over omissions. Importantly, they were also immune to the framing of the outcomes. Notably, it was not relevant for people with protected values whether the alternatives were associated with gains or losses, and risky or certain outcomes. This conclusion is also consistent with previous literature arguing that protected values are associated with deontological rules. Overall, the studies suggest that for people holding protected values, adherence to principle is important, whereas the magnitude of consequences matters less.

Reactions to Taboo Versus Tragic Trade-Offs

Another research direction has focused on reactions to trade-offs. This research has shown that people struggle to protect such values from trade-offs and respond with strong moral outrage when faced with actual violations of *taboo trade-offs* (i.e., situations that pit protected values against secular values, such as lives vs. money). Studies by Tetlock and colleagues compared people's reactions to routine trade-offs (e.g., paying someone to clean) with their reactions to taboo trade-offs (e.g., selling human body parts). The results have shown that people express intense cognitive reactions (such as harsh attributions to norm violators), emotional reactions (such as expressing anger, disgust, and contempt), behavioral reactions (such as intentions to punish violators), and moral cleansing (such as attempts to reaffirm one's own moral worthiness) when confronted with taboo trade-offs.

Despite decisions involving protected values being often negatively emotion laden and capable of triggering harsh reactions, recent research also suggests that protected values may work as an important decision tool. Protected values can help *facilitate* decisions because people can then rely on values that are precluded from trade-offs. People faced with a decision task with which they associate protected values (i.e., a taboo trade-off) need less time to make the decision and perceive the task as easier to solve than people faced with decisions that are not linked to any protected values. In contrast, people need much more time to make a decision and perceive the problem as most difficult to solve when the task reflects a *tragic trade-off* (i.e., situations that pit two protected values against each other, such as one life vs. another). For instance, imagine a situation where a hospital has to decide which of two desperately ill people should get the one liver that is currently available due to a shortage of organ donors. Such situations are not only perceived as more difficult but also as emotionally stressful.

Additional Issues

It may be important to emphasize that people should not be portrayed as absolute defenders of protected values. On the contrary, there are hints that people who claim to have protected values nevertheless would consider trade-offs and compromises under certain circumstances. It seems to be a dynamic process when people treat a trade-off as taboo or when they do not. In addition, there is

very probably vast variation in what individuals, groups, and cultures hold protected and sacred. This applies also to the context of medical decision making, where experts (physicians) and lay people (patients) are likely to have different or differently strong protected values. Overall, it seems essential to acknowledge the existence and reality of protected values, because they are an important source of conflict in decision making and interpersonal interactions.

Carmen Tanner and Daniel Hausmann

See also Decisional Conflict; Irrational Persistence in Belief; Monetary Value; Moral Choice and Public Policy; Mortality; Person Trade-Off

Further Readings

Baron, J., & Spranca, M. (1997). Protected values. *Organizational Behavior and Human Decision Processes, 70,* 1–16.

Ritov, I., & Baron, J. (1990). Reluctance to vaccinate: Omission bias and ambiguity. *Journal of Behavioral Decision Making, 3,* 263–277.

Ritov, I., & Baron, J. (1999). Protected values and omission bias. *Organizational Behavior and Human Decision Processes, 79,* 79–94.

Skitka, L. J., Bauman, C. W., & Sargis, E. G. (2005). Moral conviction: Another contributor to attitude strength or something more? *Journal of Personality and Social Psychology, 6,* 895–917.

Tanner, C., & Medin, D. L. (2004). Protected values: No omission bias and no framing effects. *Psychonomic Bulletin & Review, 11,* 185–191.

Tanner, C., Medin, D. L., & Iliev, R. (2008). Influence of deontological vs. consequentialist orientations on act choices and framing effects: When principles are more important than consequences. *European Journal of Social Psychology, 38,* 757–769.

Tetlock, P. E. (2003). Thinking the unthinkable: Sacred values and taboo cognitions. *Trends in Cognitive Sciences, 7,* 320–324.

Tetlock, P. E., Kristel, O. V., Elson, S. B., Green, M., & Lerner, J. S. (2000). The psychology of the unthinkable: Taboo trade-offs, forbidden base rates, and heretical counterfactuals. *Journal of Personality & Social Psychology, 5,* 853–870.

Tversky, A., & Kahneman, D. (1981). The framing of decisions and the psychology of choice. *Science, 211,* 453–458.

Qualitative Methods

Qualitative methods are a form of scientific inquiry involving in-depth examination of phenomena through the collection and analysis of textual data or observation and recording of events. Qualitative methods are relevant to the field of medical decision making for a variety of reasons. First, qualitative methods are useful when studying emerging topics. This is well suited for medical decision making where new questions are constantly raised as a consequence of the rapid evolution and interplay of science, medicine, and healthcare. Second, qualitative methods can help explore how people attach meaning to decisions, how decisions unfold, and the context in which decisions are made. Also, medical decision making often is concerned with subjective concepts such as values, perceptions, feelings, and attitudes. Qualitative methods are well equipped to examine and understand these areas by allowing respondents to express in their own words their unique experiences.

Qualitative methods encompass various modes of data collection (i.e., semistructured interviews, focus groups, narratives, and observation), analytic techniques (i.e., grounded theory, deconstruction, narrative, content, and discourse analysis), and presentation formats. Other considerations when using qualitative methods include sample size determinations, interrater reliability, and the availability of software to help manage the process. Qualitative methods are increasingly used in combination with quantitative methods in approaches called mixed methods. Despite the many benefits of using qualitative methods, there are drawbacks, including increased time and the lack of generalizability.

Purpose and Goals of Qualitative Methods

Qualitative methods have unique goals and objectives when compared with quantitative methods that should be considered when deciding between methodologies. A common goal of qualitative research is to explore and understand the meaning attributed to some set of issue(s) through the untangling of complex processes. Qualitative methods can help researchers, clinicians, and policymakers understand the context in which topics such as medical decision making take place. For these reasons, qualitative methods are chosen when engaging in exploratory research where not enough knowledge is present to form specific hypotheses. Qualitative methods are well suited for these topics because people speak with their own voice rather than being forced to situate a response into a set of a priori assumptions. This helps facilitate discovery of new findings and can lead to the formation of theoretically grounded hypotheses to be tested later using quantitative methods.

Qualitative inquiry is often characterized as an iterative process as opposed to a linear order of events. This process involves ongoing analysis and reflection that adapts to discoveries made throughout the process. These methods traditionally have been used by social scientists, but the unique

contributions have been noticed by researchers from other disciplines, leading to a greater acceptance of qualitative methods as a science.

Specific Qualitative Methods

Qualitative Data Collection Techniques

Qualitative methods use various modes of data collection guided by the objectives of the research. Two of the most common types used in medical decision making research are semistructured interviews and focus groups.

Semistructured interviews involve interviewers asking respondents targeted questions and respondents replying in their own words. Interviewers are trained to ask probing questions based on initial responses to uncover more detail. This is a prime example of the iterative nature and adaptability of qualitative methods. Interviews are often tape-recorded and subsequently transcribed for analysis.

Focus groups are similar to interviews in terms of questions asked, the use of follow-up questions, and the practice of tape-recording. However, focus groups comprise more than one participant and can yield different findings compared with an interview due to the group dynamic. This dynamic can dampen or enhance response contingent on whether participants know each other, sensitivity of the topic, and other factors. These factors should be considered when deciding which data collection technique best suits the research topic.

Narratives, another form of qualitative data, can be collected through semistructured interviews, life history interviews, or found on the Internet. Narratives are stories that can range from broad to very specific if the content is guided by a researcher. For example, a researcher interested in dietary decisions of people with diabetes may ask a group of diabetics to write about their food choices and the surrounding circumstances for a week. Narratives may also preexist and can be compiled because they have something in common, such as being written by patients who have decided to forgo cancer treatment to try alternative approaches. These narratives can be found in newspaper and magazine editorials or in publicly available Internet blogs and discussion forums.

Two other forms of data collection techniques involve watching events take place. One less frequently used in medical decision making is participant observation or ethnography, where an investigator gains access to a group or community through an informant or gatekeeper. The investigator then immerses herself or himself in the group, observing, interacting, and engaging in activity. The second form, nonparticipant observation, involves a researcher watching and documenting the specifics of an event. The use of video-recording equipment no longer requires the researcher to be present during these events. Nonparticipant observation is more common in the study of medical decision making and specifically has been used in the study of patient–physician interactions. Observation is particularly useful when exploring subtle events people do not realize happen or behavior people do not realize they engage in.

Qualitative Data Analysis

Textual data can be analyzed using a variety of methods, one of the most common of which is grounded theory founded by Glaser and Strauss. This approach involves reading transcribed interviews or focus group reports multiple times to identify themes in the data. These themes are assembled in a list and discussed among a group of researchers who compare and refine the lists until an agreement is reached. The data are then revisited to connect specific pieces of text to the listed themes.

Other common forms of analysis include narrative analysis, where emphasis is placed on examining how a story is told, in particular, how people structure their stories, what are key components, what is the context, and so on. Content analysis can be used to code data from observation studies whereby qualitative data are quantified in some fashion, such as the number of times a physician makes eye contact with a patient. Other types of qualitative analysis include discourse analysis, used by ethnographers to uncover rules of conversation and deconstruction, which sets out to break down and untangle existing assumptions.

Presenting Qualitative Results

Results from qualitative methods can be presented in a number of different ways. Researchers

using grounded theory as an analytic tool usually provide illustrative quotations that best capture the general content of a theme. Some investigators will provide counts or the number of times a theme appears. Counts can be viewed as the quantification of qualitative findings, and researchers and readers must be cautious not to place too much emphasis on these numbers because qualitative samples are typically small and not generalizable. Pictures and diagrams are also used to present themes found in the data and are especially helpful when the code structure is complex and has many levels.

Other Considerations

Samples

Qualitative samples are often small and not generalizable to larger populations. In contrast to quantitative methods and the use of power calculations, qualitative methods do not facilitate prior determinations of sample size. To apply for grants, qualitative researchers can estimate the sample size needed based on rules of thumb and previous studies. The process of determining that enough data have been collected is guided by the concept of theoretical saturation, which is a point when the researcher concludes that no new themes are resulting from analyzing more data.

Interrater Reliability

The analysis of text can be open to interpretation, which is why a lot of qualitative research is conducted in teams. Also, due to concerns about consistency of analysis, researchers using the grounded theory approach have sought for ways to ensure that if others were coding the data they would arrive at similar conclusions. This has led to widespread use of interrater reliability calculations in qualitative methods that are increasingly required for publication. This technique helps assure readers that two or more individuals agreed a reasonable amount of times on the presence of themes in a sample of text drawn from the study.

Computer Software

The use of grounded theory in qualitative inquiry has given rise to the development of software programs that aid in the research process. These programs help organize qualitative data under user-specified code structures, which still require the researcher to analyze the data.

Mixed Methods

As different as qualitative and quantitative methods are in purpose, planning, data collection, analysis, and presentation of findings, the two approaches are very complementary and increasingly used in combination. A common example is the sequential exploratory design that begins with a research topic for which there is little previous work. Qualitative methods are then used to collect data and analyze the topic to generate themes. The themes are viewed as concepts to be operationalized with multiple quantitative items to be tested in larger, more representative samples. Once collected, quantitative analytic techniques can be applied to the data to examine how well the created items capture the content of the themes derived from the qualitative work. The blending of qualitative and quantitative methods does not have to be sequential but can be simultaneous where both types of data are collected at the same time and analyzed separately. Qualitative methods may also be used as a follow-up to a quantitative study to help answer questions raised by quantitative findings. Many types of mixed methods designs exist and new forms are being developed as more people appreciate the unique contributions of each method.

Downsides to Qualitative Methods

Although qualitative methods have many advantages over quantitative when engaging in certain types of inquiry, there are downsides. Qualitative methods generally are more time-consuming in terms of data collection and analysis because of the in-depth nature. Connected to this issue is increased expense, which can be greater because of longer time commitments from participants, interviewers, and researchers who analyze the data. Another limitation of qualitative methods is that they are typically not generalizable because they often use small convenience samples. Therefore, researchers who at the outset are looking to generalize their findings to a larger population may not want to solely use these methods. Similarly, qualitative methods are not going to establish definitive laws of behavior or determine causality but can help in moving this process forward.

Qualitative Methods in Clinical Practice

Qualitative methods can be used beyond research settings and applied in everyday situations such as clinical practice and used in the process of medical decision making. For example, these methods can be used by physicians to enhance the process of screening through the provision of a more in-depth picture of the patient's situation and the context surrounding their medical encounter. This can lead to the development of more individualized treatment plans that are better informed. Qualitative methods have also been useful in the development of decision aids, a popular tool in the field of medical decision making.

Noah J. Webster

See also Ethnographic Methods

Further Readings

Bergman, M. M. (2007). *Mixed methods research*. London: Sage.

Crabtree, B., & Miller, W. (1999). *Doing qualitative research* (2nd ed.). Thousand Oaks, CA: Sage.

Creswell, J. W., & Creswell, J. W. (2007). *Qualitative inquiry and research design: Choosing among five approaches*. Thousand Oaks, CA: Sage.

Denzin, N. K., & Lincoln, Y. S. (Eds.). (2005). *The SAGE handbook of qualitative research* (3rd ed.). Thousand Oaks, CA: Sage.

Glaser, B., & Strauss, A. L. (1967). *The discovery of grounded theory: Strategies for qualitative research*. Chicago: Aldine.

Marshall, C., & Rossman, G. (1999). *Designing qualitative research* (3rd ed.). Thousand Oaks, CA: Sage.

Rubin, H., & Rubin, I. (1995). *Qualitative interviewing: The art of hearing data*. Thousand Oaks, CA: Sage.

Strauss, A., & Corbin, J. (1990). *Basics of qualitative research*. Newbury Park, CA: Sage.

Tashakkori, A., & Teddlie, C. (Eds.). (2003). *Handbook of mixed methods in social and behavioral research*. Thousand Oaks, CA: Sage.

Quality-Adjusted Life Years (QALYs)

The quality-adjusted life year (QALY) is a measure of the value of health outcomes. A QALY takes into account both the quantity (survival) and the quality of life generated by healthcare interventions and allows decision makers to compare diverse health interventions using a common measure.

What Is a Quality-Adjusted Life Year?

One of the problems faced by decision makers is how to compare health interventions across diseases and with different health outcomes for priority setting in healthcare. A QALY-like concept was first proposed by Herbert Klarman in 1968 in a study of chronic kidney disease that estimated that quality of life was 25% better with transplant compared with dialysis. The method and premise was further developed in the 1970s, with the term *quality-adjusted life year* and QALY first being popularized by M. C. Weinstein and W. B. Stason in 1977. The underlying premise of this metric was to refer diverse health outcomes, such as lives saved, improved life expectancy, improvements in quality of life, functionality, or symptom control, back to the same value scale such that it would be possible to compare these diverse health outcomes with each other. When combined with information about the costs of alternative healthcare interventions, QALYs form the basis of cost-utility analysis; an incremental cost-utility ratio (or a cost per QALY gained) indicates the additional cost of one intervention compared with another that is required to generate 1 extra year of perfect health. QALYs are also referred to by different names; for example, the U.S. National Center for Health Statistics calls them *years of healthy life* (YHL), and Statistics Canada uses a variety of terms, including *health-adjusted life years* (HALYs) and *health-adjusted life expectancy* (HALE).

By capturing changes in both mortality (life expectancy) and morbidity (quality of life) related to a healthcare intervention, and combining them into a single outcome, the QALY offers advantages over health outcomes measured in natural units, for example, survival, because it is (a) likely to better capture the true scope of health-related effects of an intervention and (b) also provides a common metric by which diverse programs and interventions can theoretically be compared in terms of costs and consequences. For priority setting, the QALY metric therefore allows consideration of the

relative efficiency of wide-ranging interventions across different disease states.

Calculation

A QALY is calculated by weighting the time spent in different health states by how desirable that health state is. These weights are variously referred to as *QALY weights, QOL weights, utility weights,* and *HRQOL (health-related quality of life) weights.* These terms are often used interchangeably in the literature, although they are not strictly equivalent. The term *QALY weight* is used here. To operationalize the QALY metric, weights are needed to represent the health-related quality of life of different health states. QALY weights have a number of properties:

1. They are based on preferences; more preferred health states have a higher weight than less preferred; the weights should be based on a sample of individual preferences, obtained in a way that involves a trade-off between quality and quantity of life.
2. They are bounded on perfect health and death.
3. They measure strength of preference on a cardinal (interval) scale, with equal intervals measured in such a way that they have equal value.

The convention is to weight perfect health with a value of 1, while weighting death as 0. These conventions are useful for calculation and interpretation practicalities. If death received a weight of anything other than 0, then it would mean that in all analyses death would be accruing some weight into the future for as long as the death would state lasted (infinite). Giving perfect health a value of 1 means that the QALY is then interpreted in terms of years of perfect health, such that 1 year in perfect health is equivalent to 1 QALY; a year of less than perfect health therefore generates less than 1 QALY. In addition, some health states may be worse than death and therefore can have negative QALY weights. A QALY is thus calculated:

$$\text{QALY} = \text{Duration of health state (years)} \times \text{QALY weight for health state.}$$

A gain in QALYs from a healthcare intervention can therefore be derived from an improvement in survival, an improvement in quality of life, or an improvement in both survival and quality of life. This is shown in Figure 1. Without the intervention, a person survives for 3 years, with a QALY weight of .6 for Year 1, .4 for Year 2, and .3 for Year 3, with a total number of QALYs of .6 + .4 + .3 = 1.3 QALYs. In contrast, with the intervention a person will survive for 4 years, with a QALY weight of .8 for Year 1, .6 for Year 2, .75 for Year 3, and .4 for Year 4, giving a total number of QALYs of .8 + .6 + .75 + .4 = 2.55 QALYs. With the intervention the person gains 1.25 QALYs, resulting from improved quality of life over the time that the person would have been alive anyway (3 years) and an additional year of survival, albeit at less-than-perfect quality of life.

A number of different methods can be used to elicit QALY weights; the three main methods to directly measure QOL values and utilities are the visual analog scale or rating scale, the time trade-off, and the standard gamble. These methods can be used to directly measure an individual's utilities and values, or they can be used to value hypothetical, scenario-based descriptions of specific health states or a pathway of health states. Multi-attribute measures of health status can also be used, and these are used in a similar manner as generic QOL instruments to measure and value an individual's QOL.

Assumptions

For QALYs to accurately reflect preferences, a number of assumptions must be made about the nature of the QALY weight. QALYs assume that the value of the QOL weight is constant and is unrelated to (a) the duration of the health state, (b) when the health state occurs in time (e.g., now or at age 70), and (c) where the health state occurs in relation to other health states. These are important assumptions, and it has been argued that they do not necessarily hold. It is likely that the value of a health state is altered by the length of time a person spends in the state; it therefore might be necessary to estimate separate QALY weights for health states over different durations. It has also been suggested that prognosis may influence the value that is attached to a health state; a poor but

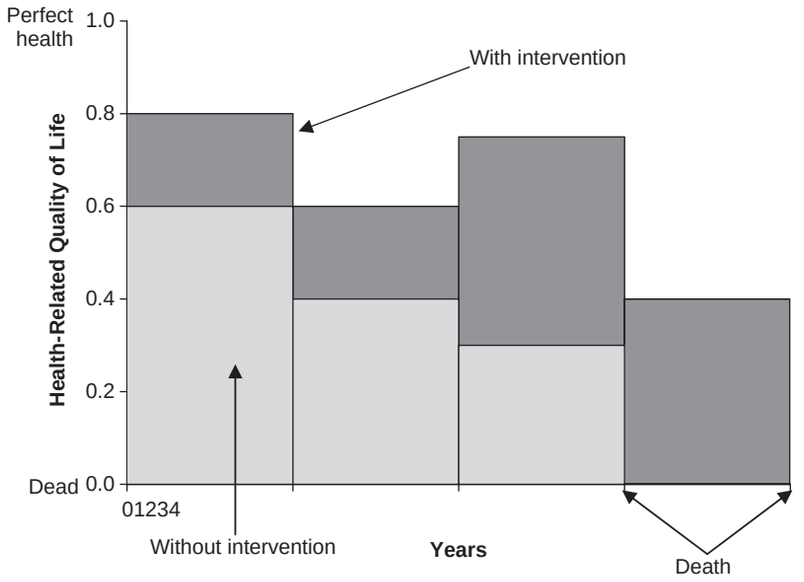


Figure 1 Quality-adjusted life years (QALYs)

temporary health state may be valued more highly than would be expected because it is perceived as a short-term inconvenience; whereas the value attached to a good health state might be diminished by the knowledge that it will eventually lead to poor quality of life and death. Most QALY applications assume that an individual's attitude toward risk is neutral.

Additionally, the QALY approach to priority setting assumes

- that the total value to society of a health intervention is the sum of all the health benefits (QALYs) that it produces in people who receive the intervention and
- that the health benefit in each person receiving the health intervention is the sum of all the gains in utility (QOL) over all the life years in which the person gets the benefits of the service.

That is, QALYs are additive, both within and across individuals. These assumptions are demonstrated in the following example: Individual A gets an increase in utility from .6 to .9 for 2 years, then from .6 to .7 for the following 3 years, giving a total of .9 QALYs gained ($2 \times .3 + 3 \times .1$). Individuals B and C get a health benefit of 2.5 and 1.6 QALYs gained, respectively. The total gain in health outcome for this population (A, B, and C) is therefore the sum of $.9 + 2.5 + 1.6 = 5$ QALYs gained. An additional implication is that this service for these three patients is valued as highly as a health service benefiting one person who gains 5 years in full health or a service that provides five patients with 1 additional year in full health each. That is, QALYs make no distinction regarding the distribution of health benefit in society. This is referred to as the assumption of *distributive neutrality*.

Shortcomings

The QALY concept is not without criticism. Criticism ranges from those who believe that the QALY approach is too complex, and decision making would be better served by the use of disaggregated health outcomes counted in natural units, to those who believe that the QALY approach is too simplistic and more complex methods should instead be used. Alternative measures have been suggested, including healthy years equivalents (HYEs), saved young life equivalents (SAVEs), and disability-adjusted life years (DALYs), although these measures are not without their own criticisms.

Criticisms also relate to more technical aspects of the calculation of QALYs, including the underlying

assumptions discussed above, such as duration and prognosis, and the measurement of QALY weights. Alternative methods of measuring QALY weights, such as the time trade-off, standard gamble, and multi-attribute utility instruments, will almost always lead to different values for a particular health state, and this variability between measures has led opponents of QALYs to criticize the QALY metric and the methods used to generate the weights. Concerns also relate to the notion that QALYs may inadequately capture true quality-of-life implications in a number of circumstances, for example, (a) the quality of life associated with emotional or mental health states; (b) the health effects of preventive programs, where the benefits in terms of health outcomes may not occur for many years in the future and may be difficult to value because the value attached to different aspects of health is likely to vary with age and life context; and (c) temporary or short-term changes in quality of life, which may also be very difficult to value using a QALY metric because the measures may not be sensitive to small changes. When using descriptions of health states, concerns have been raised over whether the responses given to hypothetical situations reflect people's real decisions.

Additional criticisms leveled at QALYs are related to broader issues, such as whether "perfect health" is interpreted consistently among respondents, whether QALYs undervalue healthcare because they do not capture the wider benefits of healthcare (externalities), and whose preferences should be used to inform decision making, as different respondent populations are likely to give different values to the same health state, for example, the community, compared with patients, compared with health practitioners.

Use in Decision Making and Resource Allocation

Despite the shortcomings of QALYs, a number of decision-making bodies, such as the National Institute for Health and Clinical Excellence (NICE) in the United Kingdom, the Pharmaceutical Benefits Advisory Committee (PBAC) in Australia, and the Canadian Agency for Drugs and Technologies in Health (CADTH), use QALYs to inform decision making about funding of pharmaceutical and nonpharmaceutical health technologies. Although research continues on the measurement and application of QALYs, they are, to date, probably the most widely accepted, and used, measure of health outcome that combines both morbidity and mortality into a single usable metric.

Kirsten Howard

See also Cost-Utility Analysis; Disability-Adjusted Life Years (DALYs); League Tables for Incremental Cost-Effectiveness Ratios; Quality-Adjusted Time Without Symptoms or Toxicity (Q-TWiST); Risk Aversion; Utility Assessment Techniques

Further Readings

Brazier, J., Deverill, M., Green, C., Harper, R., & Booth, A. (1999). A review of the use of health status measures in economic evaluation. *Health Technology Assessment, 3*(9). Retrieved February 28, 2009, from http://www.ncchta.org/fullmono/mon309.pdf

Drummond, M. F., Sculpher, M. J., Torrance, G. W., O'Brien, B. J., & Stoddart, G. L. (2005). *Methods for the economic evaluation of health care programmes* (3rd ed.). Oxford, UK: Oxford University Press.

Fox-Rushby, J. (2002). *Disability-adjusted life years (DALYs) for decision making? An overview of the literature.* London: Office of Health Economics.

Klarman, H., Francis, J., & Rosenthal, G. (1968). Cost-effectiveness analysis applied to the treatment of chronic renal disease. *Medical Care, 6,* 48–54.

Mehrez, A., & Gafni, A. (1992). Preference-based outcome measures for economic evaluation of drug interventions: Quality-adjusted life years (QALYs) versus healthy-years equivalents (HYEs). *Pharmacoeconomics, 1,* 338–345.

Nord, E. (1999). *Cost-value analysis in health care. Making sense out of QALYs.* Cambridge, UK: Cambridge University Press.

Pliskin, J., Shepard, D., & Weinstein, M. C. (1980). Utility functions for life years and health status. *Operations Research, 28,* 207–224.

Sackett, D. L., & Torrance, G. (1978). The utility of different health states as perceived by the general public. *Journal of Chronic Diseases, 31,* 697–704.

Torrance, G. W. (1986). Measurement of health state utilities for economic appraisal. *Journal of Health Economics, 5,* 1–30.

Weinstein, M. C., & Stason, W. B. (1977). Foundations of cost-effectiveness analysis for health and medical practices. *New England Journal of Medicine, 296,* 716–721.

Quality-Adjusted Time Without Symptoms or Toxicity (Q-TWiST)

Quality-adjusted time without symptoms or toxicity (Q-TWiST) is an outcome measure for cancer clinical trials, which was developed in the mid-1980s by Richard Gelber and Aron Goldhirsch to evaluate adjuvant chemotherapy. It is an adaptation of the QALY, the quality-adjusted life year. It is calculated as follows:

$$\text{Q-TWiST} = u_t \times \text{TOX} + \text{TWiST} + u_r \times \text{REL},$$

where TOX is the months spent with any burden of subjective treatment side effects, REL is the months following disease relapse, TWiST is defined as time without symptoms of disease and toxicity of treatment, and u is a utility coefficient, taking values between 0 and 1, to represent the value, relative to TWiST, of TOX and REL, respectively, and the subscript "t" is toxicity and "r" is relapse. The developers proposed to use arbitrary values for u and to show the effect of different values by performing threshold analyses, also called sensitivity analyses. Such analyses result in combinations of utilities for TOX and REL (relative to TWiST) whereby one treatment strategy is superior to the other. Clinicians should then assess these utilities with their patients to decide which treatment is superior. An example is Q-TWiST for radiation therapy in the treatment of patients with poor-prognosis resectable rectal cancer:

$$\begin{aligned}\text{Q-TWiST} &= u_t \times \text{TOX} + \text{TWiST} + u_r \times \text{REL} \\ &= .5 \times .9 + 33.2 + .5 \times 8.7 = 38 \text{ months}.\end{aligned}$$

Outcome Measurement in Oncology

TWiST

Overall survival time is the most definitive endpoint used to evaluate treatment effectiveness for cancer patients. Other measures such as tumor-free interval, duration of response, or time to progression of disease are also considered for making therapeutic decisions. A value judgment is often made by weighing benefits in terms of these measures against the risks of undesirable side effects of treatment. In the second half of the 1980s, Gelber and Goldhirsch developed a quality-of-life-oriented endpoint for assessing adjuvant therapies in oncology. This endpoint was obtained by subtracting the amount of time of poor quality of life from each unit time interval to adjust the measure of benefits. It reflected the amount of good quality time enjoyed by the patients. Specifically, the time without symptomatic relapse of cancer was adjusted by subtracting units of time during which toxic effects of treatment were experienced. The measure defined time without symptoms of disease and toxicity of treatment (TWiST). It was calculated for each patient by subtracting from overall survival periods of time during which treatment or disease reduced quality of life. These periods included months with any burden of subjective treatment side effects (TOX) and all months following disease relapse (REL). Average TWiST could be calculated for several treatments and compared over time to see when (if ever) after start of treatment the risk-benefit ratio began for a treatment with more early toxicity.

Quality-Adjusted Survival: Q-TWiST

The all-or-none analysis of TWiST was deemed somewhat unrealistic by the developers, as it assigned no value to both the period of life with toxicity and the period following relapse. A refinement was created to TWiST to include in the analysis times spent with toxicity or relapse but with intermediate weightings based on their value relative to TWiST.

Q-TWiST and QALY are equivalent concepts, and depending on the elicitation of the utility coefficient u and on the way the data are analyzed they will return similar or different results. A utility is defined as the level of desirability that people associate with a particular outcome. It is a cardinal number that represents the strength of an individual's preference for a particular outcome when faced with uncertainty. Utilities are assigned to each outcome, on a scale that is established by assigning a value of 1 to the state of optimal health and a value of 0 to death. In QALYs, each year of survival is multiplied by its utility, and the thus adjusted life years are summed. In Q-TWiST, life years are assigned to specific health state categories (TOX, REL, TWiST) and multiplied with a fixed

utility for that category. The categorization may make the analysis more appealing to clinicians.

Anne M. Stiggelbout

See also Holistic Measurement; Quality-Adjusted Life Years (QALYs); Survival Analysis; Utility Assessment Techniques

Further Readings

Gelber, R. D., Gelman, R. S., & Goldhirsch, A. (1989). A quality-of-life-oriented endpoint for comparing therapies. *Biometrics, 45,* 781–795.

Goldhirsch, A., Gelber, R. D., Simes, J., Glasziou, P., & Coates, A. S., for the Ludwig Breast Cancer Study Group. (1989). Costs and benefits of adjuvant therapy in breast cancer: A quality-adjusted survival analysis. *Journal of Clinical Oncology, 7,* 36–44.

Nooij, M. A., de Haes, J. C., Beex, L. V., Wildiers, J., Klijn, J., Becquart, D., et al. (2003). Continuing chemotherapy or not after the induction treatment in advanced breast cancer patients: Clinical outcomes and oncologists' preferences. *European Journal of Cancer, 39,* 614–621.

Quality of Well-Being Scale

Medical decision making at the organizational and societal levels usually requires that health benefits be quantified in a common unit. This enables administrators or decision makers to compare and evaluate programs that address different diseases or populations. To compare very different alternatives, measures of program benefits must be comprehensive, including all possible direct effects, whether intended or not. Therefore, the current consensus is that preference- or utility-based measures of generic health-related quality of life (HRQOL) best meet these criteria. The most frequently used measures in this class include the EuroQOL 5D (EQ-5D), the Health Utilities Index measures (HUI), and the Quality of Well-Being Scale measures (QWB and QWB-SA).

The Quality of Well-Being (QWB) scale is a generic, preference-based measure of HRQOL with well-established psychometric properties in a wide variety of diseases and subgroups. In response to limitations of the QWB, a self-administered version of the QWB (QWB-SA) has been developed and validated. The QWB-SA is quicker and easier to administer in most research and clinical assessment protocols. Both questionnaires assess the presence or absence of symptoms and functioning on specific days prior to administration. The measures produce a single score that ranges from 0 (death) to 1.0 (optimal HRQOL). The final score from the QWB-SA or the QWB can be integrated with time and mortality to calculate quality-adjusted life years (QALYs) and conduct cost-effectiveness analysis.

Health-Related Quality of Life

Health-related quality of life (HRQOL) provides a comprehensive description of health and overall well-being. HRQOL measures can be classified in a number of different ways. For example, HRQOL measures are either generic or disease-specific and can be described as psychometrically based or preference/utility based. The QWB (and QWB-SA) is a generic HRQOL measure that was designed to be used with any adult population and any health condition, including healthy individuals. The QWB and QWB-SA are preference-based measures and were not developed to assess statistically independent domains of HRQOL. They provide a single score that summarizes total HRQOL based on the mean preference ratings that health consumers gave to the health states described within it. These preferences or utilities are ratings of observable health states using a scale anchored by death and optimum health, and assuming equal intervals.

Theoretical Basis

The QWB was developed in the 1970s based on a General Health Policy model. This theoretical model focuses on mortality (death) and morbidity (health-related quality of life) and proposes that symptoms and disabilities are important for two reasons: First, illness may cause life expectancy to be shortened and, second, illness may make life less desirable at times prior to death. In assessing the impact of medical interventions or programs, the model requires data on changes in mortality as well as on changes in HRQOL. The General Health Policy model incorporates preference for observed health states (utility) and duration of stay

in health states. Preferences or utility for health states are typically measured using economic principles that ask individuals to place preferences or values on a wide variety of health states involving both symptoms and functioning. The health preferences or utilities are placed on a preference continuum for the desirability of various health states, giving a "quality" rating on an interval scale ranging from 0 for death to 1.0 for completely well.

Quality-Adjusted Life Years and Cost-Effectiveness

Once a mean QWB score is obtained that describes the level of morbidity or wellness in a sample, the score can be multiplied by the amount of time at that level of wellness to calculate QALYs. A QALY is defined as the equivalent of 1 completely well year of life or a year of life with optimal functioning and no health problems or symptoms. For example, imagine a person who has a set of symptoms and is in a state of functioning that is rated by community peers as 0.5 on a 0.0 to 1.0 scale. If the person remains in that state for 1 year, living that full year with its quality reduced to half of what is optimal, then it is considered equivalent to living 6 months with optimal quality of life (.5 year × 1.0 QOL score). Thus, a person requiring a cane or walker to get around might be hypothetically rated at .50. If he or she remained in that state for an entire year, the individual would lose the equivalent of one half of a QALY. However, a person who has the flu may also be rated as .50. In this case, the illness might only last 3 days and the total loss in QALYs might be 3/365 × .50, which is equal to .004 QALYs. The .004 QALYs may seem insignificant when compared with the person who has difficulty walking but suppose that 5,000 people in a community get the flu. The well years lost would then be 5,000 × .004 QALYs, which is equal to 20 years of perfect health for one person. An important feature of the QALY system is that it is completely generic. It can be used to compare small health consequences that affect a large number of people or large health consequences that affect a small number of people or any variation of those factors. The quality-adjusted life expectancy is the current life expectancy adjusted for diminished quality of life associated with dysfunctional states and the duration of stay in each state.

The calculation of QALYs is required for conducting cost-utility analysis, which is simply a type of cost-effectiveness analysis that uses QALYs as its unit measure of health benefit. The QWB was the first assessment instrument developed for the primary purpose of calculating QALYs in cost-effectiveness analysis. Prior to the existence of generic, preference-based measures, many different outcomes were used to represent the effectiveness side of cost-effectiveness analyses. Generic, preference-based measures and QALYs have become the recommended standard for cost-effectiveness analyses because they provide a common metric for comparing results across studies and populations.

Assessment

In the original QWB, respondents report whether or not each of 27 groups of symptoms were experienced on each of the 6 days prior to the assessment. Functioning is assessed by questions about the presence of functional limitations over the previous 6 days, within three separate domains (mobility, physical activity, and social activity). Unlike measures that ask about general time frames such as "the past 4 weeks" or "the previous month," the QWB asks whether specific symptoms or functional limitations did or did not occur on specific days. Each group of symptoms and functional limitation is weighted using preferences obtained from the ratings of 856 people randomly sampled from the general population. The domain scores (3 functioning, 1 symptom) are subtracted from 1.0 to create a total score that provides an expression of well-being that ranges from 0 (death) to 1.0 (asymptomatic optimal functioning). References on the validation of the instrument are available from the University of California, San Diego (UCSD) Health Services Research Center. The original QWB must be administered by a trained interviewer because it employs a complex system of branching questions and probes. The original questionnaire takes an average of about 15 minutes to complete. The authors believe that the administration time and complexity of the original measure requiring a trained interviewer has resulted in its underutilization.

Self-Administered Version

In 1996, a self-administered version of the questionnaire was developed to address some of the limitations of the original version. The Quality of

Well-Being Scale-Self-Administered (QWB-SA) improves on the original version in a number of ways. First, the administration of the questionnaire no longer requires a trained interviewer and can be completed in about 10 minutes. Second, the assessment of symptoms follows a clinically useful "Review of Systems" model, rather than clustering symptoms based on preference weights. Third, a wider variety of symptoms are included in the QWB-SA, making it more comprehensive and improving the assessment of mental health.

Preference weights for the QWB-SA were obtained from a new sample and studies have been published comparing the new and old versions. The QWB-SA and QWB were highly correlated and the test-retest reliability is high. The measure is not designed to be internally consistent because the factors it measures (symptoms and functioning) are interdependent. QWB-SA scores tend to be slightly lower than QWB scores, primarily because mental health symptoms are assessed in greater detail and are more likely to contribute to decreased scores.

The format for the QWB-SA includes five sections. Part 1 assesses the presence or absence of 19 chronic symptoms or problems (e.g., blindness, speech problems). Because these symptoms are chronic, they are not expected to vary over the 3-day assessment period and therefore are assessed using a yes/no response format. The chronic symptoms are followed by 25 acute (or more transient) physical symptoms (e.g., headache, coughing, pain) and 14 mental health symptoms (e.g., sadness, anxiety, irritation). The remaining sections of the QWB-SA are similar to the QWB and include assessment of mobility (including use of transportation), physical activity (e.g., walking and bending over), and social activity, including completion of role expectations (e.g., work, school, or home).

The recall period assessed by the QWB-SA is shorter than in the QWB. The QWB asked patients about symptoms and function during each of the 6 days prior to the day of administration, while the QWB-SA questions refer to each of the 3 days prior to the day of administration. This change was designed to reduce respondents' burden and recall bias without sacrificing valuable information. The impact of this change was examined by dropping information from days 4, 5, 6 and recalculating QWB scores based only on the past 3 days. No significant differences in scores were found between the overall quality of life score when using only the most recent 3 days, and the change resulted in a shorter administration time.

A total of 12 different symptom questions in the QWB-SA are related to mental health, including questions about symptoms indicative of mood, anxiety, psychotic features, appetite, energy, anhedonia, and sleep. Researchers are exploring the possibility of deriving a mental health subscale based on these questions; however, it is hard to separate out the impact that these symptoms have on functioning as opposed to that attributable to nonmental health or "physical" symptoms.

When compared with other generic, preference-based measures of HRQOL, the QWB-SA is slightly longer and is more time-consuming because it involves a more comprehensive assessment of symptoms and functioning. However, this more detailed assessment of symptoms and functioning may result in greater sensitivity to change in some populations. The QWB-SA asks about the presence or absence of specific complaints on specific days to reduce the influence of memory, or severity ratings such as pain intensity, that require personal interpretation. In addition, the distribution of QWB-SA scores in most studies is close to normal, suggesting that ceiling or floor effects are less common than with other HRQOL measures.

Uses

The QWB-SA has been used in the evaluation of many chronic disease populations, and the measure has been selected for several multisite National Institutes of Health randomized controlled trials, including Lifestyle Interventions and Independence for Elders (LIFE-P), the National Emphysema Treatment Trial (NETT), the Diabetes Prevention Program (DPP), and portions of the Prostate, Lung, Colorectal, and Ovarian Cancer (PLCO) and the Modification of Diet in Renal Disease (MRDR) trials. In addition, the QWB has been used in a variety of clinical studies for a range of medical and surgical conditions that include chronic obstructive pulmonary disease, AIDS, cystic fibrosis, diabetes mellitus, atrial fibrillation, lung transplantation, arthritis, cancer, schizophrenia, and many other conditions.

Both the QWB and QWB-SA are available free of charge to users from nonprofit organizations. A

small fee is charged to for-profit users. Information on copyright agreements and user manuals are available at http://outcomes.ucsd.edu/portalVBVS/DesktopDefault.aspx.

Erik J. Groessl and Robert M. Kaplan

See also Cost-Effectiveness Analysis; Cost-Utility Analysis; EuroQoL (EQ-5D); Health Utilities Index Mark 2 and 3 (HUI2, HUI3); Quality-Adjusted Life Years (QALYs)

Further Readings

Gold, M. R., Siegel, J. E., Russell, L. B., & Weinstein, M. C. (1996). *Cost-effectiveness in health and medicine.* New York: Oxford University Press. Retrieved January 13, 2009, from http://outcomes.ucsd.edu/portalVBVS/DesktopDefault.aspx

Groessl, E. J., Kaplan, R. M., Rejeski, W. J., Katula, J. A., King, A. C., Frierson G., et al. (2007). Health-related quality of life in older adults at risk for disability. *American Journal of Preventive Medicine, 33*(3), 214–218.

Kaplan, R., & Anderson, J. (1990). An integrated approach to quality of life assessment: The general health policy model. In B. Spilker (Ed.), *Quality of life in clinical studies* (pp. 131–149). New York: Raven Press.

Sieber, W. J., Groessl, E. J., David, K. M., Ganiats, T. G., & Kaplan, R. M. (2007). *User's manual: Quality of well-being scale, self-administered.* Retrieved January 13, 2009, from http://outcomes.ucsd.edu/portalVBVS/DesktopDefault.aspx

R

Randomized Clinical Trials

A clinical trial is defined as a controlled, prospective assessment of an intervention for a disease or condition in human beings. In general, the purposes of randomized clinical trials, or Phase III trials, are to evaluate the efficacy of a novel treatment relative to an observation/placebo arm, or a standard therapy, or to determine whether an experimental treatment is more effective than a standard therapy having lesser morbidity. According to ClinicalTrials.gov, Phase III trials

> are expanded controlled and uncontrolled trials after preliminary evidence suggesting effectiveness of the drug has been obtained, and are intended to gather additional information to evaluate the overall benefit-risk relationship of the drug and provide and *[sic]* adequate basis for physician labeling.

In essence, the ultimate goal of a Phase III clinical trial is to improve medical practice.

The purpose of this entry is to present an overview of the basic principles involved in the design and conduct of Phase III trials. Meticulously designed Phase III trials can answer important scientific questions so that valid inferences about the therapy being tested can be made. In the following sections, the discussion focuses on determining the sample size within the context of testing a hypothesis. Investigators who are interested in conducting a Phase III trial should start by specifying a hypothesis of interest. Once this is stated, then the required sample size should be calculated to reflect a realistic clinical-effect size, and an accurate estimate of the projected accrual rate should be used. In addition, these trials should include sequential guidelines for stopping a trial early so that participants can benefit from a promising therapy or are spared from a harmful treatment.

Hypothesis Testing

Every Phase III trial starts by asking an important question about the effectiveness of a new experimental treatment on outcome. For the parameter of interest, two states must be defined before testing a hypothesis: (1) the null hypothesis, which is usually a hypothesis of no difference in the parameter of interest between the groups, and (2) the alternative hypothesis, which can be either a two-tailed (all possible values of the parameter of interest are considered except the null) or a one-sided alternative (the parameter of interest will vary from the null in only one direction). Researchers must first decide the two-sided versus one-sided question as part of the hypothesis to be tested. Most Phase III trials are based on a two-sided question, and consequently, the number of patients required in such a trial is larger than when a one-sided question is tested.

The null and alternative hypotheses discussed in the section above are based on superiority studies. In some instances, however, the interest may lie in testing whether two regimens are as effective or equivalent. Equivalence and noninferiority trials

differ from superiority studies in that the objective is to test whether an experimental therapy is as effective (or not worse) than a current standard.

Types of Error

Investigators always make two types of errors in the process of statistical inference: (1) Type I and (2) Type II error rates. Type I, or the false-positive rate, is the probability of rejecting the null hypothesis when the null hypothesis is true—that is, concluding an ineffective treatment as active. The probability of committing a Type I error rate is traditionally chosen as 5%. A Type II, or false-negative rate, is the probability of not rejecting the null hypothesis when the null hypothesis is not true—that is, declaring an effective treatment as inactive. Power is defined as the complement of Type II error rate and is the probability of rejecting the null hypothesis when the alternative hypothesis is true.

Design Considerations

Endpoints

An endpoint is a criterion by which patient benefit is measured. Reliable and meaningful endpoints are inherent to well-designed Phase III studies, and it is crucial to explicitly define the primary endpoint at the design stage. Due to space limitations, two types of outcomes are discussed in this section: (1) time-to-event and (2) binary endpoints. Time-to-event endpoints are outcomes where time is measured from the date of randomization until the date of occurrence of an event of interest. The time variable is the failure time and is measured in years (months, weeks, or days), and the event is an incident of interest and may be death, death due to a certain cause, disease progression, or the development of metastases. Time-to-event endpoints must take into account a fundamental analytical element known as censoring. Censoring arises either because patients do not experience an event of interest before the trial ends or are lost during the follow-up period. As a result, information about an individual failure time will be unknown. Using time-to-event endpoints has the advantage of using all available information, including participants who fail to complete the trial.

In general, overall survival is the most common time-to-event endpoint used in Phase III trials in cancer. It is considered to be the "hardest" endpoint as it is the most objective endpoint. Overall survival is defined as the interval between date of randomization and date of death due to any cause. Other endpoints such as progression-free survival and disease-free survival are frequently employed in oncology trials.

Binary endpoints are common outcomes that are dichotomous in nature and are often based on success rates (yes or no). For example, in one trial, the primary endpoint was the presence of prostate cancer based on a biopsy performed at the end of the study.

Randomization

Randomization is the fundamental basis of all Phase III clinical trials and inherent to their validity. In addition, randomization minimizes bias and is a keystone in establishing the validity of the statistical tests of significance. Randomization helps in making the treatment groups balanced and comparable at baseline. Randomized block design is one of the most common and simplest methods of randomization. A block of size b is a series of treatment assignments that are generated randomly where patients are allocated sequentially to treatments as they enter the trial. The advantage of this method is that the imbalance between treatments is never greater than $b/2$. However, the main drawback is that if the block size is known in advance, clinicians may be able to guess the next treatment assignment and bias the results by putting "worse" patients on the control arm.

Patient response may depend on prognostic factors, and randomization helps balance such factors by treatments. Some imbalances, however, may occur by chance. One strategy is to use blocked randomization within predefined combinations of the prognostic factors (strata). Using stratified block randomization, patients are randomized to treatments using block sizes equal to b within each stratum. In recently designed trials, randomization was stratified by the predicted survival probabilities based on prognostic models. For example, in CALGB 90401, a trial that enrolled 1,050 men with castrate-resistant prostate cancer, randomization was stratified by the

predicted survival probability at 24 months: <10%, 10%–29.9%, or ≥30%.

Sample Size Determination

Perhaps the most crucial step in the design of a Phase III trial is the determination of how many participants are required to test the alternative hypothesis. It is unethical to recruit too few patients on a trial as there may be inadequate evidence to make a decision on whether the new treatment is effective. On the other hand, recruiting too many participants is both costly and time-consuming.

Comparison of Two Survival Curves

Suppose that an investigator is interested in comparing the survival curves of subjects who are treated with a standard of care or control (denoted as 1) with those who are treated with an experimental therapy (denoted as 2). A survival curve can be viewed as a graph of the probability of surviving up to a given point. Even though methods used for analyzing survival data are nonparametric—that is, assumption about the survival distribution (denoted as $S(t)$) does not need to follow any particular model (such as the Kaplan-Meier approach), assumptions are often made for sample size determination. One of the most common survival distributions used is the exponential model, where $S(t) = e^{-\lambda t}$. The hazard rate, λ, is constant over time and is defined as the instantaneous potential per unit time for the event to occur, given that the individual survived up to time t. The null hypothesis is $\lambda_1 = \lambda_2$ versus the alternative hypothesis $\lambda_1 \neq \lambda_2$. At the design stage, λ_1 may be unknown, but it is assumed that the median survival time (M_1) in the control group is known. Therefore, using the well-known relationship $\lambda_1 = (\log 2)/M_1$ for the exponential distribution, λ_1 can be determined. The hazard ratio (denoted as $\Delta = \lambda_1/\lambda_2$) is the ratio of the hazard rate in subjects assigned to the control and experimental arms. Alternatively, the hazard ratio may be defined as the ratio of the median survival time in the experimental arm to the control arm ($\Delta = M_2/M_1$). The number of deaths (d) required to be observed to test the alternative hypothesis with power $(1 - \beta)$ is

$$d = [z_{1-\alpha/2} + z_{1-\beta}]^2/W_i(\log \Delta)^2, \quad (1)$$

where

$z_{1-\alpha/2} + z_{1-\beta}$ are the quantiles obtained from the standard normal distribution for the Type I and Type II error rates and

W_i is the proportion of participants allocated to the experimental and control arms ($i = 1, 2$).

If an investigator is interested in testing a one-sided alternative hypothesis, then $z_{1-\alpha/2}$ is replaced by $z_{1-\alpha}$ in Equation 1. The basic assumption in Equation 1 is that all patients have been followed until death. The sample size needed in a clinical trial with a time-to-an-event endpoint is not only a function of Type I and Type II error rates as well as Δ but also takes into account the probability of death over the duration of the trial. It is assumed that there is an accrual period (T), where patients enter the clinical trial according to a Poisson process and are subsequently followed up for a predetermined period of time (τ).

Table 1 provides the accrual period (T) assuming various accrual rates, hazard ratios, and follow-up periods in years. Since the accrual rate is considered fixed, T is solved by using a Newton-Raphson procedure so that the power of $1 - \beta$ is obtained at the end of the trial ($T + \tau$). The required sample size is obtained by multiplying the accrual rate by the accrual period (T). To illustrate this point, let us suppose that an investigator wishes to test the alternative hypothesis that $\Delta = 1.35$, assuming a two-sided Type I error of .05 and a power of 90%. The investigator estimates based on historical data that the accrual rate per year is 120 patients. Assuming that the follow-up period is 1 year, from Table 1 the accrual period is 4.81 years, the number of participants required to test this hypothesis is 578 (4.81 years × 120 patients/year), and the total trial duration will be 5.81 years (4.81 years accrual period + 1 year of follow-up).

It is worth noting that Equation 1 assumes that the proportional hazards assumption is not violated in that the hazard ratio is constant over time. Although the log-rank statistic can still be used for analyzing survival data when the proportional hazards assumption is not satisfied, it is not optimal. Another method is proposed that does not require that this assumption to be met, and one report has shown that this approach is accurate in nonproportional hazards settings.

Table 1 Accrual period in years assuming different accrual rates and hazard ratios

	Accrual Rate (per Year)					
τ Δ	*60*	*80*	*120*	*160*	*180*	*240*
11.20	16.62 (13.27)	12.68 (10.17)	8.73 (7.05)	6.75 (5.48)	6.09 (4.96)	4.75 (3.88)
	21.95 (18.05)	16.68 (13.75)	11.40 (9.45)	8.76 (7.29)	7.88 (6.57)	6.11 (5.12)
1.25	11.42 (9.18)	8.79 (7.10)	6.14 (5.00)	4.79 (3.92)	4.34 (3.55)	3.41 (2.80)
	14.98 (12.38)	11.46 (9.50)	7.93 (6.62)	6.15 (5.16)	5.56 (4.67)	4.35 (3.67)
1.30	8.55 (6.92)	6.63 (5.39)	4.68 (3.83)	3.68 (3.02)	3.34 (2.75)	2.64 (2.17)
	11.13 (9.24)	8.57 (7.15)	6.00 (5.04)	4.70 (3.96)	4.26 (3.59)	3.35 (2.83)
1.35	6.78 (5.52)	5.29 (4.33)	3.77 (3.10)	2.98 (2.45)	2.71 (2.23)	2.14 (1.76)
	8.76 (7.31)	6.80 (5.70)	4.81 (4.05)	3.78 (3.20)	3.43 (2.90)	2.71 (2.30)
21.20	16.22 (12.88)	12.28 (9.77)	8.34 (6.66)	6.36 (5.10)	5.70 (4.57)	4.37 (3.51)
	21.55 (17.65)	16.28 (13.35)	11.01 (9.05)	8.37 (6.90)	7.49 (6.18)	5.72 (4.73)
1.25	11.02 (8.78)	8.38 (6.70)	5.74 (4.60)	4.40 (3.54)	3.95 (3.18)	3.04 (2.45)
	14.58 (11.97)	11.05 (9.10)	7.53 (6.22)	5.76 (4.77)	5.16 (4.28)	3.97 (3.29)
1.30	8.13 (6.51)	6.22 (4.99)	4.29 (3.45)	3.30 (2.66)	2.97 (2.39)	2.29 (1.84)
	10.71 (8.83)	8.16 (6.74)	5.60 (4.64)	4.30 (3.57)	3.86 (3.21)	2.98 (2.47)
1.35	6.36 (5.11)	4.89 (3.93)	3.38 (2.73)	2.61 (2.10)	2.35 (1.89)	1.81 (1.46)
	8.34 (6.90)	6.38 (5.29)	4.40 (3.66)	3.39 (2.82)	3.05 (2.54)	2.35 (1.96)

Notes: τ = Follow-up period (years), $\Delta = \lambda_1/\lambda_2$, median = 1 year in Group 1 (hazard rate $\lambda_1 = \log 2$). Upper numbers are based on Type I error rate = .05, Type II error rate = .20 (power = 80%). Lower numbers are based on Type I error rate = .05, Type II error rate = .10 (power = 90%). Numbers within parentheses are based on one-sided tests, whereas numbers outside the parentheses are based on two-sided tests.

Although most Phase III trials are designed with at least 80% power, a trial can fail to reject the null hypothesis if the effect size or difference between the two arms is too large. Some cancer trials exhibited an overly optimistic estimate of effect size. For example, in a trial of men with castrate-resistant prostate cancer, the log-rank statistic has 80% power, assuming a hazard ratio of 1.5. A 50% increase in survival is considered a large effect size, and trials in oncology would benefit from a more realistic estimate of the clinically meaningful effect size.

Binary Endpoints

The null hypothesis to be tested is that the proportion of success (such as complete response proportion) is equal in the two groups ($P_1 = P_2$) against the alternative hypothesis that proportions are not equal ($P_1 \neq P_2$). Table 2 presents the number of patients required per arm with a Type I error rate of .05, assuming one-sided and two-sided alternatives with 80% or 90% power using an approximation. An investigator is interested in determining the number of breast cancer patients needed in a Phase III

Table 2 Number of patients required per arm for testing two proportions

	$P_2 - P_1$						
P_1	*.05*	*.08*	*.10*	*.13*	*.15*	*.18*	*.20*
.10	725 (579)	319 (256)	219 (176)	142 (115)	112 (91)	84 (68)	71 (58)
	957 (787)	418 (345)	286 (236)	184 (153)	146 (121)	109 (90)	92 (77)
.15	945 (753)	401 (321)	270 (216)	171 (137)	133 (108)	98 (79)	82 (67)
	1251 (1,027)	528 (435)	354 (292)	223 (184)	174 (144)	127 (105)	106 (88)
.20	1,133 (901)	471 (376)	313 (250)	195 (156)	151 (121)	109 (88)	91 (74)
	1,503 (1,232)	622 (511)	412 (339)	255 (211)	197 (163)	142 (118)	118 (98)
.25	1,290 (1,025)	529 (421)	348 (278)	214 (172)	165 (133)	118 (95)	98 (79)
	1,714 (1,404)	699 (574)	459 (378)	281 (232)	216 (178)	154 (128)	127 (105)
.30	1,416 (1,124)	574 (457)	376 (300)	229 (184)	175 (141)	125 (101)	103 (83)
	1,882 (1,541)	760 (623)	496 (408)	301 (248)	230 (190)	163 (135)	134 (111)
.35	1,510 (1,198)	607 (483)	395 (315)	239 (192)	182 (146)	129 (104)	106 (85)
	2,008 (1,644)	804 (660)	522 (429)	315 (259)	239 (197)	169 (139)	138 (114)
.40	1,573 (1,247)	628 (500)	407 (325)	245 (196)	186 (149)	131 (105)	107 (86)
	2,092 (1,712)	832 (682)	538 (442)	322 (265)	244 (201)	171 (141)	139 (115)
.45	1,604 (1,272)	637 (506)	411 (328)	246 (197)	186 (149)	130 (105)	106 (85)
	2,134 (1,746)	843 (692)	543 (446)	324 (266)	244 (201)	170 (140)	138 (114)
.50	1,604 (1,272)	633 (504)	407 (325)	242 (194)	182 (146)	127 (102)	103 (83)
	2,134 (1,746)	838 (688)	538 (442)	319 (262)	239 (197)	166 (137)	134 (111)

Notes: Upper numbers are based on Type I error rate = .05, Type II error rate = .20 (power = 80%). Lower numbers are based on Type I error rate = .05, Type II error rate = .10 (power = 90%). Numbers within parentheses are based on a one-sided test, whereas numbers outside the parentheses are based on a two-sided test.

trial where patients will be randomly allocated with equal probability to control and experimental arms where the primary endpoint is pathologic complete response (pCR). It is assumed that the incidence of pCR on the control regimen is 35%. The investigator postulates that a 15% increase in the pCR on the experimental arm would be considered clinically meaningful. The number of patients needed is 460 patients or 230 per arm assuming a two-sided significance level of .05 and power = 90%.

Multiarm Trials

There are different approaches for designing trials that compare a time-to-event endpoint across treatment groups. One of the most common approaches is the 2 × 2 factorial design, where two different treatments are tested simultaneously in the same study without increasing the sample size. The drawback to this approach is that most factorial trials are limited by an interaction between the two treatment groups, and as a result such trials are usually underpowered.

Sequential Monitoring

Clinical trials are both costly and time-consuming, and the main motivation for monitoring clinical trials is for ethical and economical reasons. It is known that the Type I error rate increases with repeated testing of a hypothesis performed on the same data. For example, in a trial that tests the hypothesis that a new treatment prolongs survival compared with a control, the probability of committing a Type I error rate increases from 5% to 11% if the data are analyzed at three time points.

Therefore, most randomized clinical trials should include plans for stopping the trial early, if a treatment or combinations of therapies is found to be either harmful or useful to the study participants. Indeed, there are widely acceptable monitoring guidelines that are now considered part of standard statistical practice. These measures allow researchers to perform sequential analyses while the trial is still ongoing and data are maturing. The statistical tests are performed using boundaries so that the overall Type I error rate is preserved at the .05 level.

Although a Phase III trial is the definitive study providing evidence of efficacy for a novel drug, poor methodology can produce numerous biases that may invalidate the results of these studies. Such biases may arise at the design stage (such as choice of primary endpoint, number of patients required, low power, large clinical effect size, and inappropriate design), conduct of the trial (such as type of randomization used, imbalances in prognostic factors, selection of patients, and early stopping of the trial), analysis stage (such as excluding some patients who were not treated from the analysis, and violation of the proportional hazards model), and at the stage of the reporting and interpretation of results (such as test statistic for the primary endpoint that was not prespecified in the protocol, reporting *p* values on secondary endpoints, and on a subset of patients who were randomized).

Susan Halabi

See also Effect Size; Hazard Ratio; Sample Size and Power; Survival Analysis

Further Readings

Chen, T. T., & Simon, R. (1996). Extension of two-sided test to multiple treatment trials. *Communications in Statistics, Part A: Theory and Methods, 25,* 947–965.

Emerson, S., & Fleming, T. R. (1989). Symmetric groups sequential test designs. *Biometrics, 45,* 905–923.

George, S. L., & Desu, M. M. (1974). Planning the size and duration of a clinical trial studying the time to some critical event. *Journal of Chronic Disease, 27,* 15–29.

Halabi, S., & Singh, B. (2004). Sample size determination for comparing several survival curves with unequal allocations. *Statistics in Medicine 23,* 1793–1815.

Lakatos, E., & Lan, K. K. G. (1992). A comparison of sample size methods for the logrank statistic. *Statistics in Medicine, 11,* 179–191.

Lan, K. K. G., & DeMets, D. L. (1983). Discrete sequential boundaries for clinical trials. *Biometrika, 70,* 659–663.

McPherson, K., & Armitage, P. (1971). Repeated significance tests on accumulating data when the null hypothesis is not true. *Journal of the Royal Statistical Society Series A, 134,* 15–25.

O'Brien, P. C., & Fleming, T. R. (1979). A multiple testing procedure for clinical trials. *Biometrics, 35,* 549–556.

Pocock, S. J. (1979). Allocation of patients in clinical trials. *Biometrics, 35,* 183–197.

Rubinstein, L. V., Gail, M. H., & Santner, T. J. (1981). Planning the duration of a comparative clinical trial with loss to follow-up and a period of continued observation. *Journal of Chronic Disease, 34*, 469–479.

Simon, R. S. (1993). Design and conduct of clinical trials. In V. T. DeVita, S. Hellman, & S. A. Rosenberg (Eds.), *Cancer: Principles and practice of oncology* (pp. 418–440). Philadelphia: J. B. Lippincott.

Singh, B., Halabi, S., & Schell, M. (2008). Sample size selection in clinical trials when population means are subject to partial order. *Journal of Applied Statistics*, 35, 583–600.

Range-Frequency Theory

Range-frequency theory is a model of the psychological evaluation of stimuli in context developed by psychologist Alan Parducci. It posits that when a stimulus is rated alongside other stimuli, its rating will depend in part on where it ranks among the stimulus set. For example, a patient comparing doctors may provide a different rating to the same doctor when presented in a group of doctors perceived to be superior by the patient than when presented in a group of doctors perceived to be inferior by the patient.

Formally, the rating of a stimulus is described by a weighted average of the utility of the stimulus (relative to the range of utilities present in the context) and the rank of the stimulus (relative to the range of ranks present in the context). The theory is formulated mathematically as

$$R_i = J\left[w\left(\frac{x_i - x_{\min}}{x_{\max} - x_{\min}}\right) + (1-w)\left(\frac{\text{rank}(x_i) - 1}{N-1}\right)\right].$$

In this formula, R_i is the rating response given to the ith stimulus in a set of N stimuli. The first component of the formula gives some relative weight (w) to the difference between the utility of the stimulus, x_i, and the utility of the worst stimulus in the set, $x_{\min}$, relative to the range of utilities between the best and worst stimuli ($x_{\max} - x_{\min}$). The second component of the formula assigns the remaining relative weight ($1 - w$) to a similar comparison between the rank of the stimulus, rank(x_i), and the rank of the worst stimulus, relative to the range of ranks between the best and worst stimuli. The highest ranked stimulus receives a rank of N and the lowest ranked receives a rank of 1. The function J is a linear transformation. The weighting parameter, w, establishes the relative weight of the utility component and rank component of the model and is often fixed at .5 in modeling.

Range-frequency theory predicts that the same stimulus will receive different ratings depending on its relative position (rank) among the set of stimuli to be rated. For example, the rating of a health state B in a set of rank-ordered (high to low) health states A, B, C, D, E is predicted to be higher than the rating of the same health state in a set of rank-ordered (high to low) health states A, F, B, G, E, because the relative position of B is higher in the first set, where it is the second-best state, than in the second set, where it is the third-best state.

Figure 1 illustrates this key prediction of range-frequency theory. The curves plot the range-frequency theory predictions of ratings of stimuli against their underlying psychological utilities. In the solid upper curve, four stimuli are rated, with underlying utilities of .5, .6, .7, and .9; and in the lower dashed curve, four stimuli are rated, with underlying utilities of .5, .7, .8, and .9. Range-frequency theory predicts that the same stimulus with utility .7 will receive a higher rating in the first context (when it is second-ranked in the stimulus set) than in the second context (when it is third-ranked in the stimulus set). These predictions have been supported in numerous experimental studies in which stimulus context is varied.

Conversely, another important use of range-frequency theory is the recovery of "context-free" utilities from a set of ratings of stimuli made in a particular context. Because ratings are inherently contextual, they are difficult to compare across contexts. Range-frequency theory provides a "theory of the context." It can be fit to rating data in which the rankings of the stimuli are known and used to provide estimates of the underlying utilities (x_i) that are free of the impact of context and thus comparable. For example, in the health state ratings given above, the same underlying utility should be estimated for health state C despite contextual differences in ratings. Because the model is relatively parsimonious, such a modeling procedure is very often straightforward.

Alan Schwartz

See also Context Effects; Judgment; Utility Assessment Techniques

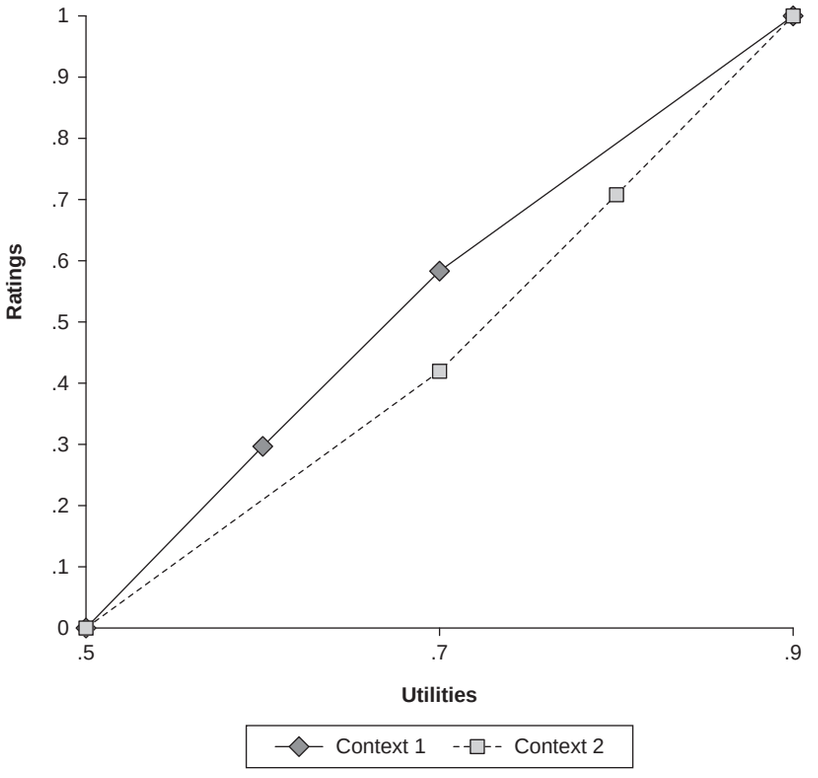


Figure 1 Range-frequency theory predictions

Further Readings

Bleichrodt, H., & Johannesson, M. (1997). An experimental test of a theoretical foundation for rating-scale valuations. *Medical Decision Making, 17,* 208–216.

Parducci, A. (1965). Category judgment: A range-frequency model. *Psychological Review, 72*, 407–418.

Parducci, A. (1974). Contextual effects: A range-frequency analysis. In E. C. Carterette & M. P. Friedman (Eds.), *Handbook of perception* (Vol. 2, pp. 127–172). New York: Academic Press.

Schwartz, A. (1997). Rating scales in context. *Medical Decision Making, 17,* 236.

Rank-Dependent Utility Theory

Originally called anticipated utility theory, rank-dependent utility theory is a generalized expected utility model. It was first developed by John Quiggin in 1982 as a solution for violations of stochastic dominance (where one outcome is always ranked above another), which expected utility theories were not able to resolve. It involves making decisions under risk and uncertainty. Expected utility theory uses the probabilities as the criteria to make decisions, while rank-dependent theory incorporates the role of weights, which is decided by the decision maker. In the condition that the decision maker only uses the probabilities for his or her weighting, the result of rank-dependent model and expected utility model would be the same. Thus, expected utility model is a special case of rank-dependent model, based on the concept of probability weighting. The theory separates probabilities from utilities, based on the assumption that decision makers rank order all outcomes to make decisions. This theory is also able to consider the characteristics of both pessimistic and optimistic decision makers.

Overview of Theory

Rank-dependent utility theory does not rely on the assumption of independence found in the expected utility theory. Rather, the theory assumes comonotonic independence. Independence in expected utility theory requires that if different common outcomes that have equal value are used, the decision maker's preference between these two choices would not be changed. That is, the decision maker would choose the same kind of options again. However, in rank-dependent utility, independence requirements are not useful. This is because each decision maker ranks the choices in a different way, based on his or her preferences. The comonotonic independence assumption indicates that using different common outcomes or treatments should not change the rank ordering of the outcomes. As a result, the decision maker's weighting of the choice is also considered. Rank-dependent utility theory relies on decision makers deriving probability weights from the entire probability distribution, not a single probability. Decisions are made after considering the entire outcome set, not one outcome at a time. Rank-dependent utility theory is able to operate under the assumption of comonotonic independence because it assumes that people rank the possible outcomes and transform the probabilities accordingly. After the ranking is complete, the order of outcomes is used to make a decision, not the individual probabilities.

Examples

Let's first consider an example of expected utility theory. Suppose a patient has an infection in her liver and needs to choose between two treatments to protect her from further infection. Treatment A has a 30% chance of protecting her for 5 years and a 70% chance of protecting her for 3 years. Treatment B has a 40% chance of protecting her for 4 years and a 60% chance of protecting her for 2 years. If this patient uses the expected utility approach, she would choose the treatment option that provides the maximum protection in the future for her, which in this example is Treatment A. Imagine that further research regarding these treatment options indicate that the actual protection levels are as follows: Treatment A has a 30% chance of 5 years' protection, a 60% chance of 3 years' protection, and a 10% chance of no protection; and Treatment B has a 40% chance of 4 years' protection, a 50% chance of 2 years' protection, and a 10% chance of 6 years' protection. Even with these modifications, Treatment A would still provide more overall protection to the patient. The independence criteria in expected utility theory indicate that since the overall value of Treatment A is still higher than Treatment B, the patient would choose Treatment A. However, using only independence neglects the fact that Treatment A has a 10% chance of not curing the patient at all. It also neglects the fact that Treatment B has actually the highest year of protection (10% chance of 6-year protection), which may be preferred by some patients.

To further clarify rank-dependent utility theory, consider the following example. A patient who suffers from asthma is offered two treatments, A and B. Treatment A has three outcomes associated with it, along with corresponding probabilities. The outcomes are no improvement (25% chance), 3 to 4 attacks per month (60% chance), and 1 attack per month (15% chance). Treatment B has three outcomes as well: no improvement (25% chance), 10 to 20 attacks per month (60% chance), and 0 attacks per month (15% chance). The patient's preference is to become completely cured. Since the patient's priority is in not having any attacks at all, the patient would choose Treatment B based on rank-dependent utility theory—even though Treatment B has a 60% chance of 10 to 20 attacks per month, which is relatively high compared with Treatment A, with a 60% chance of 3 to 4 attacks per month. As long as the patient's preference remains the same (not having any attacks at all), the treatment with the highest chance of completely curing the patient would be ranked first. With a different preference, another treatment may be chosen.

Consider a third example. A patient is told that she has a fatal disease. If she does not choose any of the following treatments, she would die immediately. She is presented with two treatment options. Treatment A has the following outcomes: a 30% chance of living 4 years, a 55% chance of living 6 years, and a 15% chance of living 8 years. Treatment B also has three outcomes: a 30% chance of living 2 years, a 55% chance of living 10 years, and a 15% chance of living 15 years.

Suppose the patient wants to be alive for at least 4 more years to finish writing her book. Based on rank-dependent utility theory, the patient would choose Treatment A. Even though Treatment B provides her with a longer life expectancy, she would choose Treatment A, as for her, living for at least 4 more years is the first priority.

Rationale for Ranking Outcomes

There are a number of reasons why rank-dependent utility theory is thought to accurately depict how people make decisions. The first explanation is that people operate under a number of perceptual biases. They put a large emphasis on extreme values, giving less attention to any that fall in the middle. This, in essence, is a way of ranking the outcomes rather than calculating an outcome's utility.

The second rationalization for ranking outcomes lies with the decision makers themselves. The individual difference between decision makers leads some to put higher weights on probable outcomes, whereas others prefer to look for a higher payout. They are often termed as *risk-seeking* and *risk-averse* decision makers. Thus, decision makers will rank all outcomes according to these scales.

The final explanation for why people rank order outcomes is situational. If a utility function for a certain problem is asymmetric or otherwise difficult to determine, the decision maker eases the process by simply rank ordering the outcomes. This adaptation allows for more efficient and effective decision making.

Lesley Strawderman and Arash Salehi

See also Expected Utility Theory; Probability

Further Readings

Oliver, A. (2003). Testing rank-dependent utility theory for health outcomes. *Health Economics, 12,* 863–871.

Perpiñán, J. M. A., & Prades, J. L. P. (2001). Testing the descriptive performance of the rank-dependent utility in the domain of health profiles. *Spanish Economic Review, 3,* 177–191.

Quiggin, J. (1982). A theory of anticipated utility. *Journal of Economic Behavior and Organization, 3,* 323–343.

Wakker, P., Erev, I., & Weber, E. U. (1994). Comonotonic independence: The critical test between classical and rank-dependent utility theories. *Journal of Risk and Uncertainty, 9,* 195–230.

Weber, E. U., & Kirsner, B. (1997). Reasons for rank-dependent utility evaluation. *Journal of Risk and Uncertainty, 14,* 41–61.

RATIONING

Healthcare rationing is making deliberate decisions about *what* beneficial healthcare will be made available and what will be withheld, *how much* of healthcare made available will be provided, and *to whom* such care will be given. Broadly construed, rationing decisions may be implicit or explicit. Such decisions involve limiting or withholding beneficial treatment from particular patients or from patient groups in a general population. It is usually because the demand for healthcare is greater than what is available that rationing becomes necessary. Discussions about rationing may include conceptual and theoretical issues, including choice among competing moral values that justify decisions. It may also include deliberations regarding frameworks, criteria, standards, and procedures for setting institutional and societal policies to guide rationing decisions in general. Rationing is often related or even synonymous to *resource allocation* and *priority setting* since these terms are similarly used in health policy and medical ethics to mean controlling access to healthcare.

The meaning of *rationing* in healthcare is different from its use during times of war when scarce goods are distributed to everyone in equal but limited portions. In contrast, healthcare rationing means setting limits or withholding access that varies between particular patients and patient groups. In both contexts, the aim is to prevent some of the undesirable societal effects of scarcity through deliberately controlling access to goods in high demand. Although some factors—not due to deliberate human decisions—limit or prevent access to effective beneficial care, these are not considered healthcare rationing decisions as defined in this entry. Rationing is a medical decision-making problem because limiting and withholding beneficial care is usually not acceptable, unless it is justified as necessary to make at least some other

beneficial care or life-saving treatment available to those in need. Rationing decisions try to make scarce resources accessible to as many as possible. The aim is to avoid depriving everyone by depriving some.

Rationing and Allocation

The term *allocation*, as suggested by Roger Evans, may be used to refer to decisions about making certain health resources available, for example, the quantity of hospital beds or diagnostic equipments, while the term *rationing* may be used to refer to decisions about what patients should receive from among previously allocated health resources. Decisions about the quantity of health resources to make available affect decisions about which patient will receive treatment. The collective decisions of physicians about which patient needs a specific health resource could also affect decisions about how many of such resources should be made available. Limiting the number of life-saving machines to be made available, as Peter Ubel and Susan Goold point out, is as tragic as deciding which needy patient should be treated. All decisions to withhold life-saving treatment from a patient are equally tragic. The term *allocation* may be used to refer to both levels of decision making as done when the terms *microallocation* and *macroallocation* are used. In this entry, both are considered rationing. Thus, microallocation refers to rationing decisions in the context of delivering specific care to patients, for example, when physicians decide which patient will be given access to scarce renal dialysis beds in a particular clinical setting. On the other hand, macroallocation refers to rationing decisions done at the level of institutions when managers, policy makers, or insurance companies make recommendations about the number of dialysis machines to avail, what criteria should be used for prioritizing potential patients, whether dialysis treatment costs should be considered reimbursable, and how much reimbursement should be given.

Patient-Level Rationing

A patient may ration his or her own access to beneficial healthcare by deciding to skip going to a physician and by purchasing nonprescription drugs to avoid spending more for consultation services. This is done so that the money saved can be used to buy other things the patient values more or perhaps to save for the patient's future needs that include medical care.

Physicians also make rationing decisions with respect to their patients. This is known as *bedside rationing*. Ubel and Goold define that physicians ration when they (a) withhold, withdraw, or fail to prescribe care they consider best for their patients; (b) make such decisions with the intention of promoting the financial interests of all others except their patient; and (c) control the use of beneficial care. Consider when a physician decides to administer Treatment A rather than Treatment B even though the latter is slightly better for the patient. If Treatment B costs much more than Treatment A, and the physician thinks that the benefits of Treatment B are not worth the additional cost, then this is a case of bedside rationing. Although the above definition seems to highlight the undesirable aspect of rationing as withholding the best treatment, the possibility that such decisions could be justified on moral grounds is not necessarily excluded. In the given case, some beneficial treatment (although not the best) was provided to the patient to save money in order to benefit society as a whole. Moral justification of rationing decisions require other reasons (other than saving money) that explain why a physician must withhold the best care to a patient for the sake of benefiting society.

Physicians are normally obligated to provide or prescribe the appropriate beneficial care to all their patients, but they could also limit or withhold care from some patients if the resources available are not enough for all of them. Rationing the use of a temporary artificial heart (as bridge device) to patients in an organ transplant waiting list is presented by Paul Menzel to illustrate three ways of reconciling a physician's obligation to his or her patients and making the most of scarce resources: (1) following predetermined guidelines in making rationing decisions, (2) allowing a patient to give prior consent to forgo the use of a bridge device, and (3) basing the rationing decisions on considerations of fairness and justice. However, it is noted that it is difficult to come up with guidelines that will always be relevant to particular patients, to accurately discern what trade-offs patients would

be willing to take, and to find agreement about what constitutes justice.

Institution-Level Rationing

The public resources available to fund social services are usually less than the demand. Deciding to allocate a certain amount of government money to fund healthcare services will affect how other social services are funded. How education or social welfare services are funded affects public health, which influences the demand for healthcare services. Thus, how the government sets the right budget for healthcare versus other public needs constitutes rationing. The more direct rationing decision at the institutional level occurs when health authorities further divide the health budget to finance different health services. It is difficult to limit and withhold funding for certain health measures versus others that will affect satisfaction of health needs of different patient groups. For example, giving more funds to set up new health facilities could increase the number of people who can be served, but funding salary increases for personnel of existing health facilities can improve quality of care. Health authorities are also faced with the decision to allocate more funds to preventive measures, such as inclusion of new effective vaccines to the national program, versus scaling up curative services. Such rationing decisions on the level of government health institutions would definitely affect how much healthcare services different patient groups could access.

In countries where patients share costs with institutions (such as government health service, insurance companies, and employers), rationing involves decisions regarding what and how much healthcare services to cover. How insurance companies decide the amount of insurance premium people should pay also constitutes rationing. Some insurance schemes allow individuals to select premium rates that correspond to the types and amount of health services they wish to have. However, the final decision on what services will actually be paid for by insurance usually depends on the judgment of the decision makers of insurance companies, who may or may not refer to predetermined guidelines. The decision on what services and medicines to cover are usually based on proven effectiveness versus cost. Even though some treatments are proven to be effective, how much they cost is an important consideration for maintaining the viability of the insurance scheme. Failure to set limits on costs to meet every need for beneficial care could decrease the capacity of insurance companies to fulfill their obligations to their clients.

The decision of hospital managers regarding how many physicians with specific expertise they will hire, how many and what types of medical equipments they will acquire, and whether they will charge patients advance payments prior to admission constitute rationing. Such decisions affect access of different patient groups to the type and amount of beneficial treatment they would need. For example, poor patients will not be able to avail services of hospitals that charge fees prior to admission, but not doing so may burden hospitals with having to shoulder the expenses of patients who cannot pay.

Rationing Approaches and Mechanisms

There are several approaches to making rationing decisions in the peer-reviewed medical literature. Some of the more dominant approaches include the following: (a) aiming to efficiently use available resources, (b) aiming to distribute resources fairly, and (c) aiming to achieve equity in distribution of resources.

Efficiency requires determining which among the healthcare options brings the most benefit per health resource unit used. To do this would be to engage in cost-effectiveness analysis (CEA). CEA determines whether Treatment A could lengthen lives or make patients healthier compared with Treatment B at the same cost. To compare health benefits with other benefits that available resources could produce is to engage in cost-benefit analysis (CBA). CBA determines whether the health benefits produced by a treatment are of greater value than other benefits that the same amount of money could produce if spent elsewhere. Choosing only to provide care that brings maximum benefit implies withholding other healthcare options or limiting access to achieve the best overall benefit with the resources available. The Oregon Health Plan priority list is a well-known attempt to achieve efficient rationing. In trying to expand coverage to more poor Oregonians, the plan ranked various

services in the Medicaid program from the most cost-effective to the least cost-effective.

An approach that emphasizes the importance of fairness in the process of making rationing decisions has been developed by Norman Daniels and James Sabin. Decision-making procedures are fair if the reasons used in the deliberation process are considered relevant and justifiable, especially by those affected. It is thus a matter of fairness to involve patients, or their representatives, in making decisions about limiting treatments that benefit them. It is also important to present relevant evidence about the comparative benefits of the healthcare options in consideration.

Limiting and withholding healthcare must be equitable. This means that equal care should be given to those with equal need. Patients in the same condition should get the same limits to care or the same denial for the sake of those in greater need. There may be other ways to define equity in rationing.

Peer-reviewed medical literature enumerates the number of ways in which rationing is implemented: (1) by deterring or obstructing demand for healthcare, (2) by delaying delivery of care, (3) by deflecting demand, (4) by diluting healthcare demand, and (5) by denying care. For example, charging fees for services could lessen demand. Waiting lists, mandatory referral from a general practitioner to gain access to a specialist, and complicated paperwork make access inconvenient. Informing patients about outcome may change their mind about demanding treatment. Prescribing cheaper generic drugs is an example of rationing by dilution. Denial of care may be done as a result of a treatment being given lesser priority.

Allen Andrew Alvarez

See also Cost-Benefit Analysis; Cost-Utility Analysis; Evidence-Based Medicine

Further Readings

Brock, D. (2007). Health care resource prioritization and rationing: Why is it so difficult? *Social Research: An International Quarterly of Social Sciences, 74*(1), 125–148.

Daniels, N., & Sabin, J. E. (2002). *Setting limits fairly: Can we learn to share medical resources?* Oxford, UK: Oxford University Press.

Evans, R. W. (1983). Health care technology and the inevitability of resource allocation and rationing decisions: Part I. *Journal of the American Medical Association, 249*(15), 2047–2053.

Hadorn, D. C. (1991). Setting health care priorities in Oregon: Cost-effectiveness meets the rule of rescue. *Journal of the American Medical Association, 265*(17), 2218–2225.

Harrison, S., & Hunter, D. J. (1994). *Rationing health care*. London: Institute for Public Policy Research.

Kitzhaber, J., & Kemmy, A. M. (1995). On the Oregon trail. *British Medical Bulletin, 51*(4), 808–818.

Menzel, P. (2007). Allocation of scarce resources. In L. P. Francis, R. Rhodes, & A. Silvers (Eds.), *The Blackwell guide to medical ethics* (pp. 305–322). Malden, MA: Blackwell.

Russell, B. J. (2002). Health-care rationing: Critical features, ordinary language, and meaning. *Journal of Law, Medicine & Ethics, 30*(1), 82–87.

Ubel, P. A., & Goold, S. (1997). Recognizing bedside rationing: Clear cases and tough calls. *Annals of Internal Medicine, 126*(1), 74–80.

Ubel, P. A., & Goold, S. D. (1998). Rationing health care: Not all definitions are created equal. *Archives of Internal Medicine, 158*(3), 209–214.

Receiver Operating Characteristic (ROC) Curve

In practice, the outcomes of diagnostic tests are mostly interpreted and operationalized as binary—that is, as positive or negative—for the presence of a target condition. However, the actual outcome of a test is rarely a binary one. For example, the results of laboratory tests are typically measured on continuous scales, and the same applies to measures summarizing scans with modern imaging modalities, such as the standardized uptake value in positron emission tomography. When test results are measured on an explicitly defined and observed scale as in these examples, a binary outcome is defined on the basis of an explicit threshold for test positivity. When tests involve interpretation by a human observer, a similar model with a threshold for test positivity has been used widely. However, in this case the test result and the threshold are conceptualized as occurring on a latent scale, measuring the

interpreter's degree of suspicion about the presence of the target condition.

Because binary test outcomes are obtained through thresholds on an observed or a latent scale, these thresholds affect *all* the usual measures of diagnostic and predictive performance, including sensitivity, specificity, positive predictive value, negative predictive value, and likelihood ratios. The dependence of measures of test performance on the threshold for test positivity is a fundamental tenet of diagnostic test evaluation. In particular, this dependence induces the well-known trade-off between test sensitivity and specificity as the threshold for positivity is moved across its possible values.

Figure 1 shows hypothetical distributions of test results for individuals with and without the target condition and a threshold for test positivity. If the likelihood of having the target condition increases with the test score, the sensitivity of the test is measured by the area under the "condition-present" curve, to the right of the threshold. Similarly, the specificity is measured by the area under the "condition-absent" curve, to the left of the threshold. In the formulation of Figure 1, the sensitivity of the test is a decreasing function of the threshold value, and the specificity of the test is an increasing function of the threshold value.

The receiver operating characteristic (ROC) curve of a test is the graph of all possible pairs of (1 – specificity, sensitivity) obtained by varying the positivity threshold across its entire range of possible values. As can be seen from Figure 1, when the threshold moves to the left end of its range, sensitivity becomes 1 and specificity becomes 0. The converse occurs when the threshold moves to the right end of its range. Figure 2 shows a typical ROC curve.

Interpretation of ROC Curve

A test is said to have a good performance if high sensitivity is achieved while maintaining high specificity. In the limiting case, if the separation of the two distributions in Figure 1 became nearly complete, a perfect test would result with both sensitivity and specificity tending to 1. In that case, the ROC curve would be degenerate and would pass through the ideal point (0, 1). Conversely, an uninformative test would result if the distributions in Figure 1 coincided. In that case, the sensitivity and specificity would add to 1 for all thresholds and the ROC curve would be the main diagonal of the square.

Summaries of the ROC Curve

The ROC curve incorporates information on the diagnostic performance of a test across the range of possible thresholds. To facilitate the evaluation and comparison of tests, several summaries of the ROC curve have been proposed in the literature. Most commonly used among them is the area under the curve (AUC), which can be interpreted as an average of test sensitivity taken over all specificity values. If only a subset of specificity or sensitivity values is of interest in a given setting, partial AUCs can be considered. For example, high values of specificity are typically of interest in

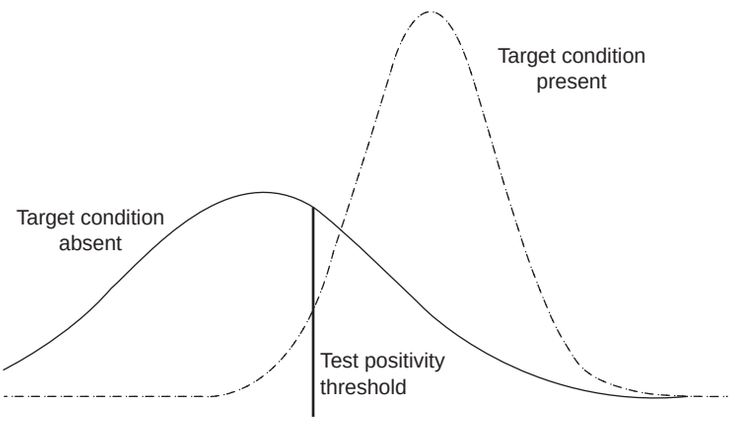


Figure 1 Fundamental conceptualization of test outcomes

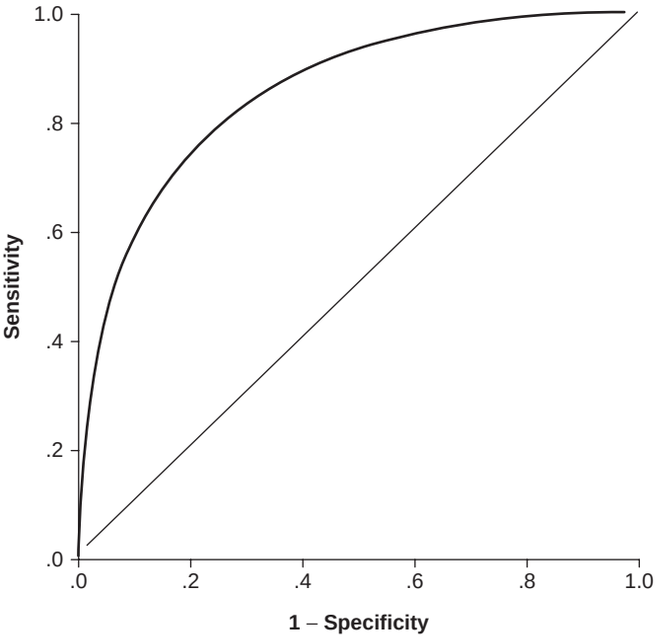


Figure 2 Typical receiver operating characteristic (ROC) curve

screening, and the corresponding partial AUC may be used as long as the range of values can be determined as part of the design of the ROC study. In signal detection theory, the AUC is the probability that, if a pair of subjects with and without the target condition is selected at random, the subject with the target condition will be ranked correctly by the test. Other summaries of the ROC curve include values of sensitivity corresponding to selected values of specificity (and vice versa) and optimal operating points, selected to minimize cost function criteria. If the curve is smooth and the average overall cost is a linear combination of the cost of each of the four categories of diagnostic decisions—true positive (TP), false positive (FP), true negative (TN), false negative (FN)—the optimal point is the point where the slope of the tangent to the curve is equal to $(1 - p)/p \times R$, where p is the prevalence and R is the ratio of cost differences = $(C_{TN} - C_{FP})/(C_{FN} - C_{TP})$.

Statistical Inference for ROC Curves

ROC curves can be estimated from studies in which tests with ordinal categorical or continuous results are evaluated against a binary reference standard indicating the presence or absence of the target condition. For example, ROC curves for diagnostic imaging modalities, which rely on human interpretation, are estimated on the basis of data on the degree of suspicion about the presence of the target condition elicited from the interpreter (reader). Such data are often collected on an ordinal categorical scale with five or seven categories, although continuous or quasi-continuous scales (e.g., percent probability of malignancy) have also been proposed. For laboratory tests, ROC curves are estimated on the basis of the typically continuous data provided by the test.

Basic Formulations and Regression Analysis

For ordinal categorical test results, ROC curves are commonly estimated using parametric models, which assume that the observed ordinal categorical responses are discretized values of an underlying latent variable. A parametric distribution, such as Gaussian or logistic, is assumed on the latent variable for cases with and without the target condition and estimation proceeds via maximum

likelihood. The "binormal" ROC model effectively assumes Gaussian distributions on the latent variable (or a monotone transformation of it) and leads to ROC curves governed by two parameters: (1) a = the difference in the means of the two distributions divided by the standard deviation of the "condition present" group and (2) b = the ratio of the standard deviations of the two distributions. The estimated curves are then used to derive summary measures and to perform comparisons. In particular, the AUC of a binormal curve is equal to $a/\sqrt{1+b^2}\Big]$.

A flexible and efficient approach to parametric ROC analysis relies on ordinal regression models in which the binary truth status is the covariate and the degree of suspicion is the response variable. Furthermore, covariates can also be incorporated in the regression model. The general form of the ordinal regression models specifies that the cumulative probability of observing a response up to suspicion level j (j = 0, 1, 2, 3, 4, when a five-category scale is used) has the following expression:

$$[P(Y \le j|X)] = h[(\theta_j - \alpha' X)\exp(-\beta' X)],$$

where θ_j is a cutoff point on an underlying latent scale, X is a vector of covariates, h is a nondecreasing link function, α is a vector of location parameters, and β is a vector of scale parameters.

One of the covariates in the model represents the binary reference standard. The link function, h, can be chosen to correspond to the distributional assumptions on the underlying scales of the ROC model. Thus, a probit link corresponds to the binormal ROC curve, and a logit link corresponds to the logistic ROC curve. The parameters (a, b) of the ROC curve are algebraic functions of the parameters of the ordinal regression model. Estimates and confidence intervals for the ROC quantities can be obtained on the basis of the corresponding quantities from the ordinal regression model by application of the delta method or bootstrapping techniques. Using the ordinal regression approach, we can examine models with covariates indicating characteristics of patients and readers, as well as interaction terms. For example, to account for possible differences in accuracy among readers, we can construct sets of indicator variables for readers and include them in the location and scale of the ordinal regression model. Similarly, appropriately constructed covariates can be included to examine the effects of patient clinical characteristics. Models are compared on the basis of deviance statistics and model fit can be assessed using analysis of appropriately constructed residuals.

For ordinal categorical data, the AUC can also be estimated nonparametrically, using the Wilcoxon statistic. Comparisons of areas can then be made using nonparametric methods.

An alternative approach to ROC analysis is based on the notion that an ROC curve can be understood as a graph of sensitivity as a function of specificity (or conversely). If Y_D and $Y_{\bar{D}}$ represent the test results in the "condition-present" and "condition-absent" groups, respectively, and Y_D and $F_{\bar{D}}$ represent the corresponding survivor functions (e.g., $F_D(x) = P(Y_D > x)$), then the sensitivity at threshold c is equal to $F_D(c)$, and the specificity is equal to $1 - F_{\bar{D}}(c)$. The ROC curve of the test is the graph $\mathrm{ROC}(t) = F_D(F_{\bar{D}}^{-1}(t))$, as t ranges in the interval [0, 1]. It can be shown that $\mathrm{ROC}(t)$ is just the conditional probability that Y_D is greater than $Y_{\bar{D}}$, when $Y_{\bar{D}}$ is the $(1-t)$th quantile of the "condition-absent" distribution. Using this formulation of the ROC curve, inference can be based on the indicator variables $I_{\{Y_D \ge Y_{\bar{D}}\}}$ using binary regression analysis. Covariates can also be considered, for example, by assuming a parametric form for the ROC curve such as $\mathrm{ROC}(t) = g\left\{\sum_{k=1}^{K}\gamma_k h_k(t) + X\beta\right\}$, for a link function g, basis function $h_1, \ldots, h_K$, unknown parameter vector γ, covariate vector X, and parameter vector β. In particular, if the link function g is equal to Φ, the cumulative normal distribution function, and the basis functions are chosen as $h_1(t) = 1$, $h_2(t) = \Phi^{-1}(t)$, and $X\beta = 0$, the ROC curve takes the binormal form.

ROC Analysis With Correlated Data

Correlated ROC data arise in many studies of diagnostic accuracy because participants are typically examined with more than one test and, in the case of imaging, scans are typically interpreted by multiple readers. Clustered test results occur in

many other settings. For example, clustering occurs when separate test results are obtained in multiple parts of a study participant's body, such as blood vessels or liver segments, or when studies are conducted in multiple institutions. Methods for the analysis of correlated and clustered ROC results account for the correlation in the data and may also address variations among clusters. Several approaches have been developed for this type of analysis, including generalized estimating equations for ordinal regression and semiparametric ROC models, hierarchical ordinal regression models, mixed models, jackknife and bootstrap methods, and fully nonparametric methods. Several of these methods have also been extended to handle missing data, particularly in the settings giving rise to verification bias.

ROC Analysis in the Evaluation of Predictive Accuracy

Methods from ROC analysis are also used in the evaluation of the predictive ability of tests and predictive models. Although the use of ROC methods in this area is a matter of considerable current attention and debate, important advances have already been made. For example, the *c* statistic, commonly used to assess the predictive ability of a logistic regression model, is a nonparametric estimate of the area under the ROC curve. This curve would be obtained by treating the model-based estimated probability of a response as the test result and the actual binary response as the reference standard.

Time-dependent ROC analysis has extended the traditional ROC model to settings in which the reference standard is not contemporaneous to the test result but will be observed at some future point in time. Thus, the new methodology permits an assessment of how the predictive ability of a marker may vary over time. For example, a biomarker assessed before the beginning of therapy may be used to predict patient survival. The predictive ability of the marker can be assessed by estimating an ROC curve for each future time point *t*. For the curve corresponding to time *t*, the test result will be the baseline value of the biomarker, and the binary reference standard will be an indicator of whether death occurred by *t*. The statistical analysis of time-dependent ROC curves accounts for patient dropout and censoring.

Software for ROC Analysis

Basic ROC analysis, including curve estimation and curve comparison, can be performed using utilities available in the commercially available packages Stata and SAS. Parametric ROC analysis for ordinal categorical and continuous data can also be performed via the widely used and freely available ROCKIT suite of programs. Hierarchical-model ROC analysis can be performed using BUGS programs, and mixed-model analysis can be performed in SAS or Stata. A variety of specialized programs are also available, including suites of subroutines in *R* and *S*-plus for performing ROC analysis under the alternative formulation described above.

Design of ROC Studies

Commonly used methods for determining sample size required for inference about a single ROC curve and its summaries make parametric assumptions about the test result, observed or latent, to develop approximations of the distribution of the ROC summary measure of interest. The approximations are then used to derive the sample size necessary to achieve the desired expected length of the confidence interval or the desire power for a hypothesis test about the true value of the ROC summary. A similar approach has been taken in the construction of methods for determining the required sample size for inference about two curves. An important consideration in this setting is whether the test results for the two curves are correlated.

The computation of sample sizes of cases and readers in multireader studies is generally considerably more complex. A commonly used method is based on a mixed model in which the response variable is an estimate of a summary of the ROC curve such as the AUC. The approach requires specifications of average values for the various pairs of correlations that are present in the data. Bayesian and simulation-based methods are also available in the literature.

Sample size calculations for ROC analysis can be carried out in the commercially available

software PASS, the freely available software ROCKIT, and a variety of specialized software available from individual researchers.

Constantine Gatsonis

See also Diagnostic Tests

Further Readings

Dorfman, D., & Alf, E. (1969). Maximum likelihood estimation of parameters of signal detection theory and determination of confidence intervals: Rating method data. *Journal of Mathematical Psychology, 6,* 487–496.

Hanley, J. (2006). Receiver operating characteristic curves. In P. Armitage & T. Colton (Eds.), *Encyclopedia of biostatistics* (pp. 4523–4529). New York: Wiley.

Hanley, J. A., & McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic curve. *Radiology, 143,* 129–133.

Heagerty, P. J., Lumley, T., & Pepe, M. S. (2000). Time-dependent ROC curves for censored survival data and a diagnostic marker. *Biometrics, 56*(2), 337–344.

Ishwaran, H., & Gatsonis, C. (2000). A general class of hierarchical ordinal regression models with applications to correlated ROC analysis. *Canadian Journal of Statistics, 28,* 731–750.

Metz, C. (1986). ROC methodology in radiologic imaging. *Investigative Radiology, 21,* 720–733.

Obuchowski, N. (1995). Multi-reader multi-modality ROC studies: Hypothesis testing and sample size estimation using an ANOVA approach with dependent observations. *Academic Radiology, 2,* S22–S29.

Obuchowski, N., & McClish, D. (1997). Sample size determination for diagnostic accuracy studies involving binormal ROC curve indices. *Statistics in Medicine, 16,* 1529–1542.

Pepe, M. (2000). Receiver operating characteristic methodology. *Journal of the American Statistical Association, 95,* 308–311.

Pepe, M. (2003). *The statistical evaluation of medical tests for classification and prediction.* New York: Oxford University Press.

Toledano, A., & Gatsonis, C. A. (1996). Ordinal regression methodology for ROC curves derived from correlated data. *Statistics in Medicine, 15,* 1807–1826.

Zhou, X. H., Obuchowski, N., & McClish, D. (2002). *Statistical methods in diagnostic medicine.* New York: Wiley.

Recurrent Events

In many biomedical studies, subjects may experience the outcome of interest more than once over a period of observation; outcomes of this sort have been termed *recurrent events*. For example, patients with cerebrovascular disease may experience repeated transient ischemic attacks, and HIV patients may experience recurrent opportunistic infections. Other examples of recurrent events include infections, myocardial infarctions, tumor metastases, and disease relapses/remissions. The structure of recurrent events is that of naturally ordered failure time data, and the different events "within" an individual are correlated. These types of processes arise frequently in medical studies, where information is available on many individuals, each of whom may experience transient clinical events repeatedly over a period of time. For instance, asthma is occurring more and more frequently in very young children. Some new prevention trials have been set up with such children randomized to placebo or drug, and the asthma events are recorded. Typically, a patient has more than one asthma event. The different events are thus clustered within a patient and are ordered in time. This ordering can be taken into account in the model. Such data can be presented using different timescales.

Time-to-Event Data

More generally, a problem frequently faced by applied statisticians is the analysis of time-to-event data. Examples of such data arise in diverse fields such as medicine, biology, public health, epidemiology, engineering, economics, and demography. Here interest is, for example, on analyzing data on the time to death from a certain cause, duration of response to treatment, time to recurrence of a disease, time to development of a disease, or simply time to death. Recurrent event data correspond to successive observations of time-to-event data.

Recurrent Events Data

Frequent objectives in analyzing recurrent event data include (a) understanding and describing individual event processes, (b) identifying and

characterizing variation across a population of processes, (c) comparing groups of processes, and (d) determining the relationship of fixed covariates, treatments, and time-varying factors to event occurrence.

Two fundamental ways of describing and modeling event occurrences are (1) through event counts and (2) through gaps between successive events or through calendar times. Models based on counts are often useful when individuals frequently experience the events of interest and their occurrence does not alter the process itself, as, for instance, the asthmatic attacks. This is different from processes where the events may substantially alter the condition of the individual, thus affecting the event process in the future, such as the development of new sites of metastatic disease in cancer trials. The framework for the analysis of event counts is the Poisson process. To analyze recurrent event data, the focus can be placed on time-between-events (i.e., gap times) or time-to-events models (i.e., calendar times).

In the gap time representation, the time at risk starts at 0 and the time at risk for a particular event is the time from the end of the previous event (e.g., asthma attack in respirology trials) or since the entry of the subject in the study from the first event to the start of the new event (start of the next asthma attack). The waiting (gap) time between successive events is statistically independent—that is, an individual is "renewed" after each event occurrence. In the calendar time representation, the start of the at-risk period is not reset to 0 but to the actual time since entry to the study, but the length of the at-risk period is the same. The gap timescale is more appropriate when studying the recurrent event rate as a function of time since the last event, whereas the calendar timescale keeps track of actual time. This is illustrated in Figure 1.

In describing recurrent event data, a complex data structure is sometimes needed to keep track of the sequence of events within a patient. A particular patient has different periods at risk during the total observation period, which are separated either by an asthmatic event that lasts one or more days or by a period in which the patient was not under observation. The start and end of each such risk period is required, together with the status indicator, to denote whether or not the end of the risk period corresponds to an asthma attack (see Figure 1).

Frequently, it is difficult to observe the precise time of events, and all that is known is how many events occurred between successive examination times. If, for instance, examination times (or the visits) vary between patients, then the times between assessments must be taken into account; this is referred to as *interval-censored data*. Such data often arise in medical contexts such as studies

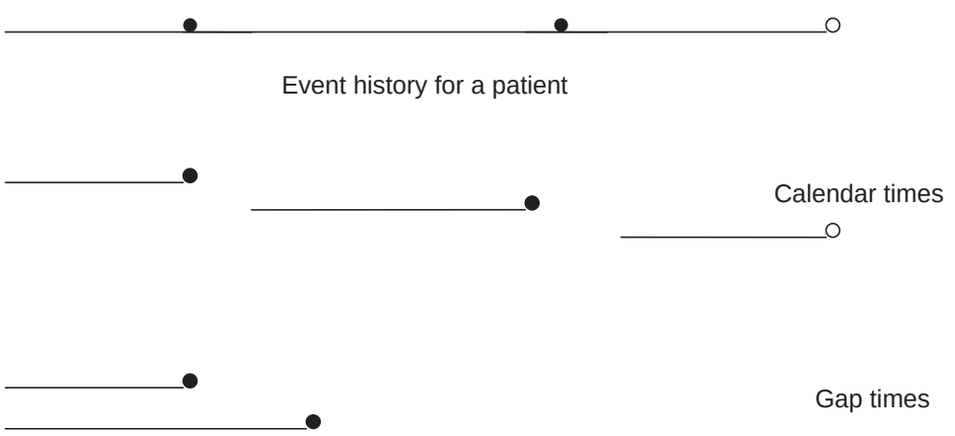


Figure 1 Event history for a patient with recurrent events together with the calendar times and the gap times

Note: ● represents an event, ○ a censoring time, and ━ the not at risk periods.

of metastatic cancer, where new metastases are detectable on magnetic resonance imaging.

Survival Analysis

Analysis of Time to Event

The specific feature that makes survival analysis different from classical statistical analysis is data censoring. Typically, the survival time is unknown for some of the subjects, the only information available being that the subject has survived up to a certain time. Thereafter, the subject is no longer followed up. This type of censoring is called *right censoring*. The analysis of survival experiments is complicated by issues of censoring, where an individual's life length is known to occur only in a certain period of time, and by truncation, where individuals enter the study only if they survive a sufficient length of time or individuals are included in the study only if the event has occurred by a given date. The classical model used to analyze survival times is the Cox proportional hazards regression model.

Analysis of Recurrent Events

Cox proportional hazards models are used for analyzing survival data; however, these methods are based on the assumption that the survival times of distinct individuals are independent of each other. This assumption may be suspect in many cohorts; for instance, if we study the times to occurrence of different nonlethal diseases within the same individual, it is quite probable that there is some association within the subject of survival times in the sample. A model that is becoming increasingly popular for modeling association between recurrent survival times is the use of a frailty model. A frailty is an unobservable random effect shared by recurrent events within each subject. In this model, subjects with a large value of the frailty will experience the event at earlier times than subjects with small values of the random effect. The most common model for a frailty is the so-called shared-frailty model extension of the proportional hazards regression model. In recent years, a number of papers appeared extending the survival models to models that are suitable to handle more complex survival data. In this context, a lot of attention has been paid to frailty models, providing a powerful tool to analyze clustered or recurrent survival data. In the frailty models, the hazard function partly depends on an unobservable random variable thought to act multiplicatively on the hazard so that a large value of the variable increases the hazard. For the jth ($j = 1, \ldots, n_i$) observation of the ith individual ($i = 1, \ldots, G$), let T_{ij} denote the survival times under study, and let C_{ij} be the corresponding right-censoring times, and the observations are $Y_{ij} = \min(T_{ij}, C_{ij})$. Our frailty model specifies that the hazard function conditional on the frailty is $\lambda_{ij}(t \mid Z_i) = Z_i \lambda_0(t) \exp(\beta' X_{ij})$, where $\lambda_0(t)$ is the baseline hazard function, $X_{ij} = (X_{1ij}, \ldots, X_{pij})$ denotes the covariate vector for the jth observation of patient i, and β' is the corresponding vector of regression parameters. Conditionally, on the frailty Z_i, the failure times $T_{i1}, \ldots, T_{in_i}$ are assumed to be independent. It is often assumed that the Z_is are independently and identically distributed from a gamma distribution with mean 1 and unknown variance θ. Large values of θ signify a closer positive relationship between the observations of the same subject and greater heterogeneity among individuals. For instance, one may be interested in data from carcinogenicity experiment on the times (T_{ij}) to development of mammary tumors for female rats. Rats are exposed to a carcinogen and further conditioned for several days prior to randomization to receive either a treatment or control (X_{ij}). A follow-up period of several days began after randomization during which they are examined every few days for the development of new tumors. The baseline hazard function $\lambda_0(t)$ corresponds to the risk of developing a new tumor at Time t for a rat having a null value of its covariate (i.e., for the control group, $X_{ij} = 0$). In this case, the random effect, Z_i, specific to each rat corresponds to a means to take into account the dependence between different times of observation for each rat. This random effect, Z_i, represents the unobservable factors that create heterogeneity in the times across rats. In this case, the times to tumors within rats are more similar than the times to tumors from different rats.

The main interest could be, for instance, to assess the effect of a new treatment (which is measured by the estimation of parameter β) and to know if this treatment has a significant effect on the development of a new tumor. There are several statistical approaches available (called inferences) to estimate these quantities and obtain conclusions. More specifically, in a frequentist approach, estimates of the parameters of interest are obtained by maximizing the marginal log likelihood. The EM algorithm approach is used for the inference in the semiparametric gamma frailty models. An alternative approach to fit semiparametric gamma frailty models is based on penalized partial likelihood maximization. It is shown, however, that this technique leads to the same estimates as the EM algorithm in the case of the semiparametric gamma frailty model. The penalized partial likelihood approach, however, can also be extended to fit a semiparametric model with normal distributed random effects. These approaches have some general drawbacks. In particular, the convergence can be slow and a direct estimate of the variance of the frailty term is not provided. Furthermore, these methods cannot be used to estimate the hazard function, which has often a meaningful interpretation in epidemiology. An alternative method is the penalized full likelihood based on the nonparametric estimation of the baseline hazard as opposed to the penalized partial likelihood. Bayesian techniques based on Gibbs sampling can also be used to fit gamma frailty models with nonparametric baseline hazard (i.e., semiparametric frailty models). In the classical Bayesian approach, the frailties are considered as parameters.

Ordering is present in recurrent event data sets. If there exists an ordering in time, we can still use techniques that do not take the ordering into account, although more specific models might be more relevant. Dependence between recurrent events is mostly modeled using time-varying covariates.

Counting Process Formulation

Modeling of recurrent events can be approached in a number of ways. For purposes of both modeling and statistical analysis, the concept of counting processes is especially useful. For a single recurrent event process starting for simplicity at $t = 0$, let $0 \leq T_1 < T_2 < \ldots < T_k$ denote the event times, where T_k is the time of the kth event. The associated counting process $\{N(t), 0 \leq t\}$ records the cumulative number of events generated by the process; specially, $N(t) = \sum_{k=1}^{\infty} 1(T_k \leq t)$ is the number of events occurring over the time interval $[0, t]$. More generally, $N(s, t) = N(t) - N(s)$ represents the number of events occurring over the interval $(s, t]$. In this notation, square or round brackets are used to indicate whether the endpoint of an interval is in or not in the interval, respectively. Counting processes, as defined here, are right continuous—that is, $N(t) = N(t+)$, with $t+$ a time infinitesimally larger than t. Figure 2 illustrates a realization of an event process in terms of its counting process $N(t)$, with jumps of height 1 and whose value at time t are known infinitesimally after t.

Extensions

Recurrent Events With Termination

The time frame for an individual's repeated event process may depend on other "terminating" events, such as death. Often the recurrence of serious events, such as tumors and opportunistic infections, is associated with an elevated risk of death. In this context, the usual assumption of noninformative censoring of the recurrent event process by death, required by most statistical analyses, can be violated. That means, for instance, that death (the censoring event) precludes any further occurrence of the recurrent event. Examples are ubiquitous and include many settings involving patients with a serious disease that is associated with both recurrent complications and high mortality. In neurovascular trials, for example, one may be interested in reducing the occurrence of transient ischemic attacks and mild strokes, but death from major strokes or any other cause may also occur. In oncology, one may be interested in characterizing the use of health services following diagnosis of cancer, but use of such services terminates in death. This dependence should be accounted for in the joint modeling of recurrent events and deaths. For instance, consider the study of patients with follicular lymphoma (FL), undergoing episodic relapses of FL. The course of this disease is usually characterized by a response to initial treatment, followed by relapses, sometimes associated with high-grade

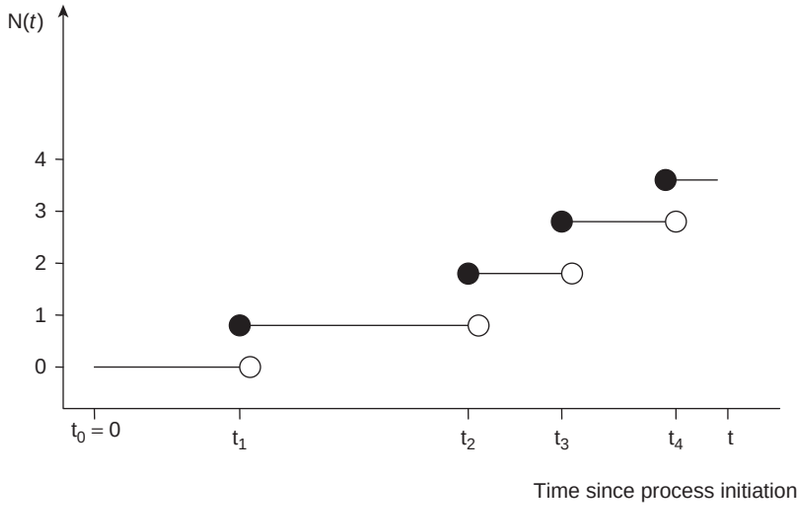


Figure 2 Counting process representation of data on recurrent events with • an event time and ○ a censoring time

non-Hodgkin's lymphomas. After the initial treatment, each patient can be monitored regularly for routine visits, and the presence of FL relapses is notified at each visit. Estimation of the risk of recurrence allows for better planning of follow-up schedules after diagnosis or first treatment and permits clinicians to determine therapeutic approaches based on the patient's risk of relapse. Furthermore, FL relapses may increase the risk of death. As a result, there is an association between the FL relapses process and the survival process, which precludes the use of standard analyses of recurrent events. Specifically, those subjects experiencing FL relapses at the highest rate are typically observed for shorter periods of observation due to mortality. We can thus consider the FL relapses and the terminal event process jointly, in a joint frailty model setting.

Time-Varying Frailties for Recurrent Events

In the models discussed previously, it is assumed that frailty is constant over time for a particular subject, but in some situations it might be desirable to allow time-varying random effects. In hazard terms, for instance, the risk of infection may increase once a first failure event occurs. We might, therefore, prefer to consider models where the dependence between survival times for a subject is taken into account using time-dependent covariates. Time-varying frailties have been proposed for recurrent events using dynamic models.

Software for Recurrent Events

A number of packages or functions for the analysis of survival data also have the capability of dealing with recurrent events, among them the functions SURVREG and COXPH in *S*-plus (Insightful Corp.) and the corresponding *R* versions (*R* Project for Statistical Computations). FRAILTYPACK under *R* can also be used to estimate the parameters in gamma frailty models with possibly right-censored, left-truncated, and stratified survival data; the procedures LIFEREG and PHREG in SAS (SAS Institute). Other major statistical packages such as Stata (Stata Corp.) also provide some procedures that will deal with recurrent events.

Virginie Rondeau

See also Cox Proportional Hazards Regression; Survival Analysis

Further Readings

Cook, R. J., & Lawless, J. F. (2007). *The statistical analysis of recurrent events*. New York: Springer.

Cox, D. R. (1972). Regression models and life tables (with discussion). *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 34,* 187–220.

Duchateau, L., & Janssen, P. (2008). *The frailty model*. New York: Springer.
Duchateau, L., Janssen, P., Kessic, I., & Fortpied, C. (2003). Evolution of recurrent asthma event rate over time in frailty models. *Journal of the Royal Statistical Society: Series C (Applied Statistics), 52,* 355–363.
Fong, D. Y., Lam, K. F., Lawless, J. F., & Lee, Y. W. (2001). Dynamic random effects models for times between repeated events. *Lifetime Data Analysis, 7,* 345–362.
Kelly, J. P. (2004). A review of software packages for analyzing correlated survival data. *The American Statistician, 58,* 337–342.
Liu, L., Wolfe, R. A., & Huang, X. (2004). Shared frailty models for recurrent events and a terminal event. *Biometrics, 60,* 747–756.
Rondeau, V., & Gonzalez, J. R. (2005). Frailtypack: A computer program for the analysis of correlated failure time data using penalized likelihood estimation. *Computer Methods and Programs in Biomedicine, 80,* 154–164.
Rondeau, V., Mathoulin-Pelissier, S., Jacqmin-Gadda, H., Brouste, V., & Soubeyran, P. (2007). Joint frailty models for recurring events and death using maximum penalized likelihood estimation: Application on cancer events. *Biostatistics, 8,* 708–721.
The R Project for Statistical Computing: http://www.r-project.org
Therneau, T. M., & Grambsch, P. M. (2000). *Modelling survival data*. New York: Springer.
Yau, K. K., & McGilchrist, C. A. (1998). Ml and Reml estimation in survival analysis with time dependent correlated frailty. *Statistics in Medicine, 17,* 1201–1213.

Recursive Partitioning

Recursive partitioning (RP), or Classification and Regression Tree (CART) analysis, represents an exploratory statistical method for the multivariable data-derived and computer-mediated analysis of uncovering a structure in a data set. It produces a graphical output in the form of a decision tree, which facilitates data interpretation. This graphical output may be used in medical decision making to stratify patients into risk categories. Unlike other individual modeling techniques such as nomograms or artificial neural networks, CART predicts a risk group or stratum on an individual basis.

Tree-structured methodologies were first put forward by Morgan and Sonquist in 1963 and Morgan and Messenger in 1973. The CART program for classification and regression trees implements an expanded and strengthened tree-structured approach due to Breiman and colleagues in 1984. Basic tree-structured classification is discussed first. Special issues in estimating the performance of tree classifiers and the implementation of data resampling procedures (such as test-set, cross-validation) are reviewed. CART detects an algorithm in the available data set. It implements a recursive partitioning procedure based on an iterative search for the best binary splits of data. Consequently, classifiers consist of binary trees whose leaves determine class labeling. The goal is to establish a statistically reliable separation of classes. The first split defines two subgroups that maximize overall class separation. Subsequently, each subgroup serves as the basis for further partitioning, independently of the others, and so on. At each step, class separation is maximized. The sequence of partitions is summarized by a binary tree (see Figure 1).

The root node of the tree corresponds to the entire data set. Partitions of the data set are associated with descendants of the root node. The leaves of the tree correspond to subgroups that are not further partitioned. Their related class label is that of the majority class in that specific node. The basis for each partition is represented by a split of a node into a left and a right branch. Splits consist of evaluating the algorithmic condition, for each case in that particular node. If the condition is true, the case "goes" to the left, or else it "goes" to the right branch. The particular algorithmic condition chosen maximally separates classes in each node. Each terminal node contains members of one class only. To classify a new case, "drop" it through the tree. Starting from the root node, the case follows a "path" determined by the splitting rules and ends at a terminal node. The label of that node is assigned to that case. CART implements such a tree-structured classification procedure by quantifying maximal class separation and selection of splits.

Tree Construction

Maximal class separation requires formation of descendant nodes that are more "pure" in class representation than their immediate ancestors.

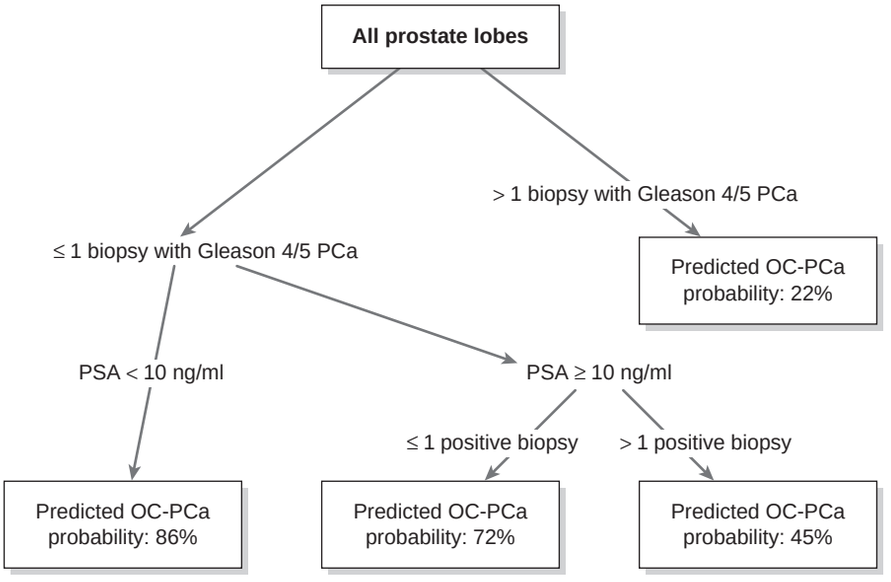


Figure 1 Example of a CART model to predict organ-confined prostate cancer based on detailed side-specific biopsy and prostate-specific antigen (PSA) information

Notes: PCa, prostate cancer; PSA, prostate-specific antigen (ng/ml); OC, organ confined.

Node purity is defined on the basis of a node impurity function, from which several definitions follow. If a node has an equal number of cases from each class, the node impurity function probability is maximal, and thus, the node is maximally impure. If the node contains cases of one class only, the node is maximally pure. When a node is split, a proportion of cases is sent to the right and to the left. The resulting change in impurity is the difference between the initial node impurity probability and a weighted sum of the impurities of the left and right branch. Terminal nodes have the greatest purity. Therefore, summing node impurity over the set of terminal nodes, weighing by the proportion of total cases in each, gives the overall tree impurity. In other words, maximal class separation at a node implies the maximization of node purity probability and minimization of node impurity probability. Maximal class separation is equivalent to reducing misclassification of cases. The Gini Diversity Index impurity function can be used to estimate the probability of misclassification and the conditional probabilities for a CART model and is readily calculated by a counting algorithm. Furthermore, the Gini Index has an advantage in that it favors the production of pure over impure descendant nodes. When all possible candidate splits have been generated for one variable, the procedure is repeated for another variable, and so forth. From this maximum set of possible single-variable splits, the split with the largest purity is applied to generate a new partition. CART repeats the above process recursively for each descendant node. The extension of this procedure to multiple dimensions is straightforward.

In summary, tree-structured classification rests on the computer implementation of a recursive partitioning of multivariate data spaces. Coupled with the quantification of node purity, the iterative search for best splits generates a tree whose overall purity is maximized.

Accuracy, Validation, and Tree Selection

Accuracy indicates the overall ability of the model to predict the outcome of interest. Current statistical methods offer the possibility of assessing a model's predictive accuracy. Usually, it is derived from the receiver operating characteristic (ROC) area under the curve (AUC). However, as opposed to the ROC, which is discriminatory, predictive accuracy combines both: discrimination and

calibration. In models that rely on time-to-event analyses and are subject to data censoring, the AUC method can be replaced with Harrell's concordance index. For both methods, predictive accuracy ranges from 50% to 100%, where 50% is equivalent to a flip of a coin and 100% represents perfect prediction. No model is perfect; the most commonly reported predictive accuracy values range from 70% to 85%.

Accuracy represents one of the most important criteria for a statistical tool that may be used in medical decision making. Predictive accuracy—that is, discrimination and calibration—should be ideally confirmed in an external cohort, which represents the gold standard method for quantifying a model's accuracy. In the absence of an external cohort, statistical methods such as data resampling techniques (e.g., bootstrapping, split sample, cross-validation, or leave-one-out) may be used to improve the estimation of accuracy. These methods rely on the same sample that was used for the model development and are termed *internal validation*. The use of the same sample to develop and validate a model may potentially be associated with an inflated accuracy. Therefore, external validation is preferred except for excessively rare pathologies where sample sizes are critically small.

Split sample and cross-validation represent robust internal validation methods. In brief, the data set is divided into learning and test sample. Trees are constructed using the learning sample, and the error rates and accuracy of classification of the test sample data are assigned to that tree. These kinds of measures form the basis for unbiased tree selection rules. Potentially, split sample estimates can be unbiased, and cross-validation accuracy estimates may tend to overestimate. Together, these accuracy measures are used to define an expected correct percentage classification score for a set of data.

Taken together, CART offers the possibility to validate its accuracy either internally by resampling or by external validation. However, external validation represents the gold standard.

Examples

In the field of prostate cancer outcome prediction, the CART methodology has been used in several scenarios covering early detection, staging, and prognosis. For example, a well-established preoperative CART model for prediction of extracapsular extension at radical prostatectomy uses detailed side-specific preoperative biopsy and serological prostate-specific antigen (PSA) information to label four different risk groups (Figure 1). On the dorsolateral side of the prostate travel autonomic nerve bundles that convey erectile function. In the case of a cancer extending over the capsule and penetrating the nerve bundle, excision is mandatory to achieve best oncological results postoperatively. However, in those individuals in whom the disease is confined to the prostate, a nerve-sparing surgical technique may maintain erectile function. The root node is split by the extent of aggressive cancer at biopsy to define the high-risk group with a probability of only 22% for organ-confined disease, followed by a PSA split of 10 ng/ml. Those with a PSA value <10 ng/ml represent those men with a high probability of organ-confined prostate cancer or low risk of extracapsular disease. Finally, the tumor extent within all biopsy cores serves to separate those individuals with a 45% versus 72% risk of organ-confined disease. This CART model serves as a decision tool to select men for a nonnerve sparing, unilateral, or bilateral nerve-sparing surgical technique. This example clearly demonstrates the useful adoption of a statistical methodology into clinical practice.

Felix K.-H. Chun, Markus Graefen, Alexander Haese, and Pierre I. Karakiewicz

See also Decision Tree: Introduction; Decision Trees, Advanced Techniques in Constructing

Further Readings

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). *Classification and regression trees*. Belmont, CA: Wadsworth.

Chun, F. K., Karakiewicz, P. I., Briganti, A., Walz, J., Kattan, M. W., Huland, H., et al. (2007). Critical appraisal of logistic regression-based nomograms, artificial neural networks, classification and regression-tree models, look-up tables and risk-group stratification models for prostate cancer. *BJU International, 99*(4), 794–800.

Efron, B., & Tibshirani, R. (1997). Improvements on cross-validation: The .632 + bootstrap method. *Journal of the American Statistical Association, 92,* 548–560.

Graefen, M., Haese, A., Pichlmeier, U., Hammerer, P. G., Noldus, J., Butz, K., et al. (2001). A validated strategy for side specific prediction of organ confined prostate cancer: A tool to select for nerve sparing radical prostatectomy. *Journal of Urology, 165*(3), 857–863.

Morgan, J. N., & Messenger, R. C. (1973). *THAID: A sequential search program for the analysis of nominal scale dependent variables.* Ann Arbor: Institute for Social Research, University of Michigan.

Morgan, J. N., & Sonquist, J. A. (1963). Problems in the analysis of survey data, and a proposal. *Journal of the American Statistical Association, 58,* 415–434.

Steuber, T., Graefen, M., Haese, A., Erbersdobler, A., Chun, F. K., Schlomm, T., et al. (2006). Validation of a nomogram for prediction of side specific extracapsular extension at radical prostatectomy. *Journal of Urology, 175*(3), 939–944.

Reference Case

A reference case is a set of methodological practices intended to enable, by means of standardization, meaningful comparisons of economic evaluation results both within and across different diseases and interventions. Such comparisons are unavoidable if economic analyses are expected to inform healthcare resource allocation decisions.

A reference case may be interpreted as a specific, highly prescriptive variant of a methodological guideline for health economic evaluations. Methodological guidelines have been developed as tools to support the conduct of scientifically consistent economic studies. Informal guidelines developed by academic groups often are differentiated from formalized guidelines issued by official bodies charged with technology appraisals to inform reimbursement and pricing decisions.

Background

In the absence of a standard, analysts were free to make choices, including (but not limited to) the form of evaluation method (e.g., cost benefit vs. cost effectiveness analysis), the appropriate measure of benefit (e.g., willingness to pay vs. health outcomes), the perspectives for valuation (i.e., the source of preference data, e.g., patients vs. a representative sample of the general public, individual vs. social, *ex ante* vs. *ex post*, or the choice of scaling instrument for utility measurement, such as standard-gamble, time-trade-off, person-trade-off, etc.) and costing (e.g., from a payer's or from a societal viewpoint), the discounting of future benefits and costs, and the reporting of their findings. The resulting variation of analytic approaches would greatly decrease the policy value of economic analyses. In response, the concept of reference case analysis was proposed to serve as a point of comparison by a common core of methodological choices across studies. It is widely acknowledged that reference case analysis, although prescriptive and generic (i.e., not disease specific) by definition, should not prevent analysts from pursuing—in addition—alternative evaluation approaches if and when they have reason to believe that the alternatives would yield more valid results or might better reflect the needs of the target audience of an analysis.

Washington Panel

A group of experts known as the Washington Panel was convened by the U.S. Public Health Service with the main task of developing standards for cost-effectiveness analysis (CEA), to ensure that differences in reported health outcomes, costs, and cost-effectiveness ratios, across studies and interventions, reflect true differences in the consequences, as opposed to artifacts due to unnecessary differences in method. Within the field of health economics, the Washington Panel introduced the notion of a reference case in 1996. The panel endorsed the use of CEA as an aid to, not a complete procedure for, decision making, on the grounds of its broader acceptance among healthcare policy makers compared with cost-benefit analysis (CBA), in light of sensibilities that a willingness to pay measure may inherently favor the wealthy over the poor. The panel recommended adopting a broad societal perspective, considering all changes in resource use and health effects due to an intervention, using a time horizon long enough to capture all the relevant future effects, applying a discount rate of 3% for both costs and effects and expressing health-related outcomes as quality-adjusted life years (QALYs). The panel proposed reporting incremental cost-effectiveness ratios (ICERs) but did not suggest an ICER threshold separating cost-effective technology from

others. The convention to exclude "indirect" productivity loss from cost calculation for reference case analysis, introduced by the Washington Panel for concerns about double counting (assuming that the full impact of morbidity was captured in the QALY measure and hence part of the denominator of the ICER), became a subject of controversial debate among health economists.

National Institute for Health and Clinical Excellence

The National Institute for Health and Clinical Excellence (NICE) was established as a special health authority within the U.K. National Health Service (NHS) in 1999 and quickly attained high international visibility. NICE evaluates 20 to 30 (mainly new and mainly pharmaceutical) technologies each year and provides mandatory guidance on their use to the NHS in England and Wales on grounds of their clinical and cost-effectiveness. To improve consistency within and between technology appraisals, NICE adopted a generic reference case with its revised methods guide in 2004. NICE justified the focus on CEA using the QALY, assumed to represent a universal and comprehensive measure of health outcomes, by its widespread use. Costing should be done from the perspective of the NHS and should include personal social services (PSS); future costs and benefits should be discounted using an annual rate of 3.5%. Since the 2004 methods guidance, parameter uncertainty should be evaluated using probabilistic sensitivity analysis. NICE indicated a most plausible range of ICERs, between £20,000 and £30,000 per QALY gained, as a benchmark for judgments about the cost-effectiveness of an intervention while recognizing that other factors such as the degree of clinical need of patients may influence its appraisals. According to NICE, estimates of the NHS (and PSS, where appropriate) budgetary impact ("affordability") of adopting a technology are not used for decision making but for implementation planning only. NICE allows additional (nonreference case) analyses if and when these can be justified.

Table 1 gives an overview of the reference case definitions.

Context and Critique

The concept of a reference case has not been universally adopted among international decision-making bodies and health technology assessment (HTA) agencies, using economic evaluations. For instance, the revised Australian guidelines, issued by the Pharmaceutical Benefits Advisory Committee (PBAC) in November 2006, expressed a general preference for cost-utility analysis (i.e., CEA using health-adjusted life years—most often QALYs—as a measure of health-related outcomes) but explicitly supported the use of CEA (with health outcomes measured in natural units, such as mmHg blood pressure reduction, episode-free days, clinical events avoided, or [unadjusted] life years gained; however, the choice of outcome measure should be justified) and cost-consequence analysis, when disaggregation of outcomes would be helpful. PBAC is also prepared to accept supplementary CBA, where outcomes are measured in monetary terms. The PBAC guidelines thus provide for an important example, where greater flexibility of analytic approaches is endorsed.

This notwithstanding, current international methodological guidelines for health economic evaluations broadly agree on many salient aspects such as the type of analysis (CEA), a strong reliance on clinical-effectiveness data and the principles of evidence-based medicine (Cochrane-style systematic reviews), choice of comparators, incremental comparisons reporting ICERs, the need to address decision uncertainty by way of sensitivity analysis, the need for and acceptance of decision analytic modeling, and adequacy of time horizon. There is less agreement among guidelines on the appropriate perspective of analysis (with a payers' perspective more often recommended in formalized official guidelines, as opposed to a societal perspective in informal academic guidelines), the relevance of Phase III efficacy trials, and the role of modeling. Ongoing academic debate concerns the valuation of health outcomes (e.g., natural units vs. QALYs vs. willingness to pay), the best way to account for uncertainty (e.g., regarding the use of probabilistic sensitivity analyses), and the role of budget impact analysis.

A major impetus behind the advocacy of a reference case approach, by the Washington Panel and by NICE, has been the basic ability to rank technologies across different disorders by their incremental cost per QALY and therefore the assumption that such rankings ("league tables") are conceptually valid. The implicit normative premises, in particular the

Table 1 Overview of reference case definitions

Issue	*Washington Panel Reference Case*	*NICE Reference Case*	*Methodological Guidelines*
Problem definition	The panel's framing recommendations are kept separate from its reference case definition	Scope from NICE	Usually expected to define indication, patient (sub) groups, comparator, and perspective
Comparator(s)	Existing practice; if not cost-effective, consider a (a) best available, (b) viable low-cost, or (c) "do-nothing" alternative	Alternative therapies routinely used within the NHS will be defined in the scope developed by NICE and will require definition and justification	Usually common practice ("f"), however, somewhat vague ("existing practice," "common practice")
Evidence on outcomes	Data should be selected from the best designed (and least biased) sources that are relevant to the question and population under study	Systematic review, with a preference for quantitative meta-analysis of randomized clinical trials data	Usually (long-term) effectiveness, not efficacy; with a broadly prevailing preference for data from randomized clinical trials
Economic evaluation	CEA	CEA	Usually CEA; sometimes more flexible (including cost-minimization and CBA)
Perspective on outcomes	All health effects, encompassing the range of groups of people affected, over a time horizon long enough to capture all relevant future effects	All direct health effects on individuals, whether patients or others (principally caregivers); time horizon should be sufficiently long to reflect any differences between the technologies being compared	Usually all relevant health outcomes
Perspective on costs	Societal perspective, long-term using opportunity cost; excluding indirect (productivity) costs; perspective should be explicitly identified	NHS and PSS	Heterogeneous; direct health care costs only or direct and indirect (productivity) costs ("f"); societal perspective is requested more often in informal guidelines ("i")
Discount rate	A real, riskless discount rate of 3.0% should be used, complemented by a sensitivity analysis (drawn from 0% to 7%, including 5%)	An annual rate of 3.5% p.a. on both costs and health effects	Often 5% discount rate ("f"); heterogeneous recommendations from 2.5% to 10% in informal guidelines ("i")
Addressing uncertainty	Univariate sensitivity analysis as a minimum; multivariate sensitivity analyses recommended	Probabilistic sensitivity analysis mandatory (or, where appropriate, stochastic analysis of patient-level data)	Sensitivity analysis

Issue	*Washington Panel Reference Case*	*NICE Reference Case*	*Methodological Guidelines*
Measure of health benefits	QALYs	QALYs	Usually including QALYs, with more flexibility as to other measures ("f," "i"), especially physical units; sometimes willingness to pay
Source of preference data for calculation of utility weights	Community preferences; if unavailable, patient preferences may be used as an approximation	Representative sample of the public (UK)	If QALYs are used, usually community preferences
Health state valuation method	Quality weights must be preference based and interval scaled	Choice-based method (e.g., time trade-off or standard gamble, not rating scale)	If QALYs are used, usually choice-based methods; often standard gamble and time trade-off; sometimes rating scales
Description of health states for calculating QALYs	A generic classification scheme, or one that is capable of being compared to a generic system	Using a standardized and validated generic instrument	Heterogeneous; sometimes disease-specific instruments allowed ("f")
Equity position	Discussion of roles and limitations of CEA in the introductory Chapter (separate from reference case definition)	Each additional QALY has equal value	NA
Budget impact analysis	NA	Impact on NHS not part of the decision-making process; however, it is required to allow effective national and local financial planning	Usually NA; Ontario: products with high budget impact will need more rigorous documentation of cost-effectiveness

Notes: For comparison, methodological guidelines may be informal ("i"; usually academic) or formalized ("f"; issued by official bodies such as HTA or pricing and reimbursement agencies).

p.a. = per annum, NICE = National Institute for Health and Clinical Excellence, CEA = cost-effectiveness analysis, CBA = cost-benefit analysis, NHS = National Health Service, PSS = personal social services, QALYs = quality-adjusted life years.

value judgment of a primary health service objective to maximize the distribution-independent sum of QALYs produced (given a budget constraint), are not universally shared and have been described as empirically flawed—that is, not reflecting prevailing public preferences. Accordingly, both the Washington Panel and NICE have acknowledged the need to consider other factors beyond those specified for reference case analysis. Some observers have noted that, in practice, adherence to a generic standard may contribute to a neglect of disease-specific information and thus contradict the aim to use the best available clinical evidence in the context of HTAs. Also concerns have been raised that high levels of standardization might foster analyses "by the cookbook" and thwart further methods development. However, the usefulness of the reference case approach is perhaps best demonstrated by the fact that the absence of a methodological standard, and therefore inconsistency of methods applied, has been cited as a reason

why CBA (using contingent valuation to establish the willingness to pay for healthcare interventions) has not yet had much policy impact—despite its theoretical advantages and a growing number of published CBAs.

Michael Schlander

See also Contingent Valuation; Cost-Benefit Analysis; Cost-Effectiveness Analysis; Cost Measurement Methods; Cost-Utility Analysis; Discounting; Pharmacoeconomics; Quality-Adjusted Life Years (QALYs); Technology Assessments; Willingness to Pay

Further Readings

Australian Government, Department of Health and Ageing. (2006). *Guidelines for preparing submissions to the Pharmaceutical Benefits Advisory Committee*. Canberra, Australia: Pharmaceutical Benefits Advisory Committee.

Dolan, P., Shaw, R., Tsuchiya, A., & Williams, A. (2005). QALY maximisation and people's preferences: A methodological review of the literature. *Health Economics, 14*(2), 197–208.

Gold, M. R., Siegel, J. E., Russell, L. B., & Weinstein, M. C. [Washington Panel]. (Eds.). (1996). *Cost-effectiveness in health and medicine* (especially Appendix A, pp. 304–311). New York: Oxford University Press.

Hjelmgren, J., Bergren, F., & Andersson, F. (2001). Health economic guidelines: Similarities, differences and some implications. *Value in Health, 4*(3), 225–250.

National Institute for Clinical Excellence. (2004). *Guide to the methods of technology appraisal*. London: Author.

Richardson, J. (1994). Cost utility analysis: What should be measured? *Social Science & Medicine, 39*(1), 7–21.

Sach, T. H., Smith, R. D., & Whynes, D. K. (2007). A "league table" of contingent valuation results for pharmaceutical interventions: A hard pill to swallow? *Pharmacoeconomics, 25*(2), 107–127.

Schlander, M. (2007). *Health technology assessments by the National Institute for Health and Clinical Excellence: A qualitative study*. New York: Springer.

Regression to the Mean

Regression to the mean can be defined as the changes that take place over time among exceptional individuals and groups, and across generations they tend to be toward the average for the larger population from which they are drawn. Many of the changes observed emerge from the fact that the measurements involved are composed of two parts: (1) a valid part and (2) a random error part. In much of the medical and healthcare research, extreme values on the first test or measurement are likely due in part to the random error component. By chance alone, that random error component is likely to contribute less on the next measurement.

Historical Background

Sir Francis Galton was the first to document regression to the mean. His grandfather Erasmus Darwin, one of the leading intellectuals of that time, and cousin Charles Darwin were both geniuses. Galton wondered if geniality was hereditary and studied famous families of geniuses such as the Darwins and the Mozarts. He noted that the children of geniuses were almost all less brilliant than their parents and that the grandchildren were even less brilliant. Children and grandchildren of geniuses, on the average, are clearly gifted but invariably closer to the general population average than their (grand)parents. More numerical evidence of this effect was provided in his studies on comparing the heights of adult children and their parents. He noted that whenever parents are well above or below average in height, their children are also likely to be similarly above or below average in height, but not by as much. He observed that the same phenomenon was true for mother sweet peas and daughter sweet peas and published a paper with the title "Regression Toward Mediocrity in Hereditary Stature." Karl Pearson, Galton's biographer and a brilliant statistician, was the first to note that Galton had created "a revolution in the scientific ideas," not the least because this phenomenon creates the false impression that all phenomena after a sufficient number of regressions will be reduced to a boring, mediocre average, with no room for individual brilliance.

Description of the Phenomenon

Regression to the mean predicts that more extreme (deviant) measures tend to be closer to the

population's average when a follow-up measurement is taken. The less reliable the measurement is, the stronger the regression. For perfectly reliable measures, there is no regression to the mean. For instance, if you measure the length of a very large stone, that stone will, in all likelihood, not have "shrunk" the next morning. For totally unreliable measures, the regression is complete; the best prediction of the next measurement is the average. If you measure blood pressure by asking, "Did you enjoy last night's television show?" an extremely high blood pressure measurement after watching the worst show of the year will almost certainly be followed by a much lower blood pressure measurement the next day. The more extreme the measurement result, the more likely noise or error terms have played a role in the measurement. So the more extreme a measurement, the more likely and stronger the regression effect will be. The chance of regression toward the mean also increases when the two measurement instruments are less than perfectly correlated and when the groups are selected on a nonrandom basis (i.e., extreme groups). It is important to note that regression is not an artifact or an observable process. It is the direct consequence of the unreliability of two (pre- and post-) measurements. It is a measurement problem.

Examples of everyday expressions that refer to regression toward the mean are "Things will even out," "It can't possibly get worse (or better) than this," and "What goes up, must go down." Most people realize that the sequel to a blockbuster will not be as successful as the blockbuster, that it is easier to win once (with luck on your side) than to repeat that twice, that an extremely high score on a test will be followed by a high but lower score on a similar test, and so on. Also remarkable besides the ubiquitousness of regression toward the mean is how commonly it is misunderstood, usually entailing undesirable consequences.

The tendency to overlook regression can lead to critical errors in judgment. Under specific conditions, medical decision makers conclude that significant differences are due to treatment, when in fact they are due to regression to the mean.

For example, subjects with extreme values in blood pressure may be selected and treated to bring their values closer to the mean. If their values are measured again, it is found that the mean of the extreme group has moved closer to the mean of the whole population. This observation is considered proof as a treatment effect. However, it is very probable that even for subjects without treatment blood pressure will drop, just because of the regression to the mean effect.

A nonmedical example where the effect of blindness for regression to the mean plays a critical role is a case in which instructors judged the effects of praising or punishing army pilots on future flight performances. In a famous study by Nobel Prize winner Daniel Kahneman, one observed that praising pilots for well-executed maneuvers causes a decline in subsequent performance and that punishing them for poorly executed maneuvers caused a gain in performance. The wrong conclusion by the drill sergeants was that praise makes pilots lazy and that punishment keeps them motivated. Every time a researcher works with extreme groups, regression to the mean can appear. The same new pedagogical tool will cause the "best" students to drop and the "worst" students to improve. It is often known as the ritual rain dance effect. There is strong evidence that the drought period is much longer before the rain dance than after. So many tribes come to the straightforward conclusion that rain dances are really effective. This phenomenon is of course well known in healthcare. Some so-called alternative treatments can report positive results because patients only visit those "doctors" when they are desperate. In line with the regression phenomenon, those treatments seldom work when the measurements are very reliable and the disease is well understood but seem to be very "successful" when the complaints are vague, chronic, and the measurements are notoriously unreliable.

Examples in Medical and Healthcare Management Decisions

Regression to the mean is one of the most common fallacies that occur whenever extreme high or extreme low groups are selected from a population based on the measurement and postmeasurement of a particular variable. Regression toward the mean occurs whenever subjects are selected on the basis of an extreme cutoff value for a certain characteristic and then undergo a postmeasurement on that same characteristic. Medical science literature is replete

with studies that report treatment effects but in fact are instances of this regression effect. For example, a study into the therapeutic effects of borage oil on people with eczema noted that people with a high atopic dermatitis score (i.e., symptoms score) showed a significant drop in the average symptoms score 2, 4, 8, and 12 weeks after the initial measurement. Moreover, the most fascinating observation in this randomized, double-blind, placebo-controlled experiment is that the average score decreased in the postmeasurements for participants from the treatment and the control group.

A common practice in large pharmaceutical and clinical labs is to try out a new treatment or drugs on clinical outliers, often yielding very positive treatment results, whereas in many cases it is regression to the mean that is the major reason for this change. A recent inquiry demonstrated how regression is embedded in research on substance use disorder treatment. In that particular study, it was noted that a group of drug addicts who were assigned to a special treatment reported a decrease in the level of substance use disorder (SUD) similar to the SUD decrease for the comparison group that received a placebo treatment. Another example where regression operates is a study on artificial insemination. It was found that subjects selected with low sperm concentration and low motility index in the first test had significantly higher means for both characteristics on the second test. In other words, even semen quality improved without any therapeutic intervention.

Regression to the mean should also be understood by policy makers. This regression fallacy continues to be missed in decisions by public health policy makers. For example, part of the 90% decrease of meningitis C in the United Kingdom was attributed to the introduction of the immunization program, where it could be explained by a very bad year being likely followed by better years. What frequently happens is that a sudden increase in a particular disease leads to changes in policies such as large-scale immunization programs, with a large decrease in disease incidence attributed to the intervention. However, policy makers should also be vigilant toward the fact that the unexpected increase in the incidence of a particular disease could be simply the result of chance. In consequence, doing nothing could have resulted in similar effects.

Dealing With Regression to the Mean

A first step for dealing with the effects of regression to the mean starts with recognizing and understanding this statistical phenomenon. In uncontrolled studies with repeated measurement designs, regression to the mean can be a tenacious problem. Changes in measurement results are often interpreted as real treatment effects. Several statistical approaches have been suggested to detect regression to the mean in uncontrolled studies. One recommended statistical approach is analysis of covariance (ANCOVA). In using ANCOVA, one can adjust each subject's follow-up measurement according to his or her baseline measurement. In doing so, a treatment effect can be predicted after controlling for the regression to the mean effect.

Although statistical control is one way to deal with the regression to the mean, a good study design can mitigate its effects. One way is the use of randomized control group designs. By random allocation of subjects to treatment and placebo conditions, the responses from all groups should be equally affected by regression to the mean, which makes it possible to estimate regression and treatment effects. An alternative approach to alleviate the effect is by improving the measurement reliability of the baseline measurement, for example, by using multiple measurements. Using only one measurement value for selecting subjects is likely to be subject to high variability. Therefore, selection of subjects on the basis of the mean of multiple baseline measurements is considered a better option.

Dave Bouckenooghe and Marc Buelens

See also Distributions: Overview; Frequency Estimation

Further Readings

Baker, H. W. G., & Kovacs, G. T. (1985). Spontaneous improvement in semen quality: Regression towards the mean. *International Journal of Andrology, 7,* 383–388.

Barnett, A. G., van der Pols, J. C., & Dobson, A. J. (2005). Regression to the mean: What is and how to deal with it. *International Journal of Epidemiology, 34,* 215–220.

Bland, J. M., & Altman, D. G. (1994). Regression towards the mean. *British Medical Journal, 308,* 1499.

Bland, J. M., & Altman, D. G. (1994). Some examples of regression towards the mean. *British Medical Journal, 309,* 780.

Breaugh, J. A., & Arnold, J. (2007). Controlling nuisance variables by using a matched-groups design. *Organizational Research Methods, 10,* 523–541.

Campbell, D. T., & Kenny, D. A. (1999). *A primer on regression artifacts*. New York: Guilford Press.

Morton, V., & Torgerson, D. J. (2003). Effect of regression to the mean on decision-making in health care. *British Medical Journal, 326,* 1083–1084.

Morton, V., & Torgerson, D. J. (2005). Regression to the mean: Treatment effect without the intervention. *Journal of Evaluation in Clinical Practice, 11*(1), 59–65.

Yudkin, P. L., & Stratton, I. M. (1996). How to deal with regression to the mean in intervention studies. *Lancet, 354,* 241–243.

Regret

Regret is the negative emotion that we experience when realizing or imagining that our present situation would have been better had we decided or acted differently. Regret originates in a comparison between outcomes of a chosen course of action and the nonchosen alternatives in which the latter outperform the former. It is clearly a painful emotion that reflects on one's own causal role in the current, suboptimal situation. The emotion regret is accompanied by feelings that one should have known better and having a sinking feeling, by thoughts about the mistake one has made and the opportunities lost, by tendencies to kick oneself and to correct one's mistake, by desires to undo the event and get a second chance, and by actually doing this if given the opportunity. Put differently, regret is experienced as an aversive state that focuses attention on one's own causal role in the occurrence of a negative outcome. It is thus a cognitively based emotion that motivates one to think about how the negative event came about and how one could change it or how one could prevent its future occurrence.

As such, regret is unique in its relation to decision making and hence to feelings of responsibility for the negative outcome. This makes regret an emotion that is highly relevant for medical decision making. One only experiences regret over a bad outcome when, at some point in time, one could have prevented the outcome from happening. Of course, other emotions can also be the result of decisions; for example, one may be disappointed with a decision outcome or happy about the process by which one made a choice. But, all these other emotions can also be experienced in situations where no decisions are made, whereas regret is exclusively tied to decisions. For example, one can be disappointed in being the carrier of a genetic disease and happy with the fact that the medication works, but one cannot regret these instances. Thus, in regret, personal agency and responsibility are central, whereas in other aversive emotions, such as anger, fear, and disappointment, agency for the negative outcomes is either undetermined or in the environment or in another agent. Hence, regret is the prototypical decision-related emotion in the sense that it is felt in response to a decision and that it can influence decision making.

Tymstra may have been the first to introduce the concept of anticipated regret into the domain of medical decision making. He described the imperative character of medical technology and argued that the mere existence of medical-technical possibilities makes it hard for doctors and patients to reject them. The success of, among others, prenatal testing and IVF (in vitro fertilization) is argued to be testimony to the effects of anticipated regret. The reasoning is that as soon as these new technologies come into play, patients (and doctors) imagine not trying them and thereby forgo a potential improvement or remedy. The accompanying sense of regret urges both doctors and patients to use this new technology. The other side of the imperative character of regret is that even the ones who opted for using the new technologies but did not obtain the hoped-for outcomes generally do not regret this, because they felt that "at least they tried." This is consistent with recent empirical findings in the science of regret by Sorum and colleagues.

The impact of regret on decision making is, of course, the prime reason for decision researchers to become interested in regret. Regret may affect decision making in two ways. First, the experience of retrospective regret may produce a behavioral inclination to reverse one's decision or undo the consequences. Second, decision makers may anticipate

possible future regret when making decisions and choose in such a way that this future regret will be minimal.

The idea that people, when making decisions, might take into account future emotional reactions to possible decision outcomes has some history in research on decision making, starting with economists studying rational choice in the early 1980s. It is now known that the influence of anticipated future regret on current decision making can take several forms. First, people may refrain from deciding in order to avoid making the wrong decision. This may cause both doctors and patients to be willing to defer responsibility. However, this inactive attitude may result in regret as well since it is known that in the long run, inactions produce the most regret. For example, patients might end up regretting having followed the advice of their doctor if the procedure does not have the expected effect. People may also avoid or delay their decisions, because they want to gather more information in order to make a better decision. This could result in extensive usage of medical tests and risky new technologies, as pointed out above.

Despite the fact that decision makers often anticipate regrets that may follow their decisions, especially when the decision is important, it does not prevent them from experiencing regret on a regular basis. Connolly and Reb discuss how both anticipated regret and retrospective regret influence cancer-related decisions. They also review research showing that postdecision regret is not unusual in cancer patients. Since the experience of regret not only reflects on the decision that produced the regretted outcome but also negatively affects the well-being that stems from the current state of affairs, understanding the psychology of regret is vital for helping patients to cope with it. Also, knowledge of how regret affects well-being and behavioral decisions may help develop interventions that prompt people to behave healthier. Extensive research has documented the success of fear appeals on health behavior. Similar studies focusing on the effects of regret appeals are wanting.

Taken together, regret is an aversive emotional state that is related to counterfactual thoughts about how the present situation would have been better had one chosen or acted differently. Therefore, decision makers are motivated to avoid or minimize postdecisional regret. This has several implications for medical decisions, because people may employ different strategies to prevent regret from happening or to cope with regret when it is experienced. In principle, the effects of regret can be considered rational, because they protect the decision maker from the aversive consequences of the experience of regret. There might be cases, however, in which an aversion to regret leads one to avoid counterfactual feedback and hence results in reduced learning from experience. This might be considered irrational. But, irrespective of this rationality question, regret has shown to be a basic emotion in the behavioral decisions of both patients and doctors. As such, it is of vital importance to take the experience of regret seriously and be aware of how it may impact these decisions.

Marcel Zeelenberg

See also Bounded Rationality and Emotions; Emotion and Choice

Further Readings

Brehaut, J. C., O'Connor, A. M., Wood, T. J., Hack, T. F., Siminoff, L., Gordon, E., et al. (2003). Validation of a decision regret scale. *Medical Decision Making, 23,* 281–292.

Connolly, T., & Reb, J. (2005). Regret in cancer-related decisions. *Health Psychology, 24,* S29–S34.

Djulbegovic, B., Hozo, I., Schwartz, A., & McMasters, K. M. (1999). Acceptable regret in medical decision making. *Medical Hypotheses, 53,* 253–259.

Sorum, P. C., Mullet, E., Shim, J., Bonnin-Scaon, S., Chasseigne, G., & Cogneau, J. (2004). Avoidance of anticipated regret: The ordering of prostate-specific antigen tests. *Medical Decision Making, 24,* 149–159.

Tymstra, T. (1989). The imperative character of medical technology and the meaning of anticipated decision regret. *International Journal of Technology Assessment in Health Care, 5,* 207–213.

Zeelenberg, M., & Pieters, R. (2007). A theory of regret regulation 1.0. *Journal of Consumer Psychology, 17,* 3–18.

Religious Factors

Religious factors can be important in the decision making of patients, healthcare professionals, and

healthcare organizations. Paradigmatic examples of decisions involving religious factors are conscientious objection to abortion and patient resistance to postmortems, but religious elements may influence a much wider range of decisions in healthcare.

Religion, Culture, and Values

Defining what counts as a religious factor in decision making is not simple. There is no generally agreed-on definition of religion; for instance, not all religions have deities. Furthermore, in specific cases there is frequently no simple way of separating religious factors from cultural factors, because the same religion has developed different traditions in different geographical regions and in different societies.

In the case of the major scriptural religions, three or sometimes four elements come together to determine the beliefs and practices of adherents:

1. The original revelation in the form of the scriptures, that is, the writings held to be authoritative
2. The tradition
3. Reason
4. In some instances, ongoing revelation

The balance between the three sources of religious guidance differ between religions and often also within religions in relation to different areas of belief. The scriptural revelation is furthermore being continuously (re)interpreted through reason, but this always takes place through the lens of a particular tradition, and that tradition may also decide who is seen as competent to perform the interpretation. This means that it will often be important to know which particular branch of a given religion a patient identifies with because (a) the religious beliefs and values may differ between different branches and (b) the patient may be more responsive to advice from his or her "own" religious leader than from someone representing another branch of what to the outsider looks like the same faith community.

The fact that a particular view is religious is not in itself a reason to respect it or accord it more weight. Religious views are not by the mere fact that they are religious intrinsically worthy of respect. But religious identification is often part of a person's core identity, and there are good reasons to treat core beliefs differently than beliefs that the person in question sees as peripheral.

Religion and Patients

Religious factors often play a significant role in the decision making of patients, especially in cases where being a member of a particular faith community is part of the patient's core identity. This may lead a person to aim at particular treatment goals and choose or refuse particular treatment modalities. A Muslim couple seeking in vitro fertilization may, for instance, have religious objections to any use of donor gametes but no objection to the creation of embryos and selection among embryos created with their own gametes.

It is generally accepted in the law of most jurisdictions that an adult person can refuse any kind of treatment even if that treatment is likely to be lifesaving. Healthcare professionals have been slightly slower to accept this, but it is now widely accepted that, for instance, an adult member of Jehovah's Witnesses should be allowed to refuse even a potentially lifesaving blood transfusion.

There is, however, in most jurisdictions no general or specific obligation to provide patients with particular treatments that they desire if these treatments are not clinically indicated.

With regard to children, the legal position is that parents cannot deny clinically indicated treatment that is lifesaving or likely of great benefit to the child even if they have strong religious objections, and even if these objections are shared by the child.

In cases where parents request minor surgical or other procedures that are religiously required, for instance, male circumcision, there is no obligation to provide such procedures, but on the other hand no reason not to perform them if they are not harming the child.

It is generally accepted that religiously required procedures that cause significant harm should never be performed.

Religion and Healthcare Professionals

Religious factors enter the decision making of healthcare professionals in two different ways:

(1) specifically, in the case of the religious healthcare professional, and (2) more generally, for all healthcare professionals in relation to patients who hold religious views.

In the case of healthcare professionals who are religious, this may influence the kind of procedures they are willing to provide or participate in and the advice they give to their patients.

Legally recognized conscientious objection is the clearest example of this influence. In many jurisdictions, healthcare professionals can legally refuse to participate in induced abortions if they have a conscientious objection to this procedure. Such objections can be based on nonreligious values but will often be religion based.

It has recently been questioned whether there should be a right to conscientious objection to any kind of legal healthcare procedure, especially in cases where widespread objection leads to problems for patients in accessing the services they want and/or need. This issue is not yet resolved.

In the more general situation of interaction with a religious patient, the healthcare professional will often need to understand something about the patient's religion and culture in order to (a) provide appropriate care and advice and (b) not behave inappropriately toward the patient. Without such knowledge, the interaction with the patient may become unsuccessful; for example, one should not touch the head of a Maori patient without permission so as to not interfere with the patient's *mana* (power or strength), and a Maori patient may lose confidence in a healthcare professional who does not know or understand this.

Healthcare professionals cannot have an obligation to know everything of relevance about all religions, because that would be impossible, but they can be reasonably expected to know something about those religious groups that are prevalent in the locality where they work and reasonably expected to be sensitive toward religious views. As mentioned above, this information will often have to be fairly specific, for example, not just knowledge about healthcare-relevant beliefs in Judaism in general but knowledge about the beliefs of the Lubavitcher branch of Hasidic Judaism.

Understanding the patient's religious views becomes especially important in relation to decisions about incompetent patients. It is generally accepted that such decisions should be made in the best interest of the patient and that the scope of best interest is broader than just "medical best interest." In the leading U.K. case, *Re S (Adult Patient: Sterilisation)*, this wider conception of best interest was expressed in the following way:

> That, once satisfied that the proposed treatment options were within the range of acceptable opinion among competent and responsible practitioners, the court should move on to the wider and paramount consideration of which of them was in the patient's best interests.

This entails that a patient's deeply held religious views may influence or determine what is in that patient's best interest.

In cases where there is conflict between what the healthcare professional believes to be the clinically indicated course of action and the patient's religiously influenced decision, the conflict may sometimes be resolved by involving a religious leader whom the patient trusts. The patient may have misunderstood what the religion actually requires, or the degree to which the requirements can be suspended in cases of illness. Most branches of Islam will, for instance, suspend the Ramadan fasting requirements for people who are ill and allow the use of products containing materials from pigs as part of necessary medical treatment, but not all Muslims know this.

Religion and Healthcare Organizations

Religious factors are important for the decision making of healthcare organizations because (a) the organization has to act appropriately in relation to the religious views of its patients and staff and (b) the organization may itself be committed to a particular religious view and have a religious identity.

Any healthcare organization needs to make conscious decisions about how it is going to accommodate the religious views of patients and staff. This becomes more urgent and also more complicated for healthcare organizations that are situated in multicultural and multireligious environments. Making room for religious views and practices becomes especially important when there is a link

between care and treatment outcomes for patients and the degree to which specific aspects of their religion is catered to. Hospitals should, for instance, as a general policy be able to meet the dietary requirements of the population that they serve because of the link between nutrition and treatment outcome.

The degree to which an organization's religious identity can determine its delivery of healthcare may differ in different healthcare systems. In a hypothetical, completely free market where patients could choose freely between a large number of healthcare providers, there would probably be few restrictions to the degree to which an organization could legitimately let its religious identity influence its healthcare delivery. Patients who do not want to go to, for instance, a Catholic hospital could just choose the Muslim or the Atheist alternative. But most healthcare systems restrict patient choice to a considerable extent, and this also limits the extent to which a healthcare organization can let its religious identity influence its actions. If the patient has no choice, then that patient will in general have a strong, legitimate presumption of being able to access a complete package of treatment options, whatever institution he or she is admitted to.

Søren Holm

See also Advance Directives and End-of-Life Decision Making; Bioethics; Cultural Issues; Shared Decision Making

Further Readings

Buchanan, A. E., & Brock, D. W. (1989). *Deciding for others: The ethics of surrogate decision making*. Cambridge, UK: Cambridge University Press.

Guinn, D. E. (Ed.). (2006). *Handbook of bioethics and religion*. Oxford, UK: Oxford University Press.

Holm, S. (2003). Conscientious objection and civil disobedience in the context of assisted reproductive technologies. *Turkiye Klinikleri Journal of Medical Ethics, Law and History, 11,* 215–220.

Holm, S., & Edgar, A. (2008). Best interest: A philosophical critique. *Health Care Analysis,* DOI 10.1007/s10728-008-0092-x

Re S (Adult Patient: Sterilisation) [2001] Fam 15.

Savulescu, J. (2005). Conscientious objection in medicine. *British Medical Journal, 332*, 294–297.

Report Cards, Hospitals and Physicians

The term *report cards* in the context of healthcare has come to be used to describe comparative data on quality and cost of care, particularly when such data are made available to the public. A common view is that greater transparency will drive improvement and guide patients to the best care. Although not the only stakeholder group that is subject to scrutiny, this entry focuses on report cards on providers, specifically hospitals and physicians.

Few would argue with the importance of provider efforts to measure performance internally for the purposes of improvement. For example, a practice may perform a peer review of physicians based on adherence to recognized care standards and make such information available to all providers within a practice. The controversy grows, however, with increasing levels of transparency. Sharing that same information with patients, health plans, purchasers, and other stakeholders brings both opportunity and risk. While several years ago the controversy centered on whether report cards should be issued at all, in more recent years the debate has centered more on what measures are best and what are the most productive ways to use the information.

Measuring and sharing data on provider performance on a large scale is a relatively new development. There were a few pioneers in the field of healthcare quality improvement who saw the importance of measuring results long ago. For example, Florence Nightingale kept meticulous records of mortality rates in military hospitals during the 19th-century Crimean War, and in so doing, she blazed trails both in quality measurement and the field of nursing. In a later example, Ernest Codman famously established the End Result Hospital, in early 20th-century Boston, in which he maintained and shared detailed records of outcomes for his surgical patients. But despite the courageous efforts of a few, who at times were subjected to scorn and ridicule, the real effects of the work of healthcare providers have often been shrouded in secrecy. Not until the soaring costs of healthcare in the 1980s and beyond has serious,

widespread attention been given to measuring quality.

One of the primary reasons for the emergence of report cards is the growing demand for *value* in healthcare. Although defined in various ways, the basic concept of value has two components: (1) quality and (2) cost. High-quality care increases the likelihood of a desirable outcome and is in line with patient preferences, according to a widely cited Institute of Medicine definition. The broad concept of cost needs no definition, but actual measurement of cost poses many challenges. Value can be thought of as the ratio of quality to cost. While the feasibility of actually decreasing healthcare costs over the long term can be debated, many would agree that purchasers of healthcare should at least expect value for the healthcare dollar spent.

There is a great deal of evidence, however, that healthcare purchasers in the United States do not consistently get value for the healthcare dollar. Errors that bring harm to patients and drive costs up, missed opportunities to provide preventive care, and failure to bring the benefits of evidence-based treatments to patients, particularly for chronic disease, are clear indicators of problems.

Content of Report Cards

Report card measures typically fall into the following categories:

1. *Process* measures are most common and reflect what action was taken during patient care. They address the question "What did you do?" Documentation of aspirin given to patients with heart attack, a perioperative antibiotic given to a surgical patient, or a test such as an echocardiogram performed for a heart failure patient are examples of process measures.
2. *Outcome* measures address the question "What was the result?" Examples include survival after a surgical procedure, rates of hospital-acquired infection, or measures of patient experience.
3. *Structural* measures address "What do you have in place?" such as personnel, information technology, or equipment. Examples include computerized physician order entry and staffing of intensive care units with intensivists.
4. *Efficiency* measures reflect cost of care. Common examples are measures of hospital length of stay or cost for specific procedures, diagnoses, or, increasingly, episodes of care (e.g., a bout of pneumonia).

Who Is Interested in Provider Report Cards?

We can consider four broad groups of stakeholders: (1) consumers (patients and potential patients), (2) providers, (3) health plans, and (4) purchasers (government and employers). At the time of this writing, it is likely that only a small minority of consumers make important healthcare decisions based on provider report cards. One reason may be availability of information. Although there is a great deal of information online, it often takes a fair amount of time and effort, as well as a certain level of sophistication, to find it. Such an effort may require not only skills in navigating the Internet but also a certain amount of medical knowledge. In addition, other factors such as where one lives, limitations on choice by the payer, and recommendations from physicians, friends, and family may be more important to consumers than information from report cards. Some experts feel that as more consumers take on the responsibility of making decisions on how to spend healthcare dollars, the amount of consumer attention to report cards will greatly increase.

Provider attention to report cards varies, likely dependent on the local environment. Some regions of the United States have seen broad coalitions of health plans and other stakeholders for purposes of provider pay for performance. Such coalitions allow some standardization so that the provider ideally only has one report card rather than a report card for each individual health plan. In other regions, where a single report card link to a provider's pay is not present, provider attention to various report cards may be less.

Health plans and purchasers are driving a great deal of provider report card activity. Commercial health plans and coalitions of health plans have created report cards, as have employer groups such as those behind Leapfrog and Bridges to Excellence. And the federal and state governments are clearly engaged, an example being the Center for Medicare and Medicaid Services (CMS) and its Hospital Compare Web site. On this Web site, the user can

make comparisons of hospitals based on process of care, outcomes such as mortality, and patient experience. CMS is formulating plans for value-based purchasing in which reimbursement will be tied to hospital performance in these areas.

Data Sources

Administrative data, specifically claims for healthcare services, are important as a data source for report cards. Reports based on administrative data are often criticized on the grounds that the data were collected for billing purposes rather than for quality measurement. Such data can be incomplete or inaccurate. However, using claims data for report cards has the advantage of being highly feasible, and for this reason, the practice is growing despite the data limitations. The feasibility of using such data is derived from the fact that the data are widely available in a fairly timely manner, and such reports are relatively inexpensive to produce.

Some efforts have been made to augment claim-based quality measures. Clinical data abstracted from the medical record can be combined with claims data, for example, to achieve more robust risk adjustment for certain conditions. In addition, specific data can be added to the claim itself expressly for the purpose of quality measurement, such as in the Physician Quality Reporting Initiative, which is CMS "pay for reporting." It appears unlikely that the use of claims data for report cards will fade in the near future, simply because of the feasibility issue.

In addition to claims and medical records, there are other sources of data for report cards. State and federal government data sources may provide information such as mortality or hospital infection rates. Providers may be surveyed for information, most notably by the Leapfrog Group, which is mainly a purchaser coalition whose voluntary hospital survey is an effort to improve hospital quality and safety. Publicly available information on accreditation of hospitals and board certification of physicians are other elements that may be found in report cards.

Challenges to the Validity of Report Cards

There are several important ways in which the numbers included in reports cards might distort the truth regarding the underlying value of healthcare. The first is data availability. Many of the intricacies of a case that can help one understand the quality of care provided are buried in free text within health records. Such information may be illegible or go undocumented altogether. In addition, information on quality from various perspectives is not generally available. For example, a physician may be most interested in the clinical outcome, such as survival after a procedure; a patient may be most interested in functional outcome or the degree to which he was treated with respect and compassion; another stakeholder, such as a purchaser, may be interested in these issues as well but also in any costly complications that may have occurred or the degree to which the care was in line with accepted guidelines. Current report cards will have some of these aspects but cannot capture quality entirely from all perspectives.

The second challenge to validity is low sample size. Health conditions and treatments are placed into many categories, and a typical physician or hospital may only treat a few cases within a category over a particular time period. Report cards often deal with this common problem either by using a disclaimer ("interpret with caution"), stating that the provider with a small sample cannot be evaluated, or simply placing small-volume providers in with low-quality providers, with the thinking that high volume is a feature common to high-quality providers. For some conditions, it may take a few years for providers to have treated a sufficient number of cases to be evaluated. Minimum sample sizes or the use of confidence intervals or other statistical approaches are at times used, but there is no standard or foolproof solution.

Gaming is another threat to report card validity. Gaming can be thought of as an action taken to improve a quality measure that does not actually improve quality but yet is within the rules. For example, on learning that risk-adjusted survival for heart failure is below average, hospital leaders may want to investigate in order to gain an understanding for this apparently poor level of performance. Such an investigation might uncover that important comorbidities are not being documented in the medical record. Knowing that such a documentation oversight leads to a falsely low number of expected deaths, the leaders take corrective action. This work results in a more favorable value

for the quality measure, has a neutral affect on patient outcomes, and is allowable. Some experts feel that the best approach to this situation is not only to use measures that are not easily gamed but to use many measures so that providers find gaming too time-consuming and difficult and concentrate on improving quality.

Gaming is distinct from fraud. Most likely a much less frequent occurrence, fraud is deliberately and knowingly providing false information. For example, a provider may document that a step in a process was completed when in fact it was not. While we cannot know the exact frequency of fraud, it is worth considering that report cards are often derived from information provided by the party whose performance is being evaluated, and therefore the risk exists.

Old data are another potential challenge to validity. If a consumer, health plan, or purchaser is to base a decision on comparative quality data, one makes the assumption that the quality of care provided in the coming year will be similar to that provided during the year(s) addressed in the report card. For example, the Hospital Compare Web site allows one to evaluate 30-day mortality for patients admitted to U.S. hospitals with certain conditions such as acute myocardial infarction. By March 2008, one could compare risk-adjusted mortality rates for hospitals caring for patients between July 2006 and June 2007. While this represents an important advancement in quality measurement in the United States, one must consider that some studies have found only a weak correlation between risk-adjusted outcomes between time periods.

Risks

Publicly reported provider report cards bring certain risks. One is that providers will avoid patients that may make them look bad. For example, a heart surgeon who is judged based on patient survival may avoid operating on a patient whose risk of postoperative death is high. Although survival after coronary artery bypass surgery is typically adjusted for differences in case mix, this adjustment is imperfect and often greatly affected by small numbers such that any death, even in a high-risk patient, may mar the record of a surgeon. Even more important than the math is the perception of the surgeon. If the provider sees a high-risk patient as someone who might tarnish his or her reputation, then this may affect the decision to operate. Reports vary within the medical literature and lay media as to the frequency of avoidance of high-risk patients. Regardless, the theoretical risk is an important consideration.

Another risk of report cards is "treating the measure." In such a case, a provider may focus on a particular measure and in good faith try to improve the care provided but meanwhile disregard other patient needs that may be even more important. For example, consider a primary care physician being judged on the percentage of patients whose "bad" cholesterol is less than 130 mg/dl. The physician may be seeing a patient who would like help with her anxiety, for which there is no quality measure. Within the usual 10- or 20-minute visit, will the doctor address the high cholesterol, perhaps at the expense of addressing the patient's main reason for coming to the office? This is analogous to the concern that student test scores used in evaluation of school teacher performance may lead to "teaching to the test." The existence of a healthcare measure used to evaluate physicians and hospitals has the potential to interfere with the doctor–patient relationship.

Interpreting report card measures may involve some subjectivity and reflects on one's values and preferences. For example, consider hospitals that are being compared on measures of value—that is, the combination of quality and efficiency (cost). A common approach to categorize a provider as "good" is to require that the provider surpass two separate thresholds of quality and efficiency, respectively, as depicted in Figure 1.

Under such a methodology, hospitals that end up in the upper-right quadrant, therefore getting past both the quality and the efficiency hurdles, are labeled as best or most desirable. Clearly, any stakeholder would agree that Hospital A is not performing well (assuming valid data and well-defined measures). Also, it is clear that Hospital C is better than Hospital E, due to better efficiency and equal quality. But what about Hospital D? It does not meet the efficiency threshold, but it is highest in terms of quality. A consumer (patient) may see this hospital as most desirable. At the same time, a purchaser of healthcare may see Hospital B as highly desirable: acceptable quality at a low cost.

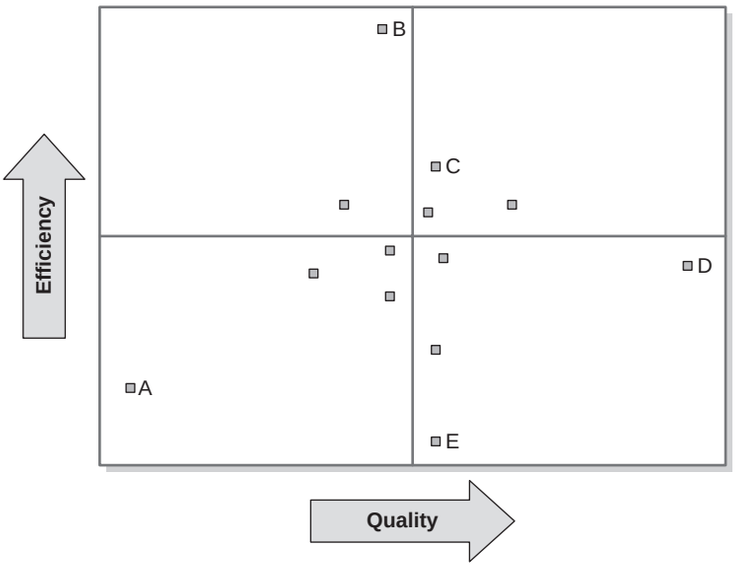


Figure 1 Comparing provider value

Note: In this hypothetical example, hospitals are compared based on varying levels of quality and efficiency (low cost is high efficiency). Arbitrary thresholds for each define the four quadrants.

If these were shoes rather than hospitals, there would be no problem. Each buyer would choose the combination of quality and cost that suits his or her preferences. But particularly in the complex U.S. healthcare system of third-party payers, agreeing on the definition of "best" is difficult.

There is no single solution to improving the value of healthcare, but it is important to consider report cards in the context of other movements of significance. A popular line of thinking is that report cards may be a partial solution that may also require some of the following: (a) pay for performance, in which provider pay is in some way dependent on quality and/or cost; (b) advances in information technology, particularly electronic health records; (c) expansion of healthcare insurance coverage; and (d) increasing consumer price sensitivity. Regarding the latter, consumer-driven insurance plans and tiered networks are examples of factors that influence consumers to choose high-value care.

These are not the only interventions that can lead to improved value for the healthcare dollar, but the relevant point is that report cards likely have their place within a multidimensional approach to improve a healthcare system rather than being an answer that works in isolation.

Christopher Hebert

See also Evidence-Based Medicine

Further Readings

Blackwell, R. D., Williams, T. E., Ayers, A. E., & Healy, B. (2005). *Consumer driven health care*. Emeryville, CA: Book Publishing Associates.

CMS/Department of Health and Human Services Hospital Compare: Retrieved January 21, 2009, from http://www.hospitalcompare.hhs.gov

Glance, L. G., Dick, A. W., Osler, T. M., & Mukamel, D. B. (2006). Accuracy of hospital report cards based on administrative data. *Health Services Research, 41*(4), 1413–1437.

Jha, A. K., & Epstein, A. M. (2006). The predictive accuracy of the New York State coronary artery bypass surgery report-card system. *Health Affairs, 25*(3), 844–855.

Neuhauser, D. (2002). Ernest Amory Codman. *Quality and Safety in Health Care, 11,* 104–105.

Porter, M. E., & Teisberg, E. O. (2006). *Redefining health care*. Boston: Harvard Business School Press.

Return on Investment

Return on investment (ROI) is the ratio of the net change in an investment's value (positive and negative) to the value of the original investment. The basic return on investment formula is ROI = (C_n / V_0), where C is the net change in value for the investment, n is the elapsed time since the initial investment, and V_0 is the initial value of that investment (at Time 0). This evaluation can be made for past, present, and future investments and their actual or projected changes in value.

Applications

While ROI can be expressed as a simple ratio, difficulties arise when a decision maker must compare investments that differ in their initial investment amounts and in the amounts and timing of their subsequent changes in investment value. Furthermore, difficulties are caused when tax rates, inflation, and expectations for what constitutes an adequate return on investment are considered in the analysis.

Medical Decision Making

ROI is an important factor in medical decision making because it provides a metric for use in comparing the relative economic attractiveness of alternative medical investments. All other factors being equal, one would assume that investments with higher ROIs would be preferred to investments with lower ROIs.

Expanding the Formula

The ROI formula can be elaborated as ROI = $(V_n - V_0/V_0)$, where V_n is the value of the investment at Time n and V_0 is the investment's initial value. This expansion makes it clear that the numerator includes all changes in the initial investment's value (gains and losses, realized and unrealized) as well as any cash flows (positive and negative) resulting from that investment. Using the example of an investment in a new clinic building, changes in the initial investment's value would include changes in the value of the building as well as cash flows from revenues and expenses involved in the ownership and operation of the building (e.g., interest payments for loans, property tax, and building maintenance costs).

Cash Flow Timing and Valuation Adjustments

When the investment period (time between V_n and V_0) is relatively short, differences in the timing of cash flows (between the initial investment and subsequent returns) is not an issue, and ROI is calculated without further adjustment. However, when cash flows occur over a longer period of time, they can have a significant impact on one's assessment of the value of alternative investment options, and special techniques are required to adjust for these differences. These adjustments take two forms:

1. *Compounding:* adjustments to the ROI estimate to account for previous cash flows and to account for the investment's changing value over time
2. *Discounting:* adjustments to individual cash flow values to account for the time lag between the initial investment and subsequent cash flow

Two additional types of adjustment may be used in ROI calculations. These are for taxes and inflation. Since both of these rates may be different for different years in the ROI analysis, it is customary to incorporate their effects in the annual calculations of net change in value (annual net cash flows).

Compounding

Assume an investment with an initial value of \$100 that reaches a final value of \$130 at the end of 3 years. One way of expressing the annual ROI is to take the arithmetic average of the change in value (\$10 = \$30/3). This yields an estimate of 10% ROI per annum 10% = \$10/\$100). However, this method does not account for the fact that if the annual ROI is actually 10%, the investment's initial value at the beginning of the second year will be the \$100 initial investment plus a 10% annual return on that investment during the first year. Therefore, using an arithmetic average overestimates the annual ROI. To calculate the annual ROI after adjusting for changes in the value of the

initial investment, one must solve the following equation for r: $V_n - V_0 = V_0 (1 + r)^n - V_0$, where r is the annualized rate of return on investment and n is the number of years in the investment period. Subtracting the initial investment value, this formula shows that the final value of an investment can be expressed as the investment's initial value compounded at an annual rate over a period of years ($V_n = V_0 (1 + r)^n$) In this example, the value for r that yields a final investment value of \$130 for a \$100 initial investment over a 3-year period is 9.1%. This is considerably smaller than the original estimate of a 10% annual ROI.

Discounting

To account for the financial opportunity that arises when investment returns are paid at the time of the initial investment, the value assigned to the different cash flow streams must be adjusted to recognize that the first option includes the potential for additional returns through the reinvestment of the earlier returns on the initial investment. This is accomplished by the use of discounting, which adjusts all cash flows to their values at Time 0 (the time of the initial investment). The formula for discounting is the reciprocal of the compounding formula. Such that the value of a cash flow at Time 0 is $CF_0 = CF_n/(1 + df)^n$, where 0 is the initial investment time, CF is the cash flow value, n is the elapsed time since the initial investment, and df is the discount factor. Using the above example, the discounted value of a \$100 initial investment with \$30 return received at the time of the initial investment is \$130; whereas, if the return is received at the end of 3 years and the expected return from alternative investments (discount factor) is 10%, it is only worth \$98.

Taxes

Frequently, different investments have different tax implications. In these situations, further adjustments should be made to account for these differences. The adjustment is as follows: $CF_{AT} = CF_{BT}(1 - t)$, where CF_{AT} is cash flow after tax, n is the elapsed time in years since the initial investment, CF_{BT} is cash flow before tax, and t is the annual tax rate.

Inflation

Inflation rates change over time and also may have implications for an investment's valuation. Here also, further adjustments can be made to account for these differences. The adjustment is as follows: $CF_{AI} = CF_{BI}(1 - i)$, where CF_{AI} is cash flow after adjustment for inflation, CF_{BI} is cash flow before adjustment for inflation, and i is the annual inflation rate.

Putting It All Together

Four types of adjustment in ROI analyses have been reviewed: (1) compounding, (2) discounting, (3) taxes, and (4) inflation. Including these calculations, the fully adjusted annual cash flow in period n is $ACF_n = CF_n(1/1 + r)^n)(1 - t_n)(1 - i_n)$. In this formula, adjusted cash flow in period n is a function of the unadjusted cash flow in that period; the rate of return, which is assumed to be constant for all periods; the annual tax rate for that period; and the inflation rate for that period. Return on investment throughout the investment period is

$$ROI = (ACF_1 + ACF_2 + ACF_3 + \cdots + ACF_n)/V_0$$

or

$$ROI = \frac{\sum_{n=j}^{t} ACF_n}{V_0}$$

where j is an individual and n is the final cash flow period.

As an example, let us consider a cash flow of \$10 that occurs 3 years after the initial investment for which there is a 10% discount rate, a 30% marginal tax rate, and a 2% inflation during that year. The adjusted cash flow value is $10(1/1.331).70 \times .98$. With these adjustments, the \$10 annual cash flow has an adjusted value of only \$5.15.

Comparing Alternative Investments

Individuals and organizations need methods for comparing the value of alternative investment options. For example, a hospital may want to assess differences in value between three options for their heart failure clinic: (1) constructing a new stand-alone building, (2) remodeling an existing

office building, or (3) leasing office space from the developer. Each of these options will require different initial investment amounts and will have different long-term cash flow requirements. The five primary methods for making such assessments are as follows:

1. Average rate of return
2. Payback period
3. Internal rate of return (IRR)
4. Net present value (NPV)
5. Profitability index

Each of these methods provides somewhat different insights into the values of competing investments; however, none of the methods alone is sufficient to give a complete assessment of the relative economic attractiveness of alternative investments.

Average Rate of Return

The average rate is an accounting method that merely averages the net change in investment value by the years of investment and divides by the initial investment amount. This is what is termed the *unadjusted ROI*. The formula is ROI = (C / V_0), where C is the arithmetic average annual cash flow.

Payback Period

A simple way of assessing an investment's return is to calculate how long it will take before the original value of the investment is recovered. This calculation is essentially the reciprocal of unadjusted ROI: PAYBACK PERIOD = (V_0 / C) As with the unadjusted ROI, this calculation assumes a constant change in investment value over each time interval, and it ignores any changes in value occurring after the end of the payback period.

Internal Rate of Return

Because of the shortcomings of the previous methods, it is generally agreed that methods are required that account for both the magnitude and the timing of cash flows. The internal rate of return (IRR) is the discount rate at which the cumulative change in investment value is equal to the value of the initial investment. Essentially, this is a break-even discount rate, where the initial investment and subsequent changes in that investment's value are equal. The formula is

$$V_0 \quad \frac{\mathrm{CF1}}{(1+df)^1} + \frac{\mathrm{CF2}}{(1+df)^2} + \frac{\mathrm{CF3}}{(1+df)^3} + \cdots + \frac{\mathrm{CF}n}{(1+df)^n}.$$

This can also be written as $\sum_{t=0}^{n} (\mathrm{CF}_t/(1+r)^t) = 0$, where the cash flow in Time 0 includes the initial investment. The advantage in the IRR decision method is that it incorporates discounting in the analysis. The primary disadvantage is that it does not account for the size of the original investment.

Net Present Value

The NPV decision method accounts for the size of the original investment by calculating the investment's estimated final value assuming a required rate of return. The formula is

$$\mathrm{NPV} = V_0 + \frac{\mathrm{CF1}}{(1+df)^1} + \frac{\mathrm{CF2}}{(1+df)^2} + \frac{\mathrm{CF3}}{(1+df)^3} + \cdots + \frac{\mathrm{CF}n}{(1+df)^n}.$$

Alternatively, this can be written as

$$\mathrm{NPV} = \sum_{t=0}^{n} \left(\frac{\mathrm{CF}t}{(1+r)^t} \right),$$

with Time 0 including the initial investment as a negative cash flow. This method yields the expected gain or loss from an investment assuming that the required rate of return is met.

Profitability Index (Benefit-Cost Ratio)

The profitability index is the adjusted ROI. In this method, the adjusted cash flows are divided by the value of the original investment. The formula is

$$\mathrm{PI} = \frac{\sum_{t=1}^{n} \frac{CFt}{(1+r)^t}}{V_0}.$$

Using this decision rule, an organization would accept investments with a profitability index greater than 1 and reject those with indices less than 1.

Let us assume that the cost of constructing a stand-alone building for a heart failure clinic is $1,000,000 and that the annual net cash flow (revenues less expenses) will be $300,000. In contrast, let us assume that $200,000 will be required to remodel an existing building for the clinic and that this alternative's annual net cash flow will be $150,000. The expected net cash flow (total annual cash flow less investment) over a 10-year investment period is $2,000,000 for the build option and $1,300,000 for the remodel option. The average rates of return are 30% and 75%, and the payback periods are 3.3 years and 1.3 years for the build and remodel options, respectively. However, these metrics do not account for differences in cash flow timing, taxes, and inflation. Assuming a 30% marginal tax rate and a 2% inflation rate, the internal rate of return for the build option is 14.1%, whereas it is 48.7% for the remodel option. If we assume a 10% discount factor, total cash flows are $176,895 ($1,176,895 less the $1,000,000 investment cost) for the build option and $388,448 ($588,448 less the $200,000 investment cost) for the remodel option. Thus, the profitability index is .176 for the build option and 1.942 for the remodel option.

Method Selection

Selecting between methods for evaluating investments may not be straightforward. Situations will arise in which there are conflicts between the investment decision recommendations from different assessment methods. Thus, the adjusted ROI (the profitability index) supplemented with the net present value, and perhaps an internal rate of return calculation, will give a better overall assessment of the investment's potential value than can be derived from using one measurement alone.

Eric L. Eisenstein

See also Cost-Comparison Analysis; Costs, Opportunity; Discounting

Further Readings

Copeland, T., Koller, T., & Murrin, J. (2000). *Valuation: Measuring and managing the value of companies*. New York: Wiley.

Kaushal, R., Jha, A. K., Franz, C., Glaser, J., Shetty, K. D., Jaggi, T., et al. (2006). Return on investment for a computerized physician order entry system. *Journal of the American Medical Informatics Association, 13,* 261–266.

Li, J. S., Eisenstein, E. L., Grabowski, H. G., Reid, E. D., Mangum, B., Schulman, K. A., et al. (2007). Economic return of clinical trials performed for pediatric exclusivity. *Journal of the American Medical Association, 297,* 480–488.

Maviglia, S. M., Yoo, J. Y., Franz, C., Featherstone, E., Churchill, W., Bates, D. W., et al. (2007). Cost-benefit analysis of a hospital pharmacy bar code solution. *Archives of Internal Medicine, 167,* 788–794.

Weeks, W. B., Wallace, A. E., Wallace, M. M., & Welch, H. G. (1994). A comparison of the educational costs and incomes of physicians and other professionals. *New England Journal of Medicine, 330,* 1280–1286.

Zelman, W. N., McCue, M. J., Millikan, A. R., & Glick, N. D. (2008). *Financial management of health care organizations: An introduction to fundamental tools, concepts, and applications*. Malden, MA: Wiley.

Risk Adjustment of Outcomes

Risk adjustment facilitates meaningful comparisons of outcomes of different groups of individuals by accounting for differences across the groups in baseline characteristics that could affect their outcomes. Groups can be defined in countless ways depending on comparisons of interest, such as patients admitted to one hospital versus another, individuals receiving Treatment X versus Treatment Y, and persons in one socioeconomic stratum versus other strata. In observation studies, individuals are not randomly assigned to the groups, and, for reasons that are sometimes poorly understood, those in one group may differ in significant ways from those in other groups. These differences might affect the likelihood that the individuals will experience the outcomes of interest. For example, if patients at the neighborhood hospital are older on average than those admitted to the downtown teaching facility, it is unclear whether the higher

mortality rate at the community hospital is caused by worse care or older patients. In these types of observation studies, the goal of risk adjustment is to take into account—or adjust for—the effect of important risk factors, so that analysts can more confidently attribute differences in outcomes to the variable of interest (e.g., hospital quality) than to underlying characteristics of the individuals in the groups (e.g., older vs. younger age).

The gold standard for determining the effects of an intervention is the randomized controlled trial (RCT). Randomization ensures that, on average, individuals assigned to receive an intervention have similar baseline characteristics (unmeasured as well as measured attributes) to those randomized to the control group that does not receive the intervention. Thus, an RCT that uses an intention-to-treat analysis—that is, compares outcomes of all persons assigned to the intervention versus control group regardless of the treatments patients actually receive—provides a theoretically sound basis for concluding that differences in outcomes are *caused by* the intention to use a different treatment.

However, RCTs are neither ethical nor practical in many situations. The key ethical concern is an unwillingness to substitute an untried treatment for one that is widely used, even if its utility is largely untested, or a treatment with well-documented but modest benefits. The numerous practical impediments include high costs and lengthy time horizons required to conduct RCTs, challenges of human subjects protections and obtaining truly informed consent, and refusals of physicians and patients to participate in randomized studies even when widely used standard therapies have little rigorous scientific evidence supporting their benefits. Another difficulty in planning RCTs is the tension between answering a narrowly defined question well (e.g., by excluding patients with extensive comorbidities) versus the ability to generalize findings more broadly (e.g., by including the full range of patients who might be candidates for the treatment should the RCT find it effective). Finally, RCTs are not well suited for answering many important questions where it is infeasible to randomly assign individuals to different groups, such as comparing the quality of care at Hospital X versus Hospital Y by contrasting patients' outcomes at these respective institutions. Randomly assigning patients to different hospitals is not possible in today's environment. When data from RCTs are not available, inferences about the value of interventions must come from observational studies. Risk-adjusted comparisons have become a standard method for using observational data to study treatment effectiveness, as well as to facilitate quality monitoring and support other health policy initiatives, such as pay-for-performance programs being implemented by government and private payers. These value-based purchasing programs aim to pay higher quality and more efficient providers more than providers offering lower value: Credible judgments about provider quality or efficiency require adjusting for differences in the types of patients seen by each provider.

Risk adjustment plays a behind-the-scenes role when comparing outcomes across providers. Another role of risk adjustment is more out-front—that is, serving as the explicit means to predict outcomes for individuals within populations. An example is risk-adjusting payment levels for healthcare services. Medicare prospective payment systems for hospitals, nursing homes, inpatient rehabilitation facilities, and home healthcare all use some form of risk adjustment, setting payment levels for services or groups of services based on specified health conditions and other characteristics of the individual Medicare beneficiary. Another example is setting payment rates for capitated healthcare plans. In these various payment examples, the goal is to create appropriate incentives to provide high-quality care to all people, regardless of the complexity or extent of a patient's health condition. By accounting for differences in illness complexity, risk-adjusted payment reduces incentives to avoid sick patients (who might have high relative costs) in favor of healthier patients (who might have much lower costs and thus produce higher profits). Without risk adjustment, providers face strong incentive to "skim" or "cherry pick" patients who require less costly care, leaving the sicker patients with fewer options.

Although the concept of risk adjustment is straightforward, designing and implementing risk adjustment confronts multiple challenges, from defining an appropriate conceptual framework to identifying data sources containing essential risk factors to using appropriate statistical methodologies. This entry briefly discusses selected issues

central to risk adjustment. One very important practical issue is not addressed: data sources. Data sources vary widely, including detailed primary data collection designed and conducted by researchers (e.g., in an observational study of treatment effectiveness) to huge data files generated by administering reimbursement systems (e.g., Medicare claims files). Though issues relating to these widely ranging data sources are not discussed, all topics presented below must eventually link back to the content, size, and nature of the intended data source. Selected conceptual issues, such as potential outcomes and candidate risk factors, and issues related to the statistical models used to do risk adjustment, are considered in what follows.

Risk of What?

As described above, risk adjustment facilitates comparisons of outcomes across populations whose baseline characteristics—especially those characteristics directly related to the outcome of interest—differ. However, the relationship between persons' characteristics, such as age and health status, and outcomes depends on the outcome. For example, widely metastatic cancer (an indicator of extent or severity of disease) could have different relationships to two policy-relevant outcomes: costs and death. Certainly, metastatic cancer increases the risk of imminent death, but it might decrease healthcare costs if persons choose relatively less expensive palliative care compared with the aggressive, potentially curative treatment sought by someone with early-stage cancer. Thus, a single risk adjustment model will not work equally well across all outcomes. Before conceptualizing a risk adjustment framework, analysts must clearly specify the outcome of interest. The next step is to specify those characteristics of individuals that increase or lower risks of experiencing that outcome.

Dimensions of Risk

Including all potential risk factors in a risk adjustment model is virtually impossible—complete information on all relevant risk factors is rarely available. Nonetheless, when designing a risk adjustment method, analysts should first go through a conceptual exercise of identifying all the various risk factors that, in an ideal situation, they would want to include. This exercise will help guide selection of data sources. And, by explicitly identifying key risk factors not included in the model, analysts have a better understanding of factors that could potentially explain differences in risk-adjusted outcomes across groups of interest. In what follows, various attributes that might be relevant to different outcomes are briefly discussed.

Age and Gender

Age has an independent effect on many outcomes even after accounting for other risk factors. However, age by itself and in addition to other risk factors often has little effect on the ability of risk adjustment models to predict outcomes. Some studies have shown that women have poorer outcomes than men following certain treatments, such as for coronary artery disease. However, like age, gender usually contributes little to risk adjustment model performance. Nevertheless, because of their face validity and easy availability, risk adjustment models usually include age and gender. Sometimes, models with only age and gender as independent variables serve as baselines for measuring improvements in model performance when other risk factors are added.

Race and Ethnicity

Disease prevalence and other health outcomes frequently differ by race and ethnicity. Measuring race and ethnicity is much more problematic than age or gender, and reliable data on these factors are often not available. Variables reflecting race and ethnicity are significant predictors of health outcomes, although the extent to which statistical significance reflects differences in socioeconomic status, patient preferences, discrimination, or biology is not clear. Due to inadequate measurement and small sample sizes, risk adjustment models may include race and ethnicity only by distinguishing black from white. Medicare data classify beneficiaries into white, black, nonblack Hispanic, and other. Until recently, most private insurers have not collected information on race and ethnicity. However, in response to continuing reports of disparities in diagnosis, treatment, and patient outcomes by race and ethnicity, and to new federal guidelines about coding multiracial identity, more

databases will contain detailed information on race and ethnicity.

Socioeconomic Status

Socioeconomic disparities in health status are well documented. For example, patients living in economically disadvantaged areas in the United States have poorer survival rates following surgery, and those without health insurance have poorer health. Often, individual-level socioeconomic data are not available and census tract or zip code level data based on the patient's residence are used instead. In some research studies, using zip code level socioeconomic data for risk adjustment has the same effect as using individual-level data. However, even when individual income measures are available, area-based income measures provide additional information about the context within which individuals receive their care since low- and high-income neighborhoods are likely to have very different environmental exposures and healthcare delivery infrastructures. Social status, such as living alone or being unmarried without close friends, is also associated with outcomes.

Functional Status

Standard functional status measures include basic activities of daily living (ADLs: eating, bathing, dressing, toileting, walking) and instrumental ADLs (IADLs: shopping, cooking, doing housework, using public transportation, balancing a checkbook). More comprehensive measures consider cognitive abilities, affective health (e.g., happiness, anxiety), and social activities. Numerous instruments exist for assessing function status; some are disease specific and others generic (independent of diagnosis). Functional status is a strong predictor of important outcomes, including in-hospital and 30-day posthospital mortality and patient satisfaction.

Health Behaviors

Health behaviors, including tobacco use, nutritional practices, level of physical activity, alcohol consumption, illicit drug use, sexual practices, societal and domestic violence, and seat belt use, are often related to outcomes. Differences in health behavior clearly contribute to some of the observed health differences by socioeconomic status.

Acute Clinical Stability

Acute clinical stability reflects patients' current physiologic functioning and is an important risk factor when examining clinical outcomes of acutely ill patients over short time frames, such as deaths in intensive care units or following urgent or semi-urgent surgery. Measuring acute clinical stability requires basic physiologic values, such as heart and respiratory rates, serum chemistry and hematology findings, arterial oxygenation, and level of consciousness.

Severity of the Principal Diagnosis

Medicare defines the principal diagnosis as the leading disease that caused a hospitalization. Clinicians have developed disease-specific severity algorithms for certain diseases. For example, oncologists stage cancers generally based on the size and characteristics of the tumor and extent of its spread throughout the body; the New York Heart Association has developed a widely used classification system for heart failure. But many conditions do not have well-established severity rating schemes.

Comorbidities

Comorbidities are diseases unrelated in etiology to the principal diagnosis. Comorbidities differ from complications, which are sequelae of the principal diagnosis. Prototypical comorbidities are chronic conditions such as diabetes mellitus, chronic obstructive pulmonary disease, or ischemic heart disease. Many studies have shown that comorbidities increase the likelihood of poor outcomes.

Decisions About Risk Factors

Many outcomes are easy to specify and their relationship to potential risk factors generates little debate. Other outcomes, however, raise important questions, often related to how the risk-adjusted outcome information will be used (e.g., setting payment rates, publicly reporting quality indicators)—questions that hold implications

about whether or not to adjust for certain risk factors. For example, experts disagree about whether to risk-adjust process of care measures, which are frequently included in public reports of provider performance. The case of screening mammography, which is recommended for women between the ages of 50 and 69, exemplifies this concern. Since all women between 50 and 69 should receive mammography, should analysts risk-adjust comparisons of mammography screening rates? Research suggests that education level is closely related to the willingness of women to undergo mammography: Less educated women are less likely to adhere to mammography recommendations from their physicians than highly educated women. Some physicians have more highly educated patient panels than do other physicians. In this situation, failure to adjust for education when comparing mammography rates across physicians could have a perverse effect. A facility with a high proportion of poorly educated patients could receive a worse rating than a second facility with mostly college-educated patients even if the first practice achieves higher screening rates with both its less educated and highly educated patients.

A broader question arises for risk factors that might reflect not only intrinsic patient characteristics but also provider attitudes. As noted earlier, many studies have documented racial disparities in healthcare service use, as well as outcomes. Researchers are still investigating the multiple factors that have produced such disparities. Should a variable such as race, that is not only often strongly associated with an outcome of interest but also perhaps with discriminatory attitudes, be included as a risk adjuster? To the extent that the different distribution of race by intervention status reflects discrimination, adjustment for race can "excuse" this discrimination. To the extent it reflects real differences in the difficulty of achieving certain outcomes by race, failure to adjust for race results in biased estimates of effectiveness. If there are sufficiently many cases in each racial group, a reasonable strategy is to run the risk adjustment models separately by racial category.

A similar type of issue arises concerning a risk factor that is not only a characteristic of the patient but may also be a reflection of provider behavior. To illustrate, consider pressure ulcer as an outcome of nursing home care. The risk of developing a pressure ulcer depends both on patients' risk factors and the quality of nursing home care. These dual factors pose challenges to using risk-adjusted pressure ulcer rates as an indicator of nursing home quality. ADL limitations are a risk factor for patients developing pressure ulcers. However, the quality of nursing home care can also affect ADL performance. Including ADL limitations in the risk adjustment model excuses the nursing home for poor care that results in ADL limitations. However, even the highest-quality nursing home care cannot prevent ADL declines for certain patients with progressive debilitating conditions. Failure to include ADL limitations as risk factors in the model predicting pressure ulcers will produce biased estimates of nursing home quality.

Recently, initiatives to measure quality of care are turning to *composites*—combinations of multiple quality indicators into a single score. Ongoing research is exploring how to best create composite measures. Whatever approach is used to combine the individual quality metrics, it is important that, where appropriate, the individual measures first be risk adjusted. In that way, variables that both conceptually and empirically are related to each dimension of quality can be included in the risk adjustment model. Composite measures can pose data challenges if different indicators within the composite require adjustment for very different risk factors. Nevertheless, failure to consider risk adjustment when specifying composites could complicate interpretations of composite scores.

Risk Adjustment Systems

A large number of risk adjustment systems have been developed, many for specific, targeted purposes. In what follows, several widely used systems that are either free or available for research purposes for a modest fee are briefly discussed.

Probably the most widely used system to measure the impact of comorbidities on outcomes is the Charlson index, a score that is the sum of whole number weights assigned to each of 19 serious medical conditions. This index can be implemented in administrative (claims) data from diagnoses coded in the ICD-9-CM classification system. Clinical Classification Software (CCS) from the Agency for Healthcare Research and Quality provides another popular way to use ICD-9-CM data

for risk adjustment. CCS maps all ICD-9-CM codes into 259 categories that also roll up into 30 broad diagnostic categories. The CCS software, which can be downloaded for free, does not provide a summary morbidity score. To account for comorbidities in a risk adjustment model, each disease category is entered as a dummy or indicator variable in a model that may also include age, sex, and other predictors. Fitting the 259-category version of CCS to a very large data set results in excellent risk prediction, but models with so many indicators (some of which flag rare diseases) should not be attempted with fewer than 100,000 observations. Smaller data sets require either fitting coarser comorbidity models to the data (e.g., using the 30 broad CCS categories) or summarizing total morbidity through a summary risk score developed on a large, benchmark data set.

Two sophisticated risk classification systems that use ICD-9-CM codes to produce summary risk scores precalibrated to large populations are used extensively in the research literature: Adjusted Clincal Groups (ACGs) and Diagnostic Cost Group or Hierarchical Condition Category models (DCG/HCC). Each of these software products offers various modeling structures for different purposes and provides guidance for best use. They are proprietary products but are available for use in research projects for a modest licensing fee.

While Charlson, ACGs, and DCGs provide all-purpose risk scores, which summarize a patient's illness burden in different settings, such as during a year of care, other methods are designed specifically for risk-adjusting hospital admissions. The most widely used of these is Diagnosis Related Groups (DRGs). Broadly, DRGs classify admissions in the same category when they have the same "principal diagnosis" and "major procedure" (if any), in some cases with distinctions based on the presence or absence of complications or comorbidities. Payers (such as Medicare) typically view hospitalizations within the same DRG as having the same expected resource consumption. However, among admissions in the same DRG, the nature and extent of secondary diagnoses present may be important. The All Patient Refined Diagnosis Related Groups (APR-DRGs) software licensed by 3M, for example, assigns "severity" categories within DRGs. One version reflects expected differences in cost; another, differences in risk of mortality. APACHE II is a widely used system that summarizes the expected effect of physiologic variables, age, and comorbid conditions on short-term mortality for patients in intensive care units.

Models for Risk Adjustment

Risk factors are often confounders—that is, variables that are both related to outcomes and that differ across groups of interest. Failure to adjust for confounders can produce biased estimates of the effect of group membership. Multivariable models are the principal method used to both address confounding and to make predictions based on the risk profiles of groups of patients.

Continuous Outcomes

Multiple regression models can be used to estimate the effect of an intervention (or group membership) on a continuous outcome, while accounting for possible confounders. The basic form of a multiple regression model is

$$E(Y_i) = a + \sum_j b_j X_{ij} + cI,$$

where Y_i is the value of the outcome for the ith patient, X_{ij} is the value of that patient's jth risk factor, and the intervention I is coded as a "dummy variable" equaling "1" if the patient receives the intervention and "0" otherwise. $E(Y_i)$ is the expected (or predicted) value of Y_i. In such a model, the effect of the intervention on Y after adjusting for the risk factors (the X_{ij}s) is estimated by the coefficient c. If the expected value of Y is Y_0 for someone without the intervention, their expected value if they received the intervention is $Y_0 + c$. The X_{ij}s may be continuous valued or dummy variables, and they may include polynomial terms or interaction terms.

Model parameters are usually estimated using ordinary least squares. Ordinary least squares estimates parameters so that predictions are as close to actual values as possible, specifically, to minimize $\sum_i (Y_i - \text{PRED}_i)^2$, where PRED_i is the predicted outcome for the ith person.

Statisticians and clinicians should work closely together to model clinically based risk factors. Clinicians provide insights regarding the reliability and validity of measured variables, relevant time

frames for measuring variables, the plausibility of nonlinear relationships, and the meaning of extreme or missing values. For example, clinicians might indicate that certain physiologic parameters may not be measured because they are presumed to be normal. In this instance, analysts might reasonably impute normal values for missing values. In other instances, clinicians may feel less comfortable assigning normal ranges to missing values.

When interest is in the quality or efficiency of a number of different providers, cI may be replaced by $\sum_k c_k I_k$, where the I_ks are dummy variables for all but one of the providers. Here, the providers are treated as so-called fixed effects, with each c_k being the estimated increment to the expected value of Y associated with the patient being in the kth provider's panel rather than the omitted provider's panel. Random effects model (also called *empirical Bayes*, *hierarchical*, or *mixed models*) may also be used to examine multiple providers. A simple example of a random effects model is

$$E(Y_i) = a_k + \sum_j b_j X_{ij},$$

where the a_ks (one for each provider) are treated as independent, identically distributed random variables that estimate the effect of provider k on the outcome after adjusting for patient-level risk factors (the X_{ij}s). Random effects models are particularly appropriate for estimating the effects of provider-level characteristics, such as being a high- versus low-volume surgeon or hospital.

Some continuous outcomes, particularly those related to resource consumption, are highly skewed; in particular, their values have a long right tail. The traditional approach to address this skewness is to fit an ordinary least squares model with the logarithm of cost, ln(cost), as the dependent variable, thus assuming that the outcome follows a lognormal distribution. This approach produces predictions in log dollars. Thus, for example, an error of $2,000 when $5,000 is expected is treated as equally serious as an error of $20,000 when $50,000 is expected. Also, the model's predictions do not readily convert to unbiased estimates in the original dollar scale. An alternative approach is to use a generalized linear model. Generalized linear models are characterized by two features, a flexible family of distribution functions (F) for the outcome variable and a choice of a link function (g), which describes the scale on which covariates in the model relate to the outcome. A typical model when the outcome is resource consumption uses a natural logarithm (ln) link function,

$$g(E(Y_i)) = \ln(E(Y_i)) = a + \sum_j b_j X_{ij} + cI,$$

with Y_i hypothesized to follow a gamma distribution. Parameters are selected to produce a best fit between the actual and predicted values in the original scale, not between actual and predicted values in a log scale (as ordinary least squares on log-transformed data does).

If $\ln(E(Y_i)) = a + \sum_j b_j X_{ij} + cI$, then $E(Y_i) = \exp(a + \sum_j b_j X_{ij} + cI) = \exp(a + \sum_j b_j X_{ij}) * \exp(cI)$. In a logarithmic model, the covariates act multiplicatively on the mean. For example, if $\exp(c)$ is .9, the expected cost of patients receiving the intervention is 10% lower than those who do not receive it.

A growing body of literature, suggests that—especially with large data sets or with only mildly skewed outcome variables (e.g., costs of hospital admissions for people admitted with similar medical problems)—simply modeling the outcome directly using ordinary least squares provides as good or better predictions than most of the more sophisticated models that have been proposed for handling skewed outcome variables. When data are very skewed, such as total costs of healthcare during a calendar year of all enrollees in a commercially insured population, which will include many people with no costs and a small number with very high costs, two part models are often used. One model predicts the likelihood of any healthcare costs (a dichotomous outcome), and a second model predicts costs of those who have some healthcare costs.

Regardless of the model used, it is important to examine the distribution of all variables and to address problems arising from the presence of extreme outliers. For example, in a general population of working adults and their families, one "million-dollar baby" can have a large effect on a regression model, even in a large data set. It is useful to top-code to reduce the influence of such outliers. For example, when top-coding at $250,000, all values larger than 250,000 are reset to 250,000.

As noted, a number of risk adjustment systems provide risk scores. These scores are particularly

useful for researchers with modest-sized study populations, where it is not feasible to include a large number of dummy variables for different conditions. To illustrate the use of such a score, imagine a risk score (RS) that expresses next year's expected healthcare costs for a person as a multiple of the average cost in a benchmark population (e.g., 6 million persons with employer-based insurance). An RS of 1.4, for example, indicates that the person's expected resource consumption is 40% greater than average. Including RS as a predictor in a regression model is an easy way to risk-adjust outcomes for differences in overall morbidity burden in a specific population of interest. However, the simplest way to convert such scores into predictions in a specific population is to multiply them by a proportionality constant k, calculated as [AVE (outcome)/AVE (RS)], where the average AVEs are computed on the cases in the specific population of interest. If each person's predicted value is $k * \text{RS}$, then the average expected outcome in the specific population will equal its average actual value.

An analyst might want to use the risk score to predict an outcome that is different from the outcome used in the original model to calculate risk scores. For example, the resource consumption-based RS above might be used to predict hospitalizations or mortality. Generally, risks for different outcomes are related, though usually not linearly. An easy way to deal with this is by *risk score bucketing*. Cases are ranked by RS and cutpoints identified that divide the population into a modest number of categories, perhaps deciles of increasing risk. Alternatively, when predicting a highly skewed variable such as cost, more refined estimates for the higher-cost cases might be more useful, which suggests using uneven-sized buckets, for example, making cuts at percentiles 20, 50, 80, 90, 95, 99, and 99.5. The expected outcome for each case is the average outcome for everyone classified in the same bucket. When using this method, each bucket should have enough cases to achieve a fairly stable average. When analyzing costs, buckets of 500 or more cases are desirable.

Dichotomous Outcomes

Logistic regression models are typically used to model dichotomous outcomes. Let p_i be the probability that person i experiences an event such as death. Then, $p_i/(1 - p_i)$ equals "the odds" that the person experiences the outcome. In a logistic regression model,

$$\ln[p_i/(1 - p_i)] = a + \sum_j b_j X_{ij} + cI.$$

Here, the effect of the intervention is modeled as multiplying the odds. For example, suppose that patients without the intervention have a 25% chance of the event occurring and $\exp(c) = 2$. Their odds are 1/3 (=.25/.75) without the intervention and 2/3 ((1/3) $*$ 2 = .40/.60) with it. That is, receiving the intervention raises their chances of the event occurring from 25% to 40%.

Time to an Event

Another fairly typical outcome is duration of survival (time-to-death). The Cox proportional hazards model is widely used when modeling this type of outcome. Let $h(t)$ = the probability that a person who has survived to time t dies prior to time $t + \Delta$, where Δ is a small interval of time. This "instantaneous hazard" of dying at time t is modeled as

$$\begin{aligned} h(t) &= h_0(t) \exp(a + \textstyle\sum_j b_j X_{ij} + cI) \\ &= [h_0(t) \exp(a + \textstyle\sum_j b_j X_{ij}] * \exp(cI). \end{aligned}$$

Here, $\exp(c)$ multiplies the hazard for people who receive the intervention. If c is negative, $\exp(c)$ is less than 1, meaning that the intervention reduces the hazard, or chance, of death. Each hazard function is uniquely associated with a survival function $S(t)$, which equals the predicted probability that a person survives from time 0 until time t. If c is negative, $S(t)$ for persons with the intervention is greater than $S(t)$ for persons with the same characteristics who do not receive the intervention; also, the gap between the two functions widens over time.

Summary Measures of Performance of Risk Adjustment Models

One frequently asked question is how well a specific risk adjustment model accounts for actual differences in patient outcomes (i.e., the model's predictive validity). No single summary measure

can fully reveal how valid a risk adjustment method is for a particular purpose. However, several widely used summary measures exist for predicting different types of outcomes.

Predicting a Continuous Outcome

Let PRED_i be the predicted value of the outcome variable for patient *i*. The standard summary measure of model performance when multiple regression is used to make the prediction is R^2, which equals

$$1 - [\sum_i (Y_i - \text{PRED}_i)^2 / \sum_i (Y_i - \text{AVE}(Y)^2],$$

where AVE(*Y*) is the average of the Y_is. R^2 is often described as the fraction of total variability in the outcome explained by differences in risk among cases in the sample. Because ordinary least squares regression produces models that maximize R^2 (since it minimizes $\sum_i (Y_i - \text{PRED}_i)^2$), ordinary least squares software calculates the model R^2. However, R^2 is a useful summary measure of how well a set of predictions "fits" the actual outcomes regardless of the modeling algorithm used. The easiest way to calculate R^2 for a set of $\{Y_i, \text{PRED}_i\}$ pairs is to calculate the Pearson correlation coefficient (ρ) for the two variables and square it.

Though useful for comparing the predictive accuracy of different models on the same data set, R^2 does not provide an intuitive indication of how well a model performs. For this purpose, it is useful to examine actual and predicted values of the outcome within deciles of the predicted outcome. The best models are both well calibrated (as evidenced by having similar average values of predicted and actual outcomes within each decile of predicted risk) and discriminate well (as evidenced by the actual mean outcomes in the highest deciles being much larger than those in the lowest deciles).

The *c* statistic is the most widely used performance measure for models predicting a dichotomous outcome. Using death as an example, one of several equivalent definitions of the *c* statistic is the following: Consider all possible pairs formed by selecting one patient who died and one who lived; the *c* statistic is the proportion of these pairs in which the predicted probability of death is higher for the patient who died than for the one who lived. The *c* statistic is close to 1.0 when the model assigns nearly all the highest scores to the patients who died and near .5 when its predictions are no better than random. The *c* statistic only measures discrimination, not calibration. A Hosmer-Lemeshow chi-square test is often used to evaluate calibration, with the aim that the test will *fail to reject* the null hypothesis that the numbers of deaths predicted by the model are a good fit for the actual numbers of deaths across the 10 deciles of predicted risk. Like other chi-square tests, it has the weakness that the null hypothesis is almost certain to be rejected, even with a good model, when there is a large number of cases.

When evaluating model performance, it is important to distinguish how well the model fits the data set used to estimate model parameters from how well it predicts outcomes in other settings. There are two main approaches for model validation. In cross-validation, analysts use part of a data set to develop the model and the remaining data to validate its performance. In independent validation, analysts apply the model to entirely new data. In cross-validation, part of the data (typically one half or two thirds of the data set) is used to develop a model; this model is then used to make predictions in the validation data, with "validated" measures of performance being calculated from the (Y_i, PRED_i) pairs. Validating a model by testing the accuracy of its predictions in an "entirely" new data set is the strongest form of validation. If the model does not validate well, the model itself may not be the problem. For example, imagine developing a model to predict mortality in one year and applying it to next year's data. Predicted mortality next year could be greater than actual because treatment approaches have improved, or because diagnosis coding is more complete, making patients appear sicker than they would have been in the earlier data. In general, one tends to have more confidence in a model's ability to capture stable relationships within a data set (this kind of patient is more likely to die than that kind) than to correctly predict actual levels of an outcome in new data. Thus, it is more common to test prediction in the new data but not model calibration. *Y* is regressed on PRED in the new data, which automatically recalibrates predictions:

$$\text{NewPRED} = a + b * \text{PRED}.$$

This forces the average of newPRED to equal the average of Y in the new data. One can then examine the closeness of fit in the new (Y_i, newPRED$_i$) pairs, as described above.

Extensions of Standard Multivariable Modeling

Propensity Scores

If the relationship between risk factors and an outcome is correctly modeled, c, the coefficient associated with the intervention is an unbiased estimate of its effect. However, all models are imperfect. Model predictions will typically underestimate "true" risk for some kinds of patients and overestimate it for others. Suppose that the model overestimates the risk of death for patients with attributes that make them more likely to receive an intervention. Then, even in the absence of an intervention, their actual outcomes will be better than predicted by the model. This better-than-expected outcome will be attributed to the intervention and will be reflected in a biased estimate of c. Propensity scores are a post hoc way to redesign studies using observational data by subsetting or reweighting the original observations in a way that substantially reduces differences in the distribution of measurable baseline risk factors between cases and controls. Because in the subset of cases used in the propensity score analysis (or in the reweighted sample of cases) similar proportions of cases and controls have each type of risk factor, similar proportions of each group will be over- or underestimated by similar amounts. As a result, propensity score analysis protects against biased estimates of treatment effects due to an imperfectly modeled relationship between the measured risk factors and the outcome.

A typical propensity score approach proceeds as follows:

1. Develop a logistic regression model to predict the probability of receiving an intervention from the risk factors. This predicted probability is the propensity score.
2. Divide the population into quantiles (e.g., quintiles or deciles) of the propensity score.
3. Within each quantile, sample equal numbers of cases and controls.

More sophisticated matching or weighting approaches have been proposed to produce the desired goal of focusing on a set of cases and a set of controls with similar distributions of the propensity score. These groups will be much more balanced on all baseline risk factors that were used in the propensity score model, a fact that can be easily verified by examining the distributions of risk factors in the subsampled (or reweighted) cases and controls. The subsample of cases and controls is then analyzed using standard techniques, although, because the groups are well balanced, multivariable modeling is more useful at this point for improving the accuracy of the estimated effect of the intervention than for removing bias due to confounding.

A propensity score approach eliminates the need for very complex modeling (e.g., including interactions and nonlinear terms) in an attempt to try to compare groups that are extremely different, and it highlights the extent to which cases and controls exist with similar baseline risk factors. Also, if there are enough cases, it allows examination of the effect of the intervention separately for the kinds of patients who rarely receive it (those in the lowest quantile of the propensity score) to others (in the higher quantiles) who more commonly receive it.

Instrumental Variables

Everything discussed so far addresses protecting against confounding from risk factors that have been observed or measured. Instrumental variable (IV) methods can be used to adjust for unmeasured as well as measured risk factors. The IV approach relies on assumptions about plausible scenarios that might capture factors that cannot be measured. The key to this approach is identifying an observed variable that (a) is associated with the likelihood of receiving the intervention but (b) does not directly affect the outcome. Such a variable is called an *instrument* (for studying this outcome). To illustrate the approach, imagine two equally skillful providers, A and B, with patient panels that are similar with respect to both measured and unmeasured variables. As a matter of personal preference, A uses intervention I on more of his patients than Provider B. In this case, the provider is a good instrument. Suppose that 80% of A's patients and 60% of B's patients receive I and that 35% of A's

patients and 29% of B's patients have a good outcome. For simplicity, assume that these percentages are estimated without error. If in fact both providers treat similar patients and provide similar quality of care, the only factor that can account for A's better success rate is his or her greater use of the intervention. The IV estimate of intervention effect is

$$\text{(Change in percent with a good outcome)/(Change in percent receiving I)} = (.35 - .29)/(.80 - .60) = .30.$$

To clarify the interpretation of the effect estimate, imagine 100 patients who switch from Provider B to Provider A. Twenty of these will get the intervention who would not have if they did not switch, and overall 6 more patients among these 100 will have a good outcome; that is, for every 10 extra people who get the intervention, 3 extra people will have a good outcome. In real applications, randomness is formally addressed and two-stage modeling is often used to estimate treatment effectiveness.

There are several things to note about IVs: (a) The estimate of treatment effectiveness applies only to those patients who are marginal with respect to the treatment: Based on their attributes, these patients might or might not get the intervention, compared with patients with characteristics about which there is widespread consensus about the intervention's value (or lack of value). In particular, the 30% success rate does not apply to patients like the 60% who would always get the treatment or to the 20% who will never get the treatment. (b) The approach relies on the assumption that patients seen by the two providers are similar in both measured and unmeasured risk factors and that Providers A and B give the same quality of care. It is not possible to directly test this assumption. However, to the extent that the patients seen by each provider have similar measured risk factors and quality on measurable dimensions, one has more confidence in the IV estimate.

Future Directions

Several studies have shown that well-done observational trials result in similar estimates of treatment effectiveness as RCTs. However, the recent experience with hormone replacement therapy (HRT) in postmenopausal women suggests the need for caution. Several observational studies had indicated that HRT decreased the risk of coronary heart disease, findings which contributed to its widespread adoption by older women. However, the results of a large, long-term randomized trial subsequently found no benefit from HRT and a second study, the Women's Health Initiative, found that HRT actually increased coronary risk. This caused a rapid decline in HRT use. There is a fascinating literature on how and why the early analyses erred. In retrospect, inadequate risk adjustment for the healthier behaviors of the women who used HRT was an important factor.

The HRT experience has led some to dismiss the findings of observational studies. However, many important questions cannot be answered with randomized controlled trials. One must learn what can be learned from observational data, keeping in mind that obtaining more comprehensive measures of baseline risk factors will provide more confidence in risk-adjusted findings. Electronic medical records will facilitate the collection of important variables. As the data improve, the need for thoughtful and comprehensive risk adjustment will increase.

Michael Shwartz, Arlene S. Ash, and Lisa I. Iezzoni

See also Analysis of Covariance (ANCOVA); Bias; Calibration; Causal Inference and Diagrams; Confounding and Effect Modulation; Cox Proportional Hazards Regression; Fixed Versus Random Effects; Logistic Regression; Ordinary Least Squares Regression; Propensity Scores

Further Readings

Adjusted Clinical Groups: http://www.acg.jhsph.edu

Clinical Classification Software: http://www.hcup-us.ahrq.gov/toolssoftware/ccs/ccs.jsp

Diagnostic Cost Group or Hierarchical Condition Category models: http://www.dxcg.com

Iezzoni, L. I. (Ed.). (2003). *Risk adjustment for measuring health care outcomes* (3rd ed.). Chicago: Health Administration Press.

3M. (n.d.). *3M All patient refined diagnosis related groups software*. Retrieved January 12, 2009, from http://solutions.3m.com/wps/portal/3M/en_US/3M_Health_Information_Systems/HIS/Products/APRDRG_Software

Risk Attitude

Risk taking is generally considered to be the expression of a personality trait called *risk attitude*. People are assumed to take a certain amount of risk across a range of situations. Some people are more risk-averse because the uncertainty and potential downside of risky options makes them anxious, while others are more risk seeking because the uncertainty and potential upside of risky options excites them.

Depending on how risk attitude is assessed, the assumption that it is consistent across situations is either true or false. Risk attitude as defined within the expected utility (EU) theory framework can vary greatly across decisions in different content domains and when outcomes are described as gains or losses. Risk taking also depends on how people know about possible decision outcomes (from personal trial-and-error learning vs. being given a statistical description). Risk attitude is more stable across situations when conceptualized within a risk-benefit framework that models risk taking as a trade-off between decision makers' *perception* of the riskiness and benefit of choice options and assumes that this trade-off is governed by their *attitude toward risk*, that is, willingness to trade off (perceived) risk for (perceived) benefits: (Preference for option X) = (Perceived benefit of X) $-$ b(Perceived risk of X). Larger values of trade-off parameter b mean that the decision maker is more risk-averse. Take the (hypothetical) example of two cancer treatments, surgery and radiation, which offer different prospects of extending life. Surgery entails a 10% chance of death during the procedure but a 90% chance of extending life by 10 years. It thus has the benefit of extending life on average by 9 years (Expected value (Surgery) = .10(0) + .90(10) = 9) but also carries some risk of immediate death. This unpredictability of outcomes is modeled in finance by the statistical variance of outcomes (which also happens to be 9), but other, less analytic factors also play a role. Radiation has no danger of immediate death but offers shorter possible life extensions, with a 50% chance of living for 2 additional years and a 50% chance of living for 5 additional years. The benefit of extending life is thus lower for radiation, namely, only 3.5 years on average (Expected value (Radiation) = .50 (2) + .50 (5) = 3.5), but its variance of outcomes is also lower, namely 2.25, which factors into perceptions of risk. If expected life extension and the variance in life extensions fully described perceived benefits and risks, respectively, then a person who had a trade-off coefficient b of .5 would have a preference of 4.5 (i.e., 9 $-$ .5(9)) for surgery and one of 2.375 (i.e., 3.5 $-$ .5(2.25)) for radiation, and thus would prefer surgery.

When modeled within a risk-benefit trade-off framework, situational differences in risk taking turn out to result from different perception of risks and benefits across situations (e.g., risks are perceived to be larger or benefits smaller in some situations compared with others) rather than from differences in the trade-off coefficient b, that is, the willingness to take on (perceived) risk to obtain (perceived) benefit, which is fairly stable for a given individual.

Understanding and appropriately assessing risk attitudes is important in medical decision making. As already common practice in the financial services industry, a client's (patient's) risk attitude should be taken into consideration when providing advice about treatment options that differ in risks and benefits. It is also crucial to understand the source(s) of an undesirable level of risk taking, when designing educational interventions to change a target group's risk taking. Sources of risk taking are not limited to risk attitude, and different reasons for the behavior call for different interventions.

Recent psychological models of risky choice identify multiple determinants of risk taking, including nonlinear marginal utility (the EU definition of risk attitude) and also loss aversion, nonlinear transformations of objective probabilities into subjective decision weights (probability weighting), perceptions of returns and risk, and finally perceived risk attitude.

Situational Differences in Risk Taking

In EU theory, the dominant economic model of risk taking, the term *risk attitude* characterizes the shape of the utility function that is estimated from a person's choices. Choice of a sure outcome over a lottery with equal expected value, for example, is modeled by a concave utility function and described as risk-averse. As shown in Figure 1a, increasing amounts of money add less and less utility. Getting $500 for sure has much more utility than having a

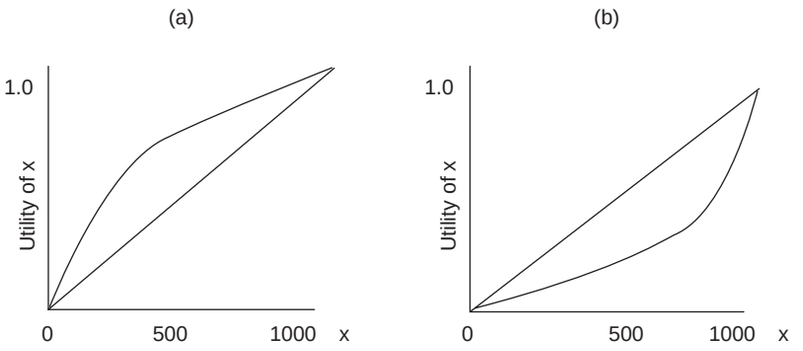


Figure 1 A concave utility function with decreasing marginal value for outcome *x* (say, dollars) on the left (a) and a convex utility function with increasing marginal value on the right (b)

50% chance of getting $1,000, otherwise $0. With a convex utility function, shown in Figure 1b, where increasing amounts of money add more and more utility, the opposite is true. The decision maker will be risk seeking and take the lottery over a sure $500, which have much lower utility than a 50% chance at $1,000, otherwise $0. Risk attitude in the EU framework is a parameter that measures the type and degree of curvature of the utility function. Even though risk attitude thus ostensibly only redescribes a set of choices, it is usually given a much more psychological interpretation.

Popular psychology and managerial folklore think of risk attitude as a personality trait, that is, as a preference for risk (either liking or not liking it) that is stable across situations. Unfortunately, there is little evidence for that, if risk attitude is defined in the EU way. A recently developed domain-specific risk-taking (DOSPERT) scale finds people's risk taking to be very different in gambling, financial investing, health decisions, recreational choices, social choices, and ethical decisions.

Psychologists Daniel Kahneman and Amos Tversky replaced the utility function of EU theory defined over total wealth with a value function defined over outcomes that are perceived as gains or losses relative to a reference point. In their model, prospect theory (PT), choices are risk-averse when outcomes are framed as gains but risk seeking when framed as losses (relative to a higher reference point). Not all apparent risk aversion is due to decreasing sensitivity on the loss and gain side. The loss function is also assumed to be steeper than the gain function, a property called *loss aversion*, which gives losses more impact (losses loom larger). Finally, PT assumes that people overweight small-probability events, which either encourages or discourages risk taking, depending on whether the rare event is desirable (winning the lottery) or not (life-threatening side effects).

Behavior in the laboratory and the real world broadly supports PT's predictions about risk aversion for perceived gains, risk seeking for perceived losses, and loss aversion. However, PT's prediction that rare events are overweighted only holds when decision makers receive a numeric or graphic description of the probability distribution of possible outcomes of all choice alternatives. Such situations have recently been called *decisions from description* and stand in contrast to *decisions from experience*, where decision makers gradually learn about available choice alternatives by personal trial and error, that is, from the feedback provided by repeated choices. All animals other than humans are restricted to this second way of learning about risks in their environment. In decisions from experience, recent outcomes carry a lot of weight, a sensible adaptation in nonstationary environments. Since rare events have a low probability of having occurred recently, they tend to get underweighted in decisions from experience, except for those rare occasions where they did recently occur, in which case they are strongly overweighted. When parents have to decide whether to inoculate their child against diseases such as the German measles, they may consult an informational brochure or Web site that provides statistical summaries of possible side effects and their (very small) probabilities, that is, they make this decision "from description" and often shy away from inoculation because they give too much weight to the possible side effects. When pediatricians contemplate the same decision, they can call to mind the

hundreds of children they have seen inoculated, with typically not a single case of negative consequences, that is, they make the decision "from experience" and may as a consequence perhaps underweight the possibility of negative side effects.

Assessment

If the purpose of the assessment is to predict risk taking in a specific future situation, risk attitude can be inferred from choices in the EU fashion, as long as the choices used in the assessment have the same characteristics as the target situation, for example, same content domain, outcome framing, and learning environment. If the purpose of the assessment is to advise or modify risk taking in a specific situation, all determinants of risk taking described above should be assessed to determine whether it is loss aversion, probability weighting, risk or benefit perception, or risk attitude that is causing the behavior one wishes to influence or change.

Elke U. Weber

See also Loss Aversion; Personality, Choices; Risk Aversion; Risk-Benefit Trade-Off, Risk Perception

Further Readings

Bromiley, P., & Curley, S. P. (1992). Individual differences in risk-taking. In J. F. Yates (Ed.), *Risk-taking behavior* (pp. 87–132). New York: Wiley.

Hertwig, R., Barron, G., Weber, E. U., & Erev, I. (2004). Decisions from experience and the effect of rare events. *Psychological Science, 15,* 534–539.

Weber, E. U., Blais, A. R., & Betz, N. (2002). A domain-specific risk-attitude scale: Measuring risk perceptions and risk behavior. *Journal of Behavioral Decision Making, 15,* 263–290.

Weber, E. U., & Johnson, E. J. (2008). Decisions under uncertainty: Psychological, economic, and neuroeconomic explanations of risk preference. In P. Glimcher, C. Camerer, E. Fehr, & R. Poldrack (Eds.), *Neuroeconomics: Decision making and the brain* (pp. 127–144). New York: Elsevier.

Risk Aversion

There are three risk attitudes (risk aversion, risk seeking, and risk neutrality). The risk attitude of "risk aversion" is distinguishable from the concept of "loss aversion." This entry discusses risk attitudes and then examines the early concepts of risk aversion in the work of mathematician Daniel Bernoulli in 1738 and the psychological research of Amos Tversky and Daniel Kahneman extending the theories in the last quarter of the 20th century.

Risk Attitudes

Risk-Averse Attitude

A risk-averse attitude is the attitude of an individual that he or she will be unwilling to accept a risk in the following situation: When presented a choice as a trade-off between a gamble and a sure thing of equal expected value, the *risk-averse* individual will be more likely to *take the sure thing* and *not take the gamble*. For example, when presented a choice between a gamble and a sure thing with the same value, for example, the choice between a 50:50 trade-off (50% chance of living an additional 1,000 days of life at the end of one's life and a 50% chance of living no more additional days of life at the end of one's life = Expected value [EV] = 500 days of life) and a sure thing (EV = 500 days of life), the risk-averse individual will choose the sure thing over the gamble.

A stronger version of risk aversion can take the following form: When presented a choice between a gamble with a higher expected value (than a sure thing) and the sure thing, the *risk-averse* individual will have a *preference for the sure thing* and *reject a gamble* even when that gamble has a higher expected value compared with the sure thing.

For example, when presented the choice between a 60:40 trade-off (60% chance of living an additional 1,000 days of life at the end of one's life and a 40% chance of living no more additional days of life at the end of one's life = EV = +600 days of life) and a sure thing (EV = +500 days of life), the risk-averse individual will still choose the sure thing over the gamble.

Risk-Seeking Attitude

A risk-seeking attitude on the part of an individual is the attitude that he or she will be willing to accept a risk in the following situation: When presented a choice as a trade-off between a gamble

and a sure thing of equal expected value, the *risk-seeking* individual will be more likely to *take the gamble*. For example, when presented a choice between a gamble and a sure thing with the same value, for example, the choice between a 50:50 trade-off (50% chance of living and losing 1,000 days of life from the end of one's life and a 50% chance of living and losing no additional days of life at the end of one's life = EV = −500 days of life) and a sure thing (EV = −500 days of life), the risk-seeking individual will choose the gamble over the sure thing.

A stronger version of risk seeking can take the following form: When presented a choice between a gamble with a lower expected value (than a sure thing) and the sure thing of a higher expected value, the *risk-seeking* individual will *reject the sure thing* and *have a preference for the gamble* even when the gamble has a lower expected value compared with the sure thing. For example, when presented the choice between a 60:40 trade-off (60% chance of losing 1,000 days of life at the end of one's life and a 40% chance of losing no days of life at the end of one's life = EV = −600 days of life) and a sure thing (EV = −500 days of life), the risk-seeking individual will still choose the gamble over the sure thing.

Risk-Neutral Attitude

A risk-neutral individual would be neutral regarding the choice, that is, he or she would be neither risk seeking nor risk-averse: When presented a choice as a trade-off between a gamble and a sure thing of equal expected value, the risk-neutral individual is more likely to be unable to choose between the two and will call the choice a toss-up.

In summary, the term *risk aversion* refers to a preference for a sure thing (certain outcome) over a gamble (risky prospect) of equal expected value. The term *risk seeking* refers to a preference for a gamble (risky prospect) over a sure thing (certain outcome) of equal expected value. According to prospect theory, people tend to be risk-averse when choosing between prospects with gains (positive outcomes) of equal expected value; people tend to be risk seeking when choosing between losses (negative outcomes) of equal expected value.

The Work of Daniel Bernoulli

Risk aversion was first worked on by Bernoulli in 1738 on a gambling dilemma (the St. Petersburg Paradox) posed to him by his cousin, Nathaniel. As part of a theoretical explanation of why gamblers are unwilling to continue betting in a sure-win game as the stakes rise in a problem, Bernoulli first formulated the concept of "risk aversion" relative to an individual's state of wealth.

In partial answer to his cousin's question, Bernoulli wrote a paper titled "Exposition of a New Theory of the Measurement of Risk." Bernoulli considered that the *value* of an item must not be based on its *price* but rather on the *utility* it yields to an individual. Bernoulli noted that the price of the item depends only on the thing itself and is equal for everyone; yet the utility of that same item depends on the particular circumstances of the person making the estimate of the worth of the same item to himself or herself. For Bernoulli, no valid measurement of the value of a risk can be obtained without consideration being given to its utility, defined as whatever gain accrues to the individual or, conversely, how much profit is required to yield a given utility. Bernoulli also suggested in his theoretical framework that people tend to evaluate outcomes in terms of their impact on the individual's resulting state of wealth.

Bernoulli's Concave Curve

Bernoulli believed that the utility resulting from a fixed small increase in wealth will be inversely proportional to the quantity of goods previously possessed. Bernoulli proposed that people have a concave utility function that captures their subjective value for money, and the preferences should be described using utility. Sarah A. Hill and William Neilson note that *diminishing sensitivity* is the property that changes in a variable have less impact the farther the variable is from a reference point.

Kahneman on Bernoulli's Perspective

Kahneman suggests that a more causal observation of the real world suggests that Bernoulli's assumption about states of wealth must be modified with consideration instead given to gains and losses. In addition, Kahneman and Tversky in their elucidation of prospect theory argued that, in

general, people tend to evaluate outcomes not in terms of their impact on an individual's resulting state of wealth but in terms of changes from a reference state with states of wealth mentioned typically only in reference to death or financial ruin.

Kahneman argues that an act of choosing may be represented as an acceptance of a gamble or the acceptance of a sure thing in a setting where a choice is made across a set of outcomes with different probabilities. For Kahneman, it is "natural" to undertake the study of decision making under risk as choices between "simple" gambles and sure things (involving, e.g., monetary outcomes) in hopes that the study of these simple problems will reveal basic attitudes of humans toward risk and value.

The Tversky-Kahneman Methodology

Tversky and Kahneman used a particular paper-pencil questionnaire methodology to extend Bernoulli's theoretical work on states of wealth. The risk attitudes—risk aversion, risk seeking, and risk neutrality—are based on a specific methodology used by Tversky and Kahneman in their examination of individual responses on paper-pencil questionnaires: (1) a decision theoretic framework in which choice between a gamble and a sure thing is the model for all decisions, where (2) choices are made (a) between simple monetary consideration involving a choice between a gamble and a sure thing with (b) objectively specified probabilities and (c) at most two nonzero outcomes.

It is important in this discussion to recognize that not all people will provide answers to questionnaires and surveys containing gambles. Some individuals may refuse to participate because they do not enter into gambles in general or they do not enter into gambles particularly related to one context, medical health issues. One may try to phrase the term *gamble* in other ways, for example, *trade-offs,* but still this group of patients may not want to participate in a survey considering gambles or trade-offs. So when considering research perspectives in the Tversky and Kahneman framework, one is examining an arena where not all humans will want to participate. And while economists may argue that these individuals in fact do enter into gambles in their life situations, they will still not complete the questionnaires or surveys and thus their responses will not be represented in the data that are typically reported in Tversky-Kahneman questionnaire or survey tasks.

One of the problems that researchers face when asking research questions in the attempt to better understand the risk attitudes is that of the expression of choice. Researchers want to attempt to provide a numerical basis for risk communication to study participants (examples of which are examined below), where the study participants may not be accustomed to communicating using such phraseology in their own lives (and may not be comfortable using the numerical framework presented in the survey questions that are asked to consider and answer). The bottom line here is that study participants asked to consider a choice between a gamble and a sure thing may not be happy with such a strong focus on numbers in the choice they are asked to consider and on which the Tversky-Kahneman framework lies.

Tversky and Kahneman's S-Shaped Curve

Tversky and Kahneman used an S-shaped curve to capture their research findings and contrasted their curve to that of Bernoulli's concave curve. Figure 1 shows an overall S-shaped function (a) somewhat concave for gains (as was Bernoulli's curve), (b) somewhat convex for losses, and (c) steeper for losses than gains.

The fact that the S-shaped curve is steeper for losses than gains (a loss of fixed amount is considered more averse as a choice option than a gain of that same fixed amount is considered appealing as a choice option) leads us into a discussion of loss aversion.

Loss Aversion

The fact that the S-shared curve is steeper for losses than gains is usually captured in the phrase *loss aversion*: Losses of a fixed amount loom larger than gains of that same fixed amount; losing a fixed amount hurts more than the pleasure derived from gaining that same fixed amount.

Loss aversion continues to be discussed and developed in the domain of marketers and consumers. Nathan Novemsky and Kahneman in 2005 described the boundaries of loss aversion in a marketing setting. Colin Camerer notes that

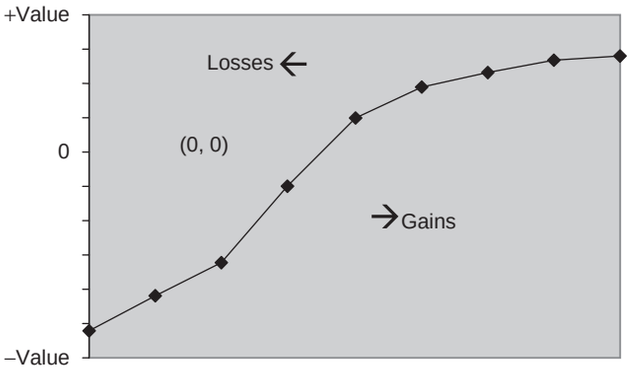


Figure 1 S-shaped curve

Novemsky and Kahneman's tour de force exploring the boundaries of loss aversion in this market setting is a reminder of both the power and delicacy of prospect theory: Prospect theory's power in explaining how people value risks and the delicacy required of its advocates to sharpen and apply the model.

Risk Attitudes and Medical Decision Making

As described earlier, the concept of a "toss-up" is where a risk-neutral individual cannot make a choice between a sure thing and a gamble with equal expected outcomes. Stephen G. Pauker and Jerome P. Kassirer argue that the toss-up is more of an atypical decision in medicine when considering a patient who is symptomatic from a medical condition or a disease process. How do the other risk attitudes (risk aversion and risk seeking) come out when they are studied in the arena of medical care when patients are asked to consider gains and losses in the domain of medical decision making?

Prospect theory—in certain of its aspects—has been tested in the domain of medicine in the setting of individuals considering hypothetical scenarios about medical conditions and disease processes.

A. B. Rosen and colleagues studied 62 study participants (mean age = 47.6 years; 47% were female, and 33% were African American). They found 37% of respondents to be decidedly risk-averse, 37% moderately risk-averse, 15% moderately risk seeking, and 11% decidedly risk seeking with increased risk aversion found in respondents of white race ($p < .01$) and lower education ($p < .05$). Women also tended to be more risk-averse ($p = .07$).

L. A. Prosser and E. Wittenberg studied differences in risk attitude across the domains of health and money for patients with multiple sclerosis ($n = 56$) and general community members ($n = 57$). Risk attitude was measured using two standard gamble questions on money and one standard gamble question on health outcomes. The authors found that risk attitude varied across domains but not by respondent type: Patients and community members were predominantly risk neutral with respect to health outcomes and risk-averse with respect to money. The authors concluded that (a) money outcomes may not be an appropriate proxy for risk preferences regarding health outcomes and (b) risk preferences may depend more on characteristics of the choice than on respondent type.

Framing and Risk Attitudes

The fact that risk attitudes are relative to whether an individual is considering a choice that is presented as a gain or as a loss brings in a new aspect of research into decision making: How does the presentation of data as gains or losses influence decision making? The influence that the presentation of data has on decision making is called *framing effects*, *presentation effects,* or *formulation effects*.

Barbara J. McNeil and colleagues investigated how variations in the way information is presented to patients influence their choices between therapeutic alternatives. The authors studied 238 ambulatory patients with different chronic medical conditions, 491 graduate students, and 424 physicians. The data presented to study participants were summaries of the results of surgery and radiation therapy for lung cancer. In this study, respondents were asked to imagine that they had lung cancer and to choose between the two therapies on the basis of both cumulative probabilities and life-expectancy data. Different groups of respondents received input data that differed only in terms of whether or not the treatments were identified and whether the outcomes were framed in terms of the probability of living or the probability of dying. The authors found that in all three populations, the attractiveness of surgery, relative to radiation therapy, was substantially greater (a) when the treatments were identified rather than unidentified, (b) when the information consisted of

life expectancy rather than cumulative probability, and (c) when the problem was framed in terms of the probability of living rather than in terms of the probability of dying. The authors suggested that an awareness of these effects among physicians and patients could help reduce bias and improve the quality of medical decision making.

Effects of Numeracy in Medical Decision Making

Two issues not explored in the above studies on risk attitudes and framing in medical contexts are the issues of literacy (the ability of individuals as study participants to work with words) and numeracy (the ability of individuals as study participants to work with numbers). Although both issues are key to risk attitudes, this section focuses on the notion of numeracy among highly educated study participants. The issue is whether (and to what extent) study participants, even highly educated study participants, have difficulty with relatively simple numeracy questions.

Isaac M. Lipkus and colleagues studied 463 men and women aged 40 and older with a three-item general and an expanded seven-item numeracy scale, assessing four numeracy skills: (1) simple mathematical operations on risk magnitudes using percentages and proportions, (2) converting percentages to proportions, (3) converting proportions to percentages, and (4) converting probabilities to proportions. The researchers found that on average, 18% and 32% of participants correctly answered all the general and expanded numeracy scale items, respectively. Approximately 16% to 20% incorrectly answered the most straightforward questions pertaining to risk magnitudes (e.g., Which represents the larger risk: 1%, 5%, or 10%?). Lipkus and colleagues conclude that if highly educated study volunteers have difficulty answering even simple numeracy questions correctly, the current methods of communicating risk numerically may be problematic and need further investigation.

Many researchers in medical decision making also have wanted to move beyond using study methods that rely on questions phrased in ways that are not replicating the actual data that is available in medical decision making. Thus, new areas of risk attitudes are being assessed within alternative study methodologies beyond the provision of summary data in terms of numbers to patients. For example, studies have been conducted using tabular and graphic data displays.

Future Research

The above considerations leads to the following question: In describing human behavior in choosing between a treatment with different benefits and risks in the short term versus the long term, how does one separate out the *external effects* that influence choice behavior in decision makers in research studies (such as framing effects) from the *internal effects* that influence choice behavior in decision makers in research studies (such as risk attitudes) from the issues of literacy and numeracy? This multipart question is one of the key questions facing future researchers in understanding how patients and physicians participate in medical decision making relevant to patients' current states of quality of life and survival.

Dennis J. Mazur

See also Decision Psychology; Human Cognitive Systems; Prospect Theory; Toss-Ups and Close Calls; Unreliability of Memory

Further Readings

Bernoulli, D. (1738). Exposition of a new theory on the measurement of risk (L. Sommer, Trans., 1954). *Econometrica, 22,* 23–26.

Hill, S. A., & Neilson, W. (2007). Inequality aversion and diminishing sensitivity. *Journal of Economic Psychology, 28,* 143–153.

Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica, 47,* 263–291.

Kahneman, D., & Tversky, A. (2007). *Choices, values, and frames*. New York: Cambridge University Press.

Littenberg, B., Partilo, S., Licata, A., & Kattan, M. W. (2003). Paper standard gamble: The reliability of a paper questionnaire to assess utility. *Medical Decision Making, 23,* 480–488.

Mazur, D. J., & Hickam, D. H. (1990). Treatment preferences of patients and physicians: Influences of summary data when framing effects are controlled. *Medical Decision Making, 10,* 2–5.

Mazur, D. J., & Hickam, D. H. (1993). Patients' and physicians' interpretations of graphic data displays. *Medical Decision Making, 13,* 59–63.

McNeil, B. J., Pauker, S. G., Sox, H. C., Jr., & Tversky, A. (1982). On the elicitation of preferences for alternative therapies. *New England Journal of Medicine, 306,* 1259–1262.

Rabin, M., & Thaler, R. H. (2001). Anomalies: Risk aversion. *Journal of Economic Perspectives, 15,* 219–232.

Starmer, C. (2000). Developments in non-expected utility theory: The hunt for a descriptive theory of choice under risk. *Journal of Economic Literature, 38,* 332–382.

Tversky, A., & Kahneman, D. (1991). Loss aversion in riskless choice: A reference-dependent model. *Quarterly Journal of Economics, 106,* 1039–1061.

Risk-Benefit Trade-Off

Risk-benefit trade-off refers to the balance of negative and positive effects on achieving a goal, such as health. For medical decisions, a risk-benefit trade-off usually refers to the perception of the anticipated balance of improvements and deteriorations in health from a given choice. For patients, caregivers, and policy makers, this can range from the balance of health in an individual to the overall balance of health experienced by a society. How trade-offs are considered is highly subjective. A risk-benefit trade-off can also consider goals outside of health.

Estimating Future Risk

Medical decisions allow for choices that can affect health. Risk can be defined as the extent to which deteriorations in health are perceived by a patient. Some medical scientists have suggested that risk be more appropriately labeled *harm,* since this is the direct opposite of benefit. Similarly, risk is sometimes inappropriately described as *safety,* which is a term used to describe the extent to which harm is absent. Risk is a term that can also refer to the chance of experiencing clinical measures of disease (e.g., disease prognosis such as risk of heart attack). In the context of a risk-benefit trade-off, risk usually refers to the harms experienced by a patient that are directly associated with the decision. They are synonymous with the adverse effects, or side effects, from a medical decision.

Absolute Harm

The chance of experiencing a side effect generally does not change according to an individual's risk of future disease. As such, medical decisions are said to be associated with an absolute risk of harm. The extent to which outcomes attributed to harm are experienced by patients may be affected by other factors, such as age, gender, presence of other diseases, or genetics. Because these adverse effects can occur rarely and are sometimes unexpected, the ability for a decision maker to factor them into a decision may be limited.

Patient Value

It is important to note that harm is subjective and is based on the perceived value of the adverse effects from a medical decision. If a patient believes that an increased chance of experiencing stomach upset is detrimental to health, then he or she would attribute a risk from the medical decision. The extent to which an outcome is perceived to deteriorate health would then equate with the perceived risk from a medical decision. It cannot be assumed that different patients would assign the same amount of risk to the same occurrence of these outcomes. Hence, risk from health effects will be perceived differently by different patients.

Multiple Consequences

Since interventions can increase the chance of more than one type of side effect and these occurrences would normally be perceived as negative, the overall perceived risk from a medical decision may need to consider the range of outcomes relevant to the patient. For example, a medical decision may involve an option that leads to a high chance of stroke and a low chance of stomach upset versus a different option that leads to a high chance of paralysis and a low chance of joint pain. These two side-effect profiles may be perceived as having similar risk to patients.

Estimating Future Benefit

Benefit can be defined as the extent to which improvements in health are perceived by a patient. Reduced risk of disease, represented by specific outcome measures such as chance of dying, chance

of heart attack, or average degree of pain relief can be used by patients and their caregivers as a means of estimating future benefit. For example, if a certain form of cancer leads to an outcome of sudden death in 3 of 100 individuals and surgery is thought to reduce the death rate by one third, then the expected consequence for any 100 individuals with this cancer who decide to receive surgery would be to avoid certain death 1 time out of 100. That is, one third of the three people who would expect to experience the outcome of sudden death would no longer experience this outcome, leading to two sudden deaths and one death averted.

Relative Benefit

The previous example illustrates that patient health outcomes are from a medical decision and are directly related to the chance of experiencing a future disease outcome (e.g., risk of sudden death) and the effect of a decision on that outcome. It is generally assumed that health interventions affect the risk of these outcomes occurring relative to the natural risk of experiencing the outcomes. This relative effect on outcomes is referred to as the *relative risk reduction*.

For example, if 100 individuals were believed to have a 30 in 100 chance rather than a 3 in 100 chance of dying of the same cancer, the intervention would still be believed to reduce sudden death by one third, leading to 10 in 100 deaths averted. Hence, 100 individuals receiving treatment would have a 1 in 10 chance of averting death, rather than a 1 in 100 chance. It should be noted that it is not always the case that a health intervention will affect undesirable outcomes in a relative fashion in different individuals with varying prognoses.

Similarity to Risk

It is important to note that, like risk, benefit is subjective and is based on the perceived value of the extent to which health is improved, which can be estimated by the occurrence of the clinical measures that characterize disease by patients and caregivers. In the previous example, if a patient perceives a reduction in the chance of dying as more valuable than an increase in the chance of dying, then he or she would be said to attribute a benefit to the intervention. Similarly, health effects will be perceived differently by different patients, and multiple outcomes that contribute to a perception of benefit may be considered.

Trade-Offs

For medical decisions, a risk-benefit trade-off usually refers to the perception of the anticipated balance of improvements and deteriorations in health from a given choice.

The trade-off has also been referred to as a *risk-benefit ratio*. Some medical scientists have suggested that because medical measures of benefit and harm are so different, that the term *risk-benefit ratio* has no literal meaning, since it cannot be calculated. In contrast, other medical scientists have used questionnaires of hypothetical scenarios, a technique called a *stated preference approach,* to understand the mathematical relationship between factors considered and their relative importance when making a medical decision. These methods are often synonymous with risk-benefit analysis, also called benefit-risk analysis, that attempts to characterize individual preferences for improvements and deteriorations in health. Risk-benefit analysis can be used by patients, caregivers, and policy makers, including market regulators, to clearly understand patient preferences and as an aid in decision making.

Single measures of disease burden that encapsulate preferences for different health states based in von Neumann and Morgenstern (NM) utility theory have also been developed. Examples of these preference-based measures include quality-adjusted life years and healthy-year equivalents. While some have suggested that these measures fail to adequately embrace preferences and preference theory, others have suggested that stated preference approaches, particularly those that examine monetary preferences such as willingness to pay, have several methodological and ethical issues that need to be resolved. More recently, hybrid approaches that are more grounded in NM utility theory have been developed.

Although risk-benefit trade-off implies that benefit is directly weighed against risk, medical decisions can involve multiple trade-offs between health outcomes attributed to benefit and risk.

This includes comparisons between benefits, between risks, between benefits and risks, or between risk-benefit trade-offs. The types of trade-offs considered and how they are compared will vary with who is making the decision. They are also highly susceptible to many known cognitive biases. A risk-benefit trade-off can also consider goals outside of health.

Perspective

When considering health trade-offs, decision makers may consider either the health of an individual or the health of a population. For example, an individual decision regarding which antibacterial to use for an infection may consider outcome measures such as the chance of relieving the infection, average time to relief of symptoms, and frequency of serious side effects. These outcome measures could contribute to a patient or caregiver's understanding of the trade-off of risk and benefit and choices that optimize the health of the infected individual. Frequently, outcome measures such as the chance of contributing to antibiotic resistance are considered when making a decision regarding antibiotics. Although these outcomes may have no direct impact on an individual's health, they can be detrimental to the future health of society, as they will lead to a situation where the antibiotic is less effective. Hence, the current risk-benefit trade-off is weighed against a future risk-benefit trade-off. Considering the value of future information on outcome measures can also be applied to any decision, and it allows decision makers to consider the effect future information will have on future risk-benefit trade-offs versus a current risk-benefit assessment.

Net Health Benefit

Healthcare third-party payers with limited resources are confronted with these similar dilemmas. That is, at any given time, the resources required to produce a given risk-benefit trade-off in one population could be used to produce a more (or less) desirable risk-benefit trade-off in a different population. Or more simply, producing health in one population may forego an opportunity to produce more health in another population. Comparing risk-benefit trade-offs with current or future risk-benefit trade-offs is a utilitarian approach, as it considers the extent to which decisions lead to the greatest amount of health for the greatest number of people.

Cognitive Bias

Human beings are susceptible to a large number of cognitive biases that make the valid identification and comparison of factors related to risk and benefit difficult. Examples of these include failure to understand probability, ignoring small differences, underestimating harm, overestimating benefit, focusing on too few attributes, or judging harmful actions worse than harmful inactions. How risk and benefit are communicated can also influence decisions. Because emotion and poor mental accounting can lead to suboptimal or irrational decisions, the field of decision analysis and its accompanying methods and tools have been developed to characterize decision problems and provide insight to decision makers for improving decisions.

Goals Outside of Health

Although risk-benefit trade-offs generally refer to a balance of factors, which can be attributed to health, other factors may enter a decision, which may be interpreted as risks or benefits. Factors such as non-health-related quality of life, equity, ethical issues such as justice, psychosocial issues such as convenience of use, and legal issues such as risk of litigation are some examples that a decision maker may attribute to benefits or risks from a decision. Although effects on resources available for healthcare are often seen as a factor outside of health, they should not be considered so, as they are directly traded for health.

Don Husereau

See also Biases in Human Prediction; Complications or Adverse Effects of Treatment; Conjoint Analysis; Contingent Valuation; Decisions Faced by Patients: Primary Care; Discrete Choice; Emotion and Choice; Evaluating Consequences; Gain/Loss Framing Effects; Heuristics; Judgment; Mental Accounting; Net Benefit Regression; Risk Aversion; Risk Perception; Utility Assessment Techniques; Value Functions in Domains of Gains and Losses

Further Readings

Baron, J. (2000). *Thinking and deciding*. New York: Cambridge University Press.

Bridges, J. F. (2003). Stated preference methods in health care evaluation: An emerging methodological paradigm in health economics. *Applied Health Economics and Health Policy, 2,* 213–224.

Drummond, M. (2005). *Methods for the economic evaluation of health programmes* (3rd ed.). Oxford, UK: Oxford University Press.

Glasziou, P. P. (1995). An evidence-based approach to individualising treatment. *British Medical Journal, 311,* 1356–1359.

Ioannidis, J. P. A. (2004). Better reporting of harms in randomized trials: An extension of the CONSORT statement. *Annals of Internal Medicine, 141,* 781–788.

Laupacis, A. (1988). An assessment of clinically useful measures of the consequences of treatment. *New England Journal of Medicine, 318*(26), 1728–1733.

Risk Communication

In a world that is fundamentally uncertain, society needs to be prepared to deal with risks and uncertainty in a proper way. However, often this is not the case, and the psychological consequences of misperceiving risks can have severe, physical consequences. First, this entry illustrates why this issue is important. Then, typical misunderstandings that happen in risk communication are explained, as well as how these misunderstandings can be avoided and insight reached.

Example

In October 1995, the U.K. Committee on the Safety of Medicines issued a warning that third-generation oral contraceptive pills containing desogestrel or gestodene increased the risk of venous thromboembolism by 100%. That is, the risk was twofold. This information was passed on in 190,000 letters to general practitioners, pharmacists, and directors of public health and also forwarded to the media. In response, many women decided not to take the pill anymore.

In the following year, the number of abortions in the United Kingdom increased by almost 9%, which makes a total of 13,600 additional abortions, against the decreasing trend in abortions in the previous years. This number is particularly interesting in comparison with the increase in conceptions, which was only 3.3%, a total of 26,000 additional conceptions. That is, the number of additional abortions amounts to more than half of the number of additional conceptions, which at least suggests that out of the additional conceptions particularly many were unwanted. Moreover, the increase both in conceptions and in abortions was particularly pronounced in teenagers. The resulting additional costs for abortion provision to the National Health Service have been estimated to be about £4 to £6 million.

A closer look at the twofold risk of thromboembolism reveals that it approximately means that the risk of thromboembolism increases from 3 in 20,000 women who take second-generation oral contraceptive pills (i.e., those containing levonorgestrel or norethisterone) to 6 in 20,000 women who take third-generation oral contraceptive pills, while the baseline risk of women who do not take oral contraceptive pills is about 2 in 20,000. That is, the relative risk increase is indeed 100%, but in absolute numbers, this means a risk increase of only 3 in 20,000. Additionally, it needs to be noted that pregnancy increases the risk to 12 in 20,000, which is again twice as high compared with taking third-generation oral contraceptive pills. Had women known these numbers, many unwanted pregnancies and subsequent abortions may have been avoided.

Risk Illiteracy

This example illustrates a larger societal problem. Many citizens are not prepared to deal rationally with risks and uncertainties. This problem is particular in that it is one of those that are not recognized as such in the public, although it may cost lives, cause abortions, or just psychological pain. Such a pill scare will likely happen again, as others did before, and people may not be prepared to react with reason, since many are statistically illiterate in the sense that they do not know about the distinction between a relative risk (100%) and an absolute risk (3 in 20,000).

It has been debated whether risk illiteracy is mainly a consequence of cognitive limitations, as

suggested by the extensive literature on risk perception. However, such an internal attribution of the causes has not led to successful treatment. If "probability blindness" were caused by our cognitive limitations, then we just would have to live with it, or, as some have suggested, to keep citizens away from important decisions. In contrast to this view, there are numerous examples showing that risk innumeracy is largely a function of the external representations used in risk communication.

In particular, there are three common representations, *relative risks*, *single-event probabilities*, and *conditional probabilities,* which may be confusing.

Relative Risks

The increased risk of venous thromboembolism by third-generation oral contraceptive pills put forward as a twofold risk, or an increase of 100%, is a relative risk. As explained before, the 100% means, an absolute risk increase from 3 to 6 in 20,000.

The problem with relative risks is that they are silent about the base rate risk. That is, the risk increase would be 100% independent of whether the increase is from 3 to 6 in 20,000 or from 3,000 to 6,000 in 20,000. However, most would agree that the societal importance of the latter risk increase would be much larger than that of the former (which matches that of third-generation pills). Relative risks thus can be used to make risks loom larger than they actually are. This similarly holds for risk reductions. In the pill example, one could argue that women who switch from third-generation pills back to second-generation pills reduce their risk of venous thromboembolism by 50%, namely, from 6 to 3 in 20,000.

However, instead of using the number of diseases as a reference class, one could also use the number of healthy women (i.e., without thromboembolism) as a reference class, and thereby make the relative risk reduction look small. Namely, instead of 19,994 in 20,000 women taking third-generation pills who are healthy, there would be 19,997 in 20,000 women with second-generation pills. The absolute increase in healthy women is again 3 in 20,000, but in relative numbers, the increase in healthy women is only .015%. Thus, a risk reduction by 50% can mean the same thing as an increase in healthy women by .015%. In absolute terms, it becomes transparent that the difference is 3 in 20,000 in both cases.

Not only are laypeople often confused about relative risks, but experts are as well. For example, decisions by health authorities on which treatment to fund have been shown to be largely affected by the representation format: Rehabilitation and screening programs were evaluated much more positively if their benefits were described in terms of relative risk reductions.

Single-Event Probabilities

An everyday life example of single-event probabilities can often be heard in the daily news when the speaker indicates the chance of rain for the next day. A statement such as that the chance of rain tomorrow is 30% remains unclear to many. In the end, it can only rain or not. The problem is that it is unclear to what the 30% refers to, that is, the *reference class* is missing. Some people believe that there will be rain in 30% of the area, others think that it is 30% of the time. The right interpretation, however, is that out of 100 days that are exactly like tomorrow, it will rain in 30 of them.

In medical contexts, single-event probabilities are often used to communicate the risks of a treatment, such as side effects. A psychiatrist often prescribed fluotexine (Prozac) to patients with mild depression and told them that the risk of having sexual problems (e.g., impotence or loss of sexual interest) as a side effect was 30% to 50%. Many of his or her patients were anxious hearing those numbers, because they interpreted them as meaning that *every* patient would have problems in about 30% to 50% of their sexual encounters. However, the numbers actually mean that out of 100 patients 30 to 50 will experience a sexual problem. Hearing this interpretation, patients were much less afraid of taking Prozac. This example illustrates again a reference class problem: While the patients had their own sexual encounters in mind as a reference class, the doctor was referring to patients as a reference class.

Therefore, the solution to such misunderstandings is obvious: clearly indicating a reference class (e.g., sexual problems will occur in 30% to 50% of patients) or using a frequentist formulation

(e.g., out of 100 patients, 30 to 50 will experience a sexual problem).

Conditional Probabilities

The chance of detecting a disease with a medical test is usually communicated as a conditional probability, namely, the sensitivity of the test: "If a woman actually has breast cancer, the chance of getting a positive result in a mammography is 90%." That is, it is the probability of testing positive *given* breast cancer. However, this is often confused with the positive predictive value of the test, the probability of having breast cancer *given* a positive test result, which is not the same. This can be illustrated with a more intuitive example. Up to 2008, every American president was male. That is, the probability of being male given that one is president of the United States was 100%. The reverse, obviously, does not hold: Given that one is male, chances of being or becoming president of the United States are still rather low.

The question is how to get from the sensitivity of the test to the positive predictive value, which is the information one really needs. Two further pieces of information are necessary. First, one needs to know the base rate of the disease; here, this is about 0.8%. Second, one needs to know the false-positive rate of the test, that is, the probability of getting a positive test result *given* that one is actually healthy, which is about 7% in this case. Formally, the sensitivity, the base rate, and the false-positive rate can be combined to calculate the positive predictive value by applying Bayes's theorem. However, both experts and laypeople often have trouble with Bayes's theorem, and it is much simpler to think about such problems in terms of natural frequencies.

That is, instead of combining conditional probabilities, imagine 1,000 women. Out of these, 8 (= .8% base rate) are expected to have breast cancer; the remaining 992 are expected not to have breast cancer. Out of the 8 women with breast cancer, about 7 (= 90% sensitivity) will test positive. Out of the remaining 992 women without breast cancer, about 69 (= 7% false positives) will also test positive. That is, there are 76 women who test positive, out of which only 7 actually do have the disease. Therefore, the probability of a woman to have breast cancer given a positive test, the positive predictive value is 7 out of 76, which is approximately 9%.

Again, being confused by conditional probabilities is not only a problem of laypeople but also of experts. Only a very small proportion of physicians who were given numbers as conditional probabilities actually combined them correctly to figure out the positive predictive value. The error that was most often observed was that the positive predictive value was confused with the sensitivity, which often resulted in overestimating the predictive power of the test (here, 90% instead of 9%). Sometimes, the false-positive rate was subtracted from the sensitivity, which still led to an overestimated predictive power (here, 83% vs. 9%). When doctors were given the same test properties in natural frequencies, they were much more likely to give the correct answer. Also, training in how to translate conditional probabilities into natural frequencies has long-lasting positive effects on the accuracy of such calculations, while training with Bayes's theorem does not seem to be very helpful.

Implications

People have to deal with risks and uncertainties every day, in particular in the medical domain. Yet the ideals of informed consent and shared decision making will not be entirely realized until medical evidence is properly understood. Appropriate risk communication is thus a necessary step toward this goal.

Wolfgang Gaissmaier and Gerd Gigerenzer

See also Bayes's Theorem; Informed Consent; Numeracy; Risk Perception; Shared Decision Making

Further Readings

Furedi, A. (1999). The public health implications of the 1995 "pill scare." *Human Reproduction Update, 5,* 621–626.

Gigerenzer, G. (2002). *Calculated risks: How to know when numbers deceive you.* New York: Simon & Schuster.

Gigerenzer, G., Gaissmaier, W., Kurz-Milcke, E., Schwartz, L. M., & Woloshin, S. (2007). Helping doctors and patients make sense of health statistics. *Psychological Science in the Public Interest, 8,* 53–95.

Gigerenzer, G., & Hoffrage, U. (1995). How to improve Bayesian reasoning without instruction: Frequency formats. *Psychological Review, 102,* 684–704.

Risk Neutrality

See Risk Aversion

Risk Perception

Risk perception refers to people's subjective impression of riskiness. Objective ways of defining risk involve the uncertainty of outcomes, their negativity, or both ("the risk of dying during surgery"). Subjective perception of risk (related to health, safety, or financial outcomes) is determined not only by these readily quantifiable outcome dimensions but also by more qualitative characteristics such as decision makers' familiarity with choice options, perceived control, and institutional trust.

Understanding risk perception matters in the context of medical decision making for multiple reasons. Informed consent requires an accurate understanding of risks and benefits. The design of risk communication messages needs to consider people's processing of risk information to ensure that risks are understood in their correct magnitude. Since policy makers and physicians often differ from the general public in familiarity with choice options, perceived control, and institutional trust, they may perceive risks in different ways even when all sides agree on associated mortality and morbidity. Understanding the source of individual or group differences in risk perception can thus explain and perhaps align disagreements about risk. Risk perceptions are also important because they influence people's choices and actions, which are often made by informal risk-benefit trade-offs.

Two pathways of risk perception are described below: risk as a statistic versus risk as a feeling. Behavioral economics and psychology have identified which type of statistical summaries and which feelings best predict people's subjective assessment of risk.

Two Pathways

Expert quantifications of morbidity or mortality risks posed by medical conditions or procedures are based on objective data and/or theoretical models. The general public can evaluate the same options very differently, sometimes with serious consequences. Public perception of an autoimmune disease risk of silicone breast implants, for example, resulted in bankruptcy for the manufacturer, despite no scientific evidence of implant-related illnesses. Neuroscience and behavioral research show that risk perception is determined by both analytic and emotional processes. The relative balance of these two pathways is affected by professional training and cognitive capacity. While emotional processes are hardwired and automatic, analytic evaluations are effortful and need to be learned.

Risk as a Statistic

Actions that result in a guaranteed, known outcome are typically described as riskless. The more disparate the range of possible outcomes, the more risky the action. This greater sense of unpredictability of outcomes is captured by the variance as a statistical index of degree of risk:

$$\text{Variance}(X) = \sum_x (x - \text{EV}(x))^2 p(x),$$

where x denotes the possible outcomes, $p(x)$ the probability that each outcome will occur, and $\text{EV}(x)$ the expected value of these outcomes.

Take the hypothetical example of two cancer treatments, surgery and radiation, which offer different prospects of extending life. Surgery entails a 10% chance of death during the procedure but a 90% chance of extending life by 10 years. The average or expected life extension of surgery therefore is $\text{EV(Surgery)} = .10(0) + .90(10) = 0 + 9 = 9$, and its variance is $\text{Variance(Surgery)} = (0 - 9)^2(.1) + (10 - 9)^2(.9) = 81(.1) + 1(.9) = 8.1 + .9 = 9$. Radiation has no chance of immediate death but offers shorter life extensions. Half of the patients live for 2 additional years, and the other half for 5 additional years, so the expected life extension is $\text{EV(Radiation)} = .50(2) + .50(5) = 1 + 2.5 = 3.5$, and its variance is $\text{Variance(Radiation)} = (2 - 3.5)^2(.5) + (5 - 3.5)^2(.5) = 2.25(.5) + 2.25(.5) = 2.25$. While radiation offers a shorter average life extension than surgery (3.5 years as opposed to 9 years), it also has a smaller variance (2.25 vs. 9), that is, less uncertainty or risk about the outcome.

The variance (or its square root, the standard deviation [*SD*], i.e., the average deviation of possible outcomes from their mean) is the most common index of risk in financial contexts, partly for its mathematical properties. However, studies of people's judgments of the riskiness of financial gambles show the variance to be a poor index of perceived risk. While downside variability affects perceived risk of an option much more than upside variability, this asymmetry is not captured by the variance. Other indices, such as conjoint expected risk (CER), allow for downside variability to have greater impact and capture both *similarities* in risk judgments (with a common way in which outcome probabilities and values are combined) and individual or group *differences* (with model parameters that capture differences in the relative weight of model components). Behavioral research also shows that the riskiness implied by variability is perceived relative to average returns. A standard deviation of ±$100 is seen as huge when the mean return is $50 but amounts to rounding error for a risky option with a mean return of $1 million. The coefficient of variation (CV),

$$\mathrm{CV}(X) = SD(X)/\mathrm{EV}(X),$$

provides a relative measure of risk, by dividing the standard deviation (*SD*) by average return (EV). A risky financial investment with an average deviation of *SD* = $100 and an average return of EV = $50 has a CV of 100/50 = 2, whereas a risky financial investment with the same average deviation of *SD* = $100 but an average return of EV = $1 million has a CV of 100/1,000,000 = .0001, a much smaller value. The CV describes perceived risk far better and is widely used as a statistical index of risk in many applied areas.

Risk as a Feeling

Psychological research has shown that the risk perceptions by the lay public have often little to do with morbidity or mortality statistics. People overweight risk associated with infrequent, catastrophic, and involuntary events, and underweight risk associated with frequent, familiar, and voluntary events. These deviations between objective statistics and subjective impression are mediated by two factors that elicit anxiety, which is interpreted as risk. The first factor, *dread,* relates to perceived lack of control and catastrophic potential. Flying is seen as a riskier mode of transportation than driving, because plane crashes are outside one's control and cause a large number of fatalities. The second factor, *risk of the unknown,* is determined by the extent to which a hazard is unobservable or unknown. Radiation is seen as risky because it can kill without the victim being aware of any exposure. Risk perception has been described as a collective process. Cultural theory distinguishes groups by their patterns of interpersonal relationships, which are seen as affecting perceptions of risk. More hierarchical groups perceive industrial and technological innovations as opportunities, whereas more egalitarian groups perceive them as threats to their social structure. Culture is seen as teaching its members where their interests lie and what events pose risks to their way of life. Cultural differences in institutional trust affect perceived risk. Minority group members have less trust in social institutions, see themselves as having less control over their life, and perceive greater risks. The relationship between trust and risk perception is also mediated by an emotional pathway, with reduced trust resulting in stronger negative affective responses to potential hazards and increased perceptions of risk.

Elke U. Weber

See also Dual-Process Theory; Emotion and Choice; Intuition Versus Analysis; Risk Attitude

Further Readings

Bontempo, R. N., Bottom, W. P., & Weber, E. U. (1997). Cross-cultural differences in risk perception: A model-based approach. *Risk Analysis, 17,* 479–488.

Loewenstein, G. F., Weber, E. U., Hsee, C. K., & Welch, E. (2000). Risk as feelings. *Psychological Bulletin, 127,* 267–286.

Slovic, P. (1997). Trust, emotion, sex, politics, and science: Surveying the risk-assessment battlefield. In M. Bazerman, D. Messick, A. Tenbrunsel, & K. Wade-Benzoni (Eds.), *Psychological perspectives to environmental and ethical issues in management* (pp. 277–313). San Francisco: Jossey-Bass.

Weber, E. U., & Hsee, C. K. (1999). Models and mosaics: Investigating cross-cultural differences in risk

perception and risk preference. *Psychonomic Bulletin & Review, 6,* 611–617.
Yates, J. F., & Stone, E. R. (1992). Risk appraisal. In J. F. Yates (Ed.), *Risk-taking behavior* (pp. 49–85). Chichester, UK: Wiley.

Risk Seeking

See Risk Aversion

S

Sample Size and Power

Often the main aim of a medical study is to estimate or test an unknown parameter of interest, such as the incidence of a certain disease, the effect of a certain treatment, or the relative risk associated with a certain exposure. An important issue in the design of a study is the choice of the number of subjects to include. The larger the study sample, the more precise a parameter estimate will be, and the choice of the sample size will depend on how much precision is required. If the aim of the study is to demonstrate that a certain treatment is effective, the power of the study is very important. The power is the chance that the study, through a statistically significant treatment effect, will prove that the treatment is effective, if the treatment really is effective. The sample size should be chosen large enough to have sufficient power in case the real treatment effect is clinically relevant. Small studies can be inadequate because of too small power or too low precision. On the other hand, very large studies can have more precision than really needed or have high power even against treatment effects that are too small to be clinically relevant, leading to a waste of money. Therefore, the choice of the sample size should be well balanced. Sample size and power are strongly related; the larger the sample size, the higher the power. This relationship is quantified by what are called sample size formulae. Below, we give the ones that are most important in practice. These formulae can be very helpful in determining the sample size of a study. A number of different cases are discussed below.

In sample size formulae, there are always one or two z scores involved. One of them regards the significance level of the statistical test or the confidence level of the confidence interval, and the other one regards the power of the test. These z scores are associated with the standard normal distribution, which is probably the most important distribution in statistics. It has mean 0 and standard deviation 1 and is graphically represented by the well-known bell-shaped curve. The total area under the curve is equal to 1. A z score z_β is defined as the point in the distribution such that on the right of z_β, the area under the curve is equal to β, and therefore the area on the left-hand side of z_β is $1 - \beta$. Consequently, $z_{\alpha/2}$ is the value such that the area on the right-hand side is equal to $\alpha/2$. A table giving the value of z_β for different values of β or vice versa can be found in any introductory statistics book.

Estimation of a Proportion

Suppose the aim of a study is to estimate a certain unknown proportion π, for instance, the prevalence of a certain disease in a certain population. Once a sample is drawn, π can be estimated by the sample proportion p, being the number of subjects with the disease in the sample divided by the sample size n. An approximate 95% confidence interval for π is given by

$$p - 1.96\sqrt{\frac{p(1-p)}{n}} < \pi < p + 1.96\sqrt{\frac{p(1-p)}{n}}.$$

This means that with 95% probability, the difference $|p - \pi|$ between the estimate and the true parameter value is smaller than $1.96\sqrt{p(1-p)/n}$. Suppose one wishes to have the sample size n sufficiently large to get an estimate p that has an error $|p - \pi|$ not larger than some chosen margin of precision δ, with high probability, say 95%. Then n should be chosen so large that $1.96\sqrt{p(1-p)/n} < \delta$, which implies that n should be larger than $\left(1.96\sqrt{p(1-p)}/\delta\right)^2$. However, before the study, p is not yet known, thus an estimate p^* of π is needed—based on external data sources or just an informed guess based on expert opinion. In general, to estimate a proportion π such that the estimation error is smaller than δ with probability $1 - \alpha$, the sample size should fulfill

$$n > \left(z_{\alpha/2}\frac{\sqrt{p^*(1-p^*)}}{\delta}\right)^2,$$

where $z_{\alpha/2}$ is as defined above.

Example

Suppose one wishes to estimate the incidence π of personality change after a coronary bypass operation. The investigator anticipates that the incidence will be in the order of magnitude of .30, and wants to have an estimation error not larger than .02 with probability 90%. Thus, δ = .02, p^* = .30, $1 - \alpha$ = .90, and $z_{0.05}$ = 1.645. Then the sample size should be at least $n = 1.645^2 \times .30((1 - .30)/.02^2 = 1421$.

Estimation of a Mean

If the aim of a study is to estimate the population mean μ of a variable, the sample size formula is

$$n > \left(z_{\alpha/2}\frac{\sigma}{\delta}\right)^2.$$

Here, σ is an a priori estimate of the population standard deviation. Notice that in essence this is the same formula as in the case of estimating a proportion, since the standard deviation of a dichotomous variable is $\sqrt{\pi(1-\pi)}$.

Example

Suppose a medical investigator wants to estimate the mean diastolic blood pressure in a population of 20-year-old men with an error that is with probability 95% less than 2 mmHg. Suppose it is known that the standard deviation of diastolic blood pressure in young male populations is in the order of magnitude of 8 mmHg. Then the number of 20-year-old men to be included in the sample should be at least

$$n = \left(1.96 \times \frac{8}{2}\right)^2 = 62.$$

Testing Equality of Two Means

Suppose the aim of a study is to compare the mean of a continuous outcome variable between two populations. More specifically, the aim is to test the null hypothesis H_0, $\mu_1 = \mu_2$, where μ_1 and μ_2 stand for the two population means. For instance, μ_1 is the mean decrease in body weight after following the Atkins diet for 3 months, and μ_2 is the mean decrease in weight after a 3-month standard control diet. Suppose the investigators think that a minimal clinically worthwhile difference in decrease in body weight should be at least .75 kg in favor of the Atkins diet. In general, the minimal clinically relevant difference is denoted by δ. If the true difference is δ or more, the power, which is the probability that H_0 is rejected, is desired to be large. The power is often denoted by $1 - \beta$, where β is the probability of not rejecting H_0 when the real difference is δ. Typical choices for β in practice are .20 or .10, corresponding with 80% or 90% power. The sample size needed to have power $1 - \beta$ against a true difference δ is given by the following formula

$$n_{\text{per group}} = \frac{(z_{\alpha/2} + z_\beta)^2 2\sigma^2}{\delta^2}.$$

Here, α is the significance level of the test, in practice almost always set at .05; thus, $z_{\alpha/2}$ = 1.96; σ is an a priori guess of the standard deviation of the outcome variable, assumed to be equal in the two populations.

Typical choices for β and the corresponding values for z_β are as follows:

β	.01	.05	.10	.15	.20	.25	.30
z_β	2.33	1.65	1.28	1.04	.84	.67	.53

Example

Suppose the test is done at the usual 5% level; thus, $z_{\alpha/2} = 1.96$. Based on the literature, the investigators anticipate that the standard deviation of the decrease in weight is about 1.5 kg. If a power of 90% is required against a difference $\delta = .75$, the number of subjects per group to be included into the study is $n = (1.96 + 1.28)^2 \times 1.7^2 / .75^2 = 60$. Thus, in total, 120 subjects are needed.

Testing Equality of Two Proportions

Suppose the aim of a study is to test the equality of two proportions. For instance, π_E is the incidence of some outcome event under an experimental treatment, and π_C is the incidence under a standard control treatment. Then, the null hypothesis to test is H_0: $\pi_E = \pi_C$. The sample size needed to have power $1 - \beta$ against a treatment effect δ is

$$n_{\text{per group}} = (z_{\alpha/2} + z_\beta)^2 \frac{\pi_C(1-\pi_C) + \pi_E(1-\pi_E)}{\delta^2}.$$

Here, π_C is an a priori guess of the incidence of the event in the control group, and $\pi_E = \pi_C + \delta$.

Example

Suppose a randomized clinical trial is designed to test the effect of a certain cholesterol-lowering drug against placebo. The outcome is the occurrence of a myocardial infarction within a follow-up period of 5 years. From the literature, it is known that $\pi_C \approx .04$. A 25% decrease of this incidence under the experimental treatment is deemed clinically relevant and realistic; thus, δ is chosen to be .01. The significance level of the test is set at the usual $\alpha = .05$, thus $z_\alpha = 1.96$; and the investigators want to have a power of 85%, thus $z_\beta = 1.04$. The number of subjects needed per treatment group is

$$(1.96+1.04)^2 \frac{.04(1-.04)+.03(1-.03)}{.01^2} = 6075.$$

Thus, in total 12,150 subjects have to be included in the study.

Comparing Time-to-Event Outcome Between Two Groups

Suppose a study is designed to compare two differently treated groups of patients who are followed for the outcome of some event of interest. The duration of follow-up may vary between patients. Mostly the treatments are compared through the hazard ratio θ, a parameter that is defined such that at each short time interval the probability of the event in the treatment group is θ times the probability of the event in the control group. Thus, $\theta = 1$ means that the chances of the event are identical in the two groups. The null hypothesis of no treatment effect, H_0: $\theta = 1$, is tested with the log-rank test. In such a situation the power of the study depends on the number of observed events rather than on the number of patients. To have power $1 - \beta$ against a hazard ratio θ, the number of events (d) needed to be observed is

$$(1.96+1.04)^2 \frac{.04(1-.04)+.03(1-.03)}{.01^2} = 6075.$$

Next, based on an a priori guess of the incidence in one of the groups, the number of subjects to be included and the length of the follow-up can be determined.

Example

Suppose a clinical trial is designed to compare the incidence of myocardial infarction of subjects receiving a certain cholesterol-lowering drug with subjects using placebo. A reduction of 25% in incidence of the outcome event is deemed clinically relevant by the investigators, and a power of 80% is wished. According to the above formula, the number of events required to have 80% power against a hazard ratio of .75 is

$$d = (z_{\alpha/2} + z_\beta)^2 \left(\frac{\theta+1}{\theta-1}\right)^2.$$

Thus, the number of subjects to be included in the trial should be large enough to observe 385 events. Suppose it is planned to have a 3-year intake period and to follow subjects until 4 years after the last subject has entered the trial. It is expected that in the placebo group the incidence is about 1% per year. A rough reasoning along the following lines leads to the required sample size. If subjects are entering the trial at a constant rate, the intended follow-up per subject is on average 5.5 years. Per subject in the control we expect,

therefore, .055 events, and in the experimental group .75 × .055 = .041 events. Thus, per pair of subjects, one from the placebo and one from the experimental group, .0951 events are expected. Therefore, to get an expected number of events of 385 events, 385/.0951 = 4,048 subjects are required per group.

Remarks Regarding Formulae

- In this entry, only formulae are provided for the most occurring situations in practice. Sample size formulae are available for many other, less frequently occurring situations as well.

- The above formulae assume a two-sided test. That is, the null hypothesis states that the effects of the experimental and control treatment are identical. In practice, the testing is mostly two sided. Then the test result can be threefold: (1) There is no statistically significant difference, (2) the experimental treatment is statistically significantly better than the control, or (3) the experimental treatment is statistically significantly worse than the control. In one-sided testing, the null hypothesis states that the experimental treatment is equal or worse than the control treatment, against the alternative that the experimental treatment is better. Then the test result is twofold: (1) The experimental treatment is significantly better than the control, or (2) the experimental treatment is not significantly better than the control. For one-sided testing, $z_{\alpha/2}$ has to be replaced in the formulae by z_{α}. For instance, 1.96 is replaced by 1.645.

- The above formulae assume equal group sizes and can be modified to unequal group sizes. If $n_2 = kn_1$ for some choice of k, the numbers of subjects needed are

$$n_1 = \frac{1}{2}n(1+k) \text{ and } n_2 = \frac{1}{2}n\left(1+\frac{1}{k}\right),$$

where n is the result of the original formulae.

For instance, in the above example comparing the Atkins diet with a control diet, 60 subjects per group were needed. In case one wishes to randomize two times as many subjects in the Atkins group as in the control group, 1/2 × 60(1 + 2) = 90 subjects in the Atkins group and 1/2 × 60(1 + 1/2) = 45 subjects in the control group are required.

- The sample size depends very strongly on the choice of δ. It is always inversely proportional to the square of δ. For instance, if the size of δ is halved, four times as many subjects are needed. The choice of δ is mostly rather subjective, but it is crucial and therefore has to be well motivated.

- The sample size depends very strongly on the choice of the power 1 − β. For instance, 90% instead of 80% requires 34% more subjects. The choice of the power is mostly subjective, but it is crucial and therefore has to be well argued.

- Sample size formulae can be rewritten such that they yield the power for a given specified sample size.

Theo Stijnen

See also Experimental Design; Statistical Notations

Further Readings

Chow, S. C., Shao, J., & Wang, H. (2007). *Sample size calculations in clinical research* (2nd ed.). Boca Raton, FL: CRC Press.

SCALING

Scaling is the process of numerically measuring a health state utility. Utility is a global, composite, preference-based measure of health-related quality of life. Utility-based measures ask respondents to indicate their preference or desire for a health state, either their own or a hypothetical description. Utilities are scaled from 0 (*death*) to 1 (*full or perfect health*), although negative values can be assigned to health states considered to be worse than death. Utility is particularly valuable as a quality weight for length of life. In many decision analyses, cost-effectiveness analyses, and clinical studies, the main outcome is quality-adjusted life years, or QALYs, which are calculated by multiplying length of life by utility. Thus, 10 years in perfect health equal 10 QALYs, while 10 years in a health state with a utility of .75 equal 7.5 QALYs.

There are several standard scaling methods to obtain health state utilities. The most frequently used are the standard gamble, time trade-off (TTO), and rating scale. They differ in theoretical background, methodology, and outcome.

Theoretical Perspective

In the 19th century, utilitarian philosophers defined *utility* as the pleasure, good, or happiness, or prevention of pain, evil, or unhappiness produced by an object. Economists subsequently adapted *utility* to mean the satisfaction or pleasure that a consumer derived from a commodity or service. In both usages, utility was considered to be subjective, summable across individuals, and a motive for behavior.

In 1944, a mathematician and an economist, John von Neumann and Oskar Morgenstern, respectively, proposed that to the extent that utilities were preferences (i.e., an individual can say which object he or she prefers over another), utility could be numerically measured. Their method of measuring utilities involved making choices between alternative outcomes, where one included a risk. The value of any outcome could be inferred from how much risk an individual would take to avoid it. The axioms of von Neumann and Morgenstern defined how a rational individual *ought* to make decisions under conditions of uncertainty, that is, when decisions involved risk or chance, and provided proof of the existence of numerical utilities. These axioms became the foundation of expected utility theory, according to which an individual will behave or make choices to maximize his utility.

Scale Properties of Utilities

To be used as QALY weights, utilities must be measured on an interval scale. An interval scale is one in which changes of the same size have the same meaning anywhere on the scale, but 0 is an arbitrary value (such as temperature). Thus, a change in utility from .2 to .4 must be the same as a change from .7 to .9. The usual end points are *death* (0) and *full health* (1), but sometimes worst health is anchored at 0. An interval scale is a type of cardinal scale and allows all parametric statistical calculations.

Direct Scaling Methods for Measuring Utilities

Standard Gamble

The standard gamble (SG) offers respondents a series of choices between the certainty of spending a specified time period in the health state of interest and taking a hypothetical treatment that has an X% chance of immediate death and a (100 – X)% chance of full health. The health state of interest can be a patient's own health or a description of a hypothetical but plausible health state, often prototypical of a disease or condition. The chances of full health and death are varied, either by direct titration or by "ping-pong." In the titration procedure, the first choice offered is between the health state of interest and a treatment that gives a 100% chance of full health and 0% chance of death. Once this is accepted, the chance of full health is decreased, usually by 5% at a time (95, 90, 85, etc.), and the chance of death increased (5, 10, 15, etc.), until the respondent indicates that he will not accept the gamble or cannot decide between the two choices. The point of indifference is either at this indecision or midway between the chance of full health that is accepted and the chance that is not accepted. In the ping-pong approach, the chance of full health is varied from high to low: 100%, 5%, 95%, 10%, 90%, and so on. The chance of death is always 100 minus the chance of full health. The ping-pong approach gradually closes in on the respondent's point of indifference between the choice of the gamble and the health state of interest. Utility for the health state is defined as 1 minus the probability of death at the point of indifference between the certainty and the risk. Thus, indifference between staying in current health and a potentially curative treatment with a 20% chance of death yields a utility of .80 for current health. If a health state is very undesirable the respondent should be willing to take a high risk of death to avoid it, and the respondent's utility for that health state will be low. Utilities for health states worse than death can be elicited by asking respondents to make a choice between certain death and a gamble in which full health occurs with a probability X and staying in the health state occurs with a probability of 1 – X. If the health state is very undesirable, the person would not take the gamble unless X is very high. Utility is calculated as $-X/(1 - X)$ and therefore has a much larger range than the 0 to 1.0 limit for utilities for health states preferred to death. This can be corrected by assigning –1 to the least preferred health state and scaling the others between it and 0. Another method is to divide the negative utility by 1 minus itself; for

example, if $X = .95$, utility = −19, and −19/(1 − (−19)) = −.95. The resulting values should be interpreted with caution.

Rating Scale

Also known as a visual analog scale or feeling thermometer, the rating scale (RS) originated in psychophysics to measure response to sensory stimuli such as light. It is used to measure feelings and attitudes in psychometrics, and it has been widely used to measure health status. For the RS, respondents position a health state on a linear scale with 0, representing *death*, at one end and 100, representing *full health*, at the other. The utility for the health state is determined by dividing the distance on the scale from death to the health state by 100. Thus, a health state rated three quarters up the scale yields a utility of .75.

For health states considered to be worse than death, the low anchor (0) is changed to the least preferred health state, and the other health states, including death, are placed between it and full health. Some transformation is required to convert the resulting utilities to the standard scale on which 0 equals *death*.

Time Trade-Off

The TTO asks the respondent to make a series of choices between spending some period of time (often the rest of his life) in the health state of interest or spending fewer years in full health. The number of years in full health is varied until the respondent is indifferent to the two choices. The less desirable a health state is perceived to be, the more time will be traded off to achieve perfect health. Utility is calculated by dividing the number of years to be spent in full health by the number of years that would otherwise be spent in the less than perfect health state, yielding a number between 0 and 1.

For health states worse than death, the choice is between immediate death and spending a variable number of remaining life years in full health (y years) followed by the remainder of life ($T - y$) in the health state of interest. For example, if remaining life years (T) were 10, the first choice would be immediate death or 9 (y) years in full health and 1 ($10 - y$) year in the health state worse than death. If life was accepted, fewer years of full health and more years in the worst health state would be offered, using either titration or ping-pong, until the point of indifference between the two alternatives could be determined. Utility is calculated as $y/(y - T)$ at this point. The transformations to a scale of 0 to −1, described above for the SG, can be used but should be interpreted with caution.

Utility Versus Value

The SG is the original utility scaling method, as it is based directly on the axioms of von Neumann and Morgenstern. It asks respondents to make a choice involving an uncertain outcome. The TTO was developed as a choice-based alternative that would be easier for people who do not understand probabilities, but technically it does not yield true utilities because it does not involve uncertainty or risk. The RS is probably the simplest method of health state valuation, but it does not involve risk, has no roots in expected utility theory, and is not choice based.

Purists distinguish between "utilities" derived from the SG and following the axioms of expected utility theory, and the "values" derived from the other scaling methods. However, much of the literature fails to make this distinction and refers to all these measures of health status as utilities.

Factors Affecting Utility Values

Scaling Method

Different scaling methods do not produce the same ratings for a given health state. In general, utilities derived from the SG tend to be higher than those from the TTO which are, in turn, higher than those from the RS.

While there is no universally accepted gold standard utility measure, the SG is the only one that is based on expected utility theory. However, some researchers have proposed that people do not always behave according to expected utility theory; that is, they may not always choose to act in a way that maximizes their utility. All scaling methods have been reported to have some kind of measurement bias. Either "end-of-scale" or "spacing out" bias may affect how respondents use the RS; as a result, the intervals between health states do not represent

the true differences as perceived by the respondent. The higher utilities obtained with the SG are said to be due to *risk aversion*; that is, the certainty of a very undesirable health state is preferred over even a small risk of death. The TTO, like the SG, presents a choice and may be subject to *loss aversion* (not wanting to give up years of life). Other issues with the TTO include the effects of time duration (years of life being considered) and time preference (years in the near future are valued more highly than years in the distant future) on utilities.

Although different scaling methods usually produce different utilities, the utilities are generally correlated. If several health states are rated, the order of preference will be the same regardless of scaling method. RS values are sometimes converted to SG or TTO utilities, but there is no agreement on the best method.

Respondents

The general consensus is that patients rate their own health states higher than individuals who have not experienced the disease rate a scenario describing it, presumably because patients adapt to their limitations. In cost-utility analysis, the consequence is that the maximum potential utility gain from a treatment is smaller if patients' utilities are used. The Panel on Cost Effectiveness in Health and Medicine, appointed by the U.S. Public Health service, recommends that societal (nonpatient) utilities be used for "reference case" analyses carried out from a societal perspective, while patients' utilities may be appropriate when interventions for the same condition are being compared. For individual decision making, the utilities of each patient should be used.

Presentation of Health States

When utilities are being elicited for health states other than patients' own, the method used to present the health states (e.g., point form, narrative text, or multimedia) can influence the utility values obtained. Regardless of respondents, framing (presenting an outcome as a loss vs. a gain), anchoring (starting points for SG or TTO), and selection of upper bound ("normal" or "perfect" health, or absence of symptoms) and lower bound ("worst health" or "death") may also affect utilities.

Measurement Properties

All three scaling measures demonstrate good reliability, validity, and responsiveness. The SG and TTO are more time-consuming than the RS, but subjects are generally willing to complete them. Computer programs are available to ease administration, scoring, and data entry. If subjects are extremely risk-averse, the SG may be less responsive to changes in health status than the other methods.

Choosing the Method

Which scaling method should be used? The RS is easy to understand and administer and has been widely used to measure health status. It elicits a value rather than a utility, and the true relationship between the RS and the SG or TTO has not been determined, so use of the RS as a surrogate is not recommended. The SG and TTO require different and more complex thought processes. Props (Chance Board for the SG and TTO Board) facilitate explanation, administration, and comprehension of these tasks. The SG is the classic method for decision making under uncertainty. Most healthcare interventions involve uncertainty and risk, an argument in favor of the SG.

Given the differences in theoretical background and the effects of scaling method, respondent, and presentation on utilities, decision makers should try to use similarly derived utilities for each decision analysis or other study where possible. Sensitivity analysis should be used to explore the effects of variation in utility values on the results.

Karen E. Bremner

See also Cost-Utility Analysis; Expected Utility Theory; Gain/Loss Framing Effects; Quality-Adjusted Life Years; Utility Assessment Techniques

Further Readings

Brazier, J. E., Deverill, M., Green, C., Harper, R., & Booth, A. (1999). A review of the use of health status measures in economic evaluation (Full report). *Health Technology Assessment*, *3*(9).

Drummond, M. F., O'Brien, B., Stoddart, G. L., & Torrance, G. W. (1997). Cost-utility analysis. In M. F. Drummond, B. O'Brien, G. L. Stoddart, & G. W.

Torrance (Eds.), *Methods for the economic evaluation of health care programmes* (pp. 139–204). New York: Oxford University Press.

Froberg, D. G., & Kane, R. L. (1989). Methodology for measuring health-state preferences-II: Scaling methods. *Journal of Clinical Epidemiology*, *42,* 459–471.

Gold, M. R., Patrick, D. L., Torrance, G. W., Fryback, D. G., Hadorn, D. C., Kamlet, M. S., et al. (1996). Identifying and valuing outcomes. In M. R. Gold, L. B. Russell, J. E. Siegel, & M. C. Weinstein (Eds.), *Cost-effectiveness in health and medicine* (pp. 82–134). New York: Oxford University Press.

Patrick, D. L., Starks, H. E., Cain, K. C., Uhlmann, R. F., & Pearlman, R. A. (1994). Measuring preferences for health states worse than death. *Medical Decision Making, 14,* 9–18.

Torrance, G. W. (1989). Utility approach to measuring health-related quality of life. *Journal of Chronic Diseases, 40,* 593–600.

Torrance, G. W. (2006). Utility measurement in healthcare: The things I never got to. *Pharmacoeconomics, 24*(11), 1069–1078.

Torrance, G. W., & Feeny, D. (1989). Utilities and quality-adjusted life years. *International Journal of Technology Assessment in Health Care, 5,* 559–575.

von Neumann, J., & Morgenstern, O. (1944). *Theory of games and economic behavior*. Princeton, NJ: Princeton University Press. (The notion of utility, its measurement, and the standard gamble are put forth in chap. 1, sec. 3.3, pp. 17–20.)

Screening Programs

Screening programs are used for the early detection of disease to decrease mortality and to increase quality of life. This subject is important in both clinical practice and public health, since screening involves a substantial part of the population. As people are already ill when a disease is detected, it is used for secondary prevention. Early detection means detecting a disease at an earlier (presymptomatic) stage than would usually occur in standard care, as patients have no clinical complaints and therefore no reason to seek medical care. A necessary condition for screening tests is the availability of intervention for the detected illness and a better effectiveness of the intervention when provided early. This entry describes important criteria to assess the pros and cons of screening programs as the basis for informed medical decision making for patients, care providers, and public health experts and policy makers. This entry illustrates these criteria using the practical context of cancer screening, which is similar to screening in the case of other diseases.

Criteria to Evaluate Screening Programs

The question whether patients benefit from early detection of disease includes the following components:

1. Can the disease be detected early? Are there feasible and practical measurements available?
2. What are the sensitivity, specificity, and predictive values of the test? How serious is the problem of false-positive test results?
3. What are the costs and harms?
4. Do the individuals in whom disease is detected early benefit from early detection, and is there an overall benefit to those who are screened?

To evaluate screening programs, several outcome and process measures are used: The reduction of incidence (if possible, e.g. cervical cancer or colon cancer) and of disease-specific and overall mortality in the population screened; the increase of cases detected at an earlier stage, leading to less case fatality in screened individuals; the reduction in complications; the prevention or reduction in recurrences or metastases; and finally, the improvement of quality of life in screened individuals.

Process variables contain the number of people screened, the proportion of target population screened and number of times screened, the detected prevalence of preclinical disease, the total costs of the program, the costs per case and previously unknown cases found, the proportion of positive screenees brought to final diagnosis and treatment, and the predictive value of a positive test in populations screened.

It is important to note that screening has benefits and harms. An important benefit is a reduction of mortality or an increased quality of life. Important harms include costs, false alarms, and overdiagnosis. Costs include not only financial costs but also nonfinancial costs to patients, including anxiety,

emotional distress, inconvenience, and overtreatment due to overdiagnosis. Thus, the "cost" of a test is not just the cost of the test procedure.

Natural History, Progression, and Regression of Disease

When evaluating the benefits of screening programs, it is essential to place screening in the appropriate timeline of the natural history of the disease. At some point, biologic onset of disease occurs without symptoms as, for instance, a subcellular change or alteration of DNA. The time between onset and the symptoms' start is called the preclinical phase. Later, the disease becomes symptomatic, when it moves into the clinical phase and the patient seeks care, gets a diagnosis, and is treated, which in turn will end up in cure, disease control, disability, or death. Screening is assigned to the asymptomatic, preclinical phase for early detection of the disease in the hope of more effective treatment than would be possible in a later stage. That means that the diagnosis is advanced to an earlier stage (called lead time, an inherent concept in screening). Early detection does not mean final diagnosis but suspicion of the disease, which needs further investigation to verify it.

Another inherent concept in screening is the critical point in the natural history of the disease before which care is more effective and easier to administer. This point might be, for breast cancer, the time when it has not spread to the axillary lymph nodes, so that prognosis is better before than after spread has taken place.

Test Accuracy of Screening Programs and Their Understanding

Each test can make two errors: false positives and misses. The false-positive rate is the proportion of positive tests among patients without the condition (Table 1). The miss rate (false-negative rate) is the proportion of negative tests among patients who actually have the condition. The specificity is the proportion of negative tests among patients without the condition. The false-positive rate and the specificity add up to 1 (100%). The sensitivity is the proportion of positive tests among clients with the condition. False-negative rate and sensitivity, again, add up to 1 (100%; Table 1).

Sensitivities and specificities continue to confuse physicians and patients alike. Figure 1 illustrates how these can be translated into natural frequencies to facilitate deriving the positive predictive value (PPV) of the test, which means the proportion of ill people among all clients with positive tests. For interpretation, it is relevant to notice that the PPV varies depending on disease prevalence in different populations.

Table 1 Four possible test outcomes (percentages refer to Figure 1)

	Disease	
Test result	*Yes*	*No*
Positive	*a* (90%) Sensitivity	*b* (9%) False-positive rate
Negative	c (10%) False-negative rate	*d* (91%) Specificity
	Sensitivity = $a/(a + c)$	Positive predictive value = $a/(a + b)$
	Specificity = $d/(b + d)$	Negative predictive value = $d/(c + d)$

Notes: Testing for a disease can have four possible results: a positive result given disease, a positive result given no disease, a negative result given disease, and a negative result given no disease. The rates with which these four results occur are called sensitivity (true-positive rate), false-positive rate, false-negative rate, and specificity (true-negative rate). The two shaded areas indicate the two possible errors, false positives and false negatives.

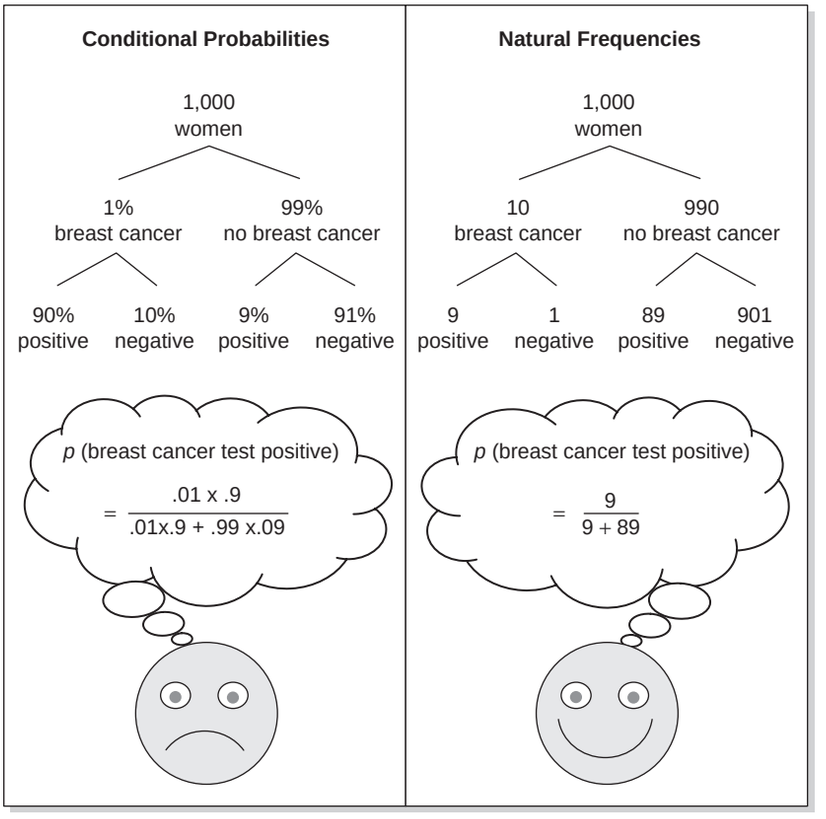


Figure 1 What is the probability that a woman who tests positive in mammography screening actually has breast cancer (positive predictive value)?

Source: Gigerenzer, G., Gaissmaier, W., Kurz-Milcke, E., Schwartz, L. M., & Woloshin, S. (2007). Helping doctors and patients make sense of health statistics. *Psychological Science in the Public Interest, 2,* 53–95. Reprinted with permission.

Note: The left side illustrates the calculation with conditional probabilities and the right side with natural frequencies. The four probabilities at the bottom of the left tree are conditional probabilities. Each one is normalized on base 100. The four frequencies at the bottom of the right tree are natural frequencies. The calculation is simpler (smiling face) because natural frequencies are not normalized relative to base rates of breast cancer, whereas conditional probabilities (or relative frequencies) are, and they need to be multiplied by the base rates. The formula is known as Bayes's rule.

To assess sensitivity and specificity reliably for all cells of Table 1, complete data would be needed. However, often only those with positive results ($a + b$) are sent for further testing with a well-proven gold standard. Data for negative tests are frequently missing in routine testing because these patients do not receive further testing. In particular, when the gold standard means invasive procedures for further testing (e.g., transrectal ultrasound and biopsy of prostate in prostate-specific antigen [PSA] test), those data are hard to obtain, resulting in unreliable measures of test characteristics.

Sources of Bias When Measuring the Effectiveness of Screening

Prognostic Selection Bias

Screening programs tend to detect disease in people with long preclinical phase, since the chance of detection is higher than for those with very short preclinical phase. Thus, slowly progressive, less malignant tumors are more likely to be detected, in contrast to fast-growing, aggressive tumors, which are less likely to be detected. Therefore, better prognoses are more likely to be

detected, so that it is not clear whether a possible benefit stems from screening or just prognostic selection as length-biased sampling. In consequence, the ability to detect a disease does not necessarily mean that the screening has benefits.

The assumption of benefit from screening only holds when

1. the disease has a detectable preclinical phase and
2. without intervention, all or most cases in the preclinical phase progress to the clinical stage.

According to Zahl, Mæhlen, and Welch, this is questionable, since the preclinical stage may be too short for early detection in extremely rapid progression. Also, spontaneous regression may occur, so that not every preclinical precancerous lesion progresses to the full cancer. Consider screening initiatives provided once a year from age 50 on. Rapid progressive tumors are harder to detect early at the screening time points than slowly progressive ones, because the rapid ones can progress into the clinical stage in less than a year, rendering early detection very difficult. Moreover, those who earlier in life had rapidly progressive tumors would not benefit at all.

Lead Time Bias

Another problem occurs when benefits are evaluated by comparisons of 5-year survival rates in screened and unscreened people. Such rates are defined as the proportion of survivors out of diseased patients 5 years from diagnosis. Increased survival rates may not be a result of screened people living longer but rather a result of diagnosis being made at an earlier point in the natural history of the disease. This is called lead time bias. In such circumstances, patients do not derive any benefit from earlier detection. Indeed, they may lose out regarding quality of life, as they do not live longer, but live longer with the diagnosis and are at a higher risk of being overtreated. To find out whether early detection is beneficial, one should rather compare mortality from the disease in the entire screened group with that in the unscreened group. Unlike survival rates, mortality includes all people in the denominator, not only those with the disease. Thus, survival and mortality rates are not simply opposites. Welch, Schwartz, and Woloshin found that across the 20 most common solid tumors in the United States, changes in 5-year survival over the last 50 years were completely uncorrelated with changes in mortality. Thus, one should not draw any conclusions about the effectiveness of screening from improved 5-year survival rates, although very often benefits are misleadingly claimed to be proven based on this statistic. For instance, many smokers, current and past, wonder whether to get a CT scan to screen for lung cancer. While advertisements (misleadingly) promote screening because of higher survival rates, there is no evidence for reduced mortality rates; yet it is known that CT screening will harm (and even kill) some people through unnecessary biopsies, overdiagnosis, and overtreatment. That is why no professional group currently recommends the test (in fact, the American College of Chest Physicians' recommendations argue against routine CT screening).

Overdiagnosis Bias

Another source of bias is true-positive detection of pseudodiseases. These are low virulent cancers, which never progress to clinical disease or even spontaneously regress, but which are still detected by screening and often unnecessarily treated. Thus, overdiagnosis can lead to overtreatment with harmful interventions that are not necessary. For instance, Schwartz and Woloshin found that about 25% of breast cancers detected by mammography are overdiagnoses.

Moreover, false-positive testing of healthy individuals screened and diagnosed as having cancer when in reality they do not can also lead to overdiagnosis and overtreatment.

Overdiagnosis could convey the false impression of increased rates of detection and diagnosis of early stage cancer as a result of screening. Beyond this, many people diagnosed with cancer in the screened group would actually not have cancer, and would therefore have good survival; the results would inflate survival, resulting in a mistaken conclusion that screening would have been shown to improve survival from cancer in the population. To avoid overdiagnosis bias, it is essential to standardize diagnostic processes as rigorously as possible.

Referral or Volunteer Bias

In deriving conclusions about the benefits of screening, one needs to ensure that both people who are screened and people who are not screened otherwise have the same characteristics. Often, volunteers who participate are healthier than the general population and are more likely to comply with medical recommendations. Thus, a lower mortality would be observed even if early detection played no role in improving the prognosis. Vice versa, it is possible that people with high risk have higher participation rates due to family history or risky lifestyles. Both phenomena are called referral or volunteer bias, of which the direction is often difficult to determine. This bias can strongly affect correct interpretation of benefits.

The best solution is, therefore, randomized controlled trials (RCTs), in which the randomization procedure ensures that the two groups have comparable initial prognostic profiles. A recent Cochrane review on mammography screening including six trials involving half a million women illustrates that the lack of proper randomization could result in an overestimation of the benefits. While two trials with adequate randomization did not show a significant reduction in breast cancer mortality, four trials with suboptimal randomization did.

Conclusion

Screening programs should entail quality-proven, practical, and feasible tests for early detection of diseases, systematically addressed to all people of a predefined target population, and accompanied by standardized documentation and continuous, independent quality assessment from the first invitation up to the evaluation of the predefined end points. Moreover, there is a need for information on pros and cons for informed decision making of the participants. The target population contains asymptomatic people who feel healthy, and among whom indeed most are definitively healthy, but does not include individuals with symptoms seeking care. Therefore, side effects are less tolerable and benefits have to be rigorously proven to outweigh possible harms. The most important benefits are the reduction of incidence and mortality of cancer and the improvement of quality of life. The most important harms are false alarms, overdiagnosis, and overtreatment. Programs proven to decrease disease-specific and overall mortality are extraordinarily effective. Evidence should be based on RCT, and all the quality criteria should be evaluated systematically before the programs are invented.

Angela Neumeyer-Gromen
and Wolfgang Gaissmaier

See also Bias in Scientific Studies; Cost-Benefit Analysis; Costs, Direct Versus Indirect; Diagnostic Tests; Informed Decision Making; Quality of Well-Being Scale; Randomized Clinical Trials; Risk Communication

Further Readings

Giersiepen, K., Hense, H. W., Klug, S. J., Antes, G., & Zeeb, H. (2007). Planning, implementation and evaluation of cancer screening programs. *Zeitschrift für ärztliche Fortbildung und Qualitätssicherung, 101,* 43–49.

Gigerenzer, G., Gaissmaier, W., Kurz-Milcke, E., Schwartz, L. M., & Woloshin, S. (2007). Helping doctors and patients make sense of health statistics. *Psychological Science in the Public Interest, 2,* 53–95.

Gordis, L. (2008). *Epidemiology.* Philadelphia: Saunders Elsevier.

Gøtzsche, P. C., & Nielsen, M. (2006). Screening for breast cancer with mammography. *Cochrane Database of Systematic Reviews, 4,* CD001877.

Hulka, B. S. (1988). Degrees of proof and practical application. *Cancer, 62,* 1776–1780.

Schwartz, L. M., & Woloshin, S. (2007). Participation in mammography screening. *British Medical Journal, 335,* 731–732.

UK National Screening Committee. (2003). *Criteria for appraising the viability, effectiveness and appropriateness of a screening programme.* Retrieved November 26, 2008, from http://www.nsc.nhs.uk/pdfs/criteria.pdf

Welch, H. G., Schwartz, L. M., & Woloshin, S. (2000). Are increasing 5-year survival rates evidence of success against cancer? *Journal of the American Medical Association, 283,* 2975–2978.

Zahl, P. H., Mæhlen, J., & Welch, H. G. (2008). The natural history of invasive breast cancers detected by screening mammography. *Archives of Internal Medicine, 168*(21), 2311–2316.

SF-6D

The SF-6D provides a method for estimating health state utility values from data obtained from patients using the SF-36 (or SF-12) health status questionnaires. The SF-6D has two parts. First, patients who have completed the SF-36 are assigned to a health state classification (i.e., SF-6D). This classification describes health on six multilevel dimensions: physical functioning, role limitations, social functioning, pain, mental health, and vitality. Second, there are algorithms for scoring each state based on values obtained from general population surveys using the standard gamble (SG). These health state utility values can be used to calculate quality-adjusted life years (QALYs) for cost-effectiveness analysis. There are two versions of the SF-6D, one for use with the SF-36 and the other for the SF-12.

Derivation of the SF-6D From SF-36

The SF-36 is the most widely used generic measure of health status in the world. It yields scores across eight dimensions and two summary scores. The SF-36 is useful for assessing changes in self-perceived health status or health-related quality of life across these eight dimensions. However, these dimension scores are derived by either simply summing responses to the SF-36 items or by the use of weights from factor analysis; as neither of these is likely to reflect people's preferences for the health states, they cannot be used in economic evaluation. Furthermore, there is no means of combining across the dimensions or combining with survival for cost-effectiveness analysis. The SF-6D was developed to estimate health state utility values from the large number of data sets collected using the SF-36.

The development of the SF-6D involved three stages. The first was the development of the SF-6D health state classification from the SF-36. The second was a valuation survey to value a sample of states defined by the SF-6D. The third stage was the econometric analysis of the health state valuation data to estimate an algorithm for scoring all states defined by the SF-6D. The first version of the SF-6D was reported in 1998, but this was substantially revised in the publication of 2002 and again in 2004. This entry only reports the latter.

SF-6D Health State Classification

The SF-6D was constructed from a selection of 11 items drawn from the SF-36. These items were selected from the SF-36 to minimize the information loss within the constraint that the resultant health state classification must be amenable to valuation. The item selection process was undertaken using evidence of the psychometric properties of the items and the factor analyses undertaken by John Ware and his colleagues in developing the SF-12.

The SF-6D has six multilevel dimensions: physical functioning, role limitation, social functioning, pain, mental health, and vitality (Table 1). The number of levels per dimension is between four and six levels of functioning or well-being, depending on the response choice categories of the original items from the SF-36. The SF-36 version of the SF-6D defines 18,000 states. The version of the SF-6D derived from the SF-12 has the same six dimensions and only differs in having just three levels for physical functioning and five for pain, which in all define 7,500 states. The SF-12 version of the SF-6D (i.e., SF-6D (12)) uses 7 items of the SF-12 (a subset of the 11 used in SF-6D (36)). Both versions of the SF-6D can be used with Versions 1 and 2 of the SF-36 and SF-12.

Valuation Survey

The UK value set for both versions of the SF-6D comes from a representative sample of 836 members of the UK general population (response rate 65%) who were interviewed and asked to value a total of 249 states defined by the SF-6D using the standard gamble (each respondent valued 6 states). Five health states were valued against full health and the worst state was defined by the SF-6D. The worst state was then valued against full health and death to transform the valuations of the intermediate states onto the full health/death scale required to calculate QALYs. There were 225 respondents excluded for either failing to value the pits state,

Table 1 SF-6D (SF-36 version)

	Physical Functioning		*Pain*
1	Your health does not limit you in vigorous activities	1	You have no pain
2	Your health limits you a little in vigorous activities	2	You have pain but it does not interfere with your normal work (both outside the home and housework)
3	Your health limits you a little in moderate activities	3	You have pain that interferes with your normal work (both outside the home and housework) a little bit
4	Your health limits you a lot in moderate activities	4	You have pain that interferes with your normal work (both outside the home and housework) moderately
5	Your health limits you a little in bathing and dressing	5	You have pain that interferes with your normal work (both outside the home and housework) quite a bit
6	Your health limits you a lot in bathing and dressing	6	You have pain that interferes with your normal work (both outside the home and housework) extremely
	Role Limitations		*Mental Health*
1	You have no problems with your work or other regular daily activities as a result of your physical health or any emotional problems	1	You feel tense or downhearted and low none of the time
2	You are limited in the kind of work or other activities as a result of your physical health	2	You feel tense or downhearted and low a little of the time
3	You accomplish less than you would like as a result of emotional problems	3	You feel tense or downhearted and low some of the time
4	You are limited in the kind of work or other activities as a result of your physical health and accomplish less than you would like as a result of emotional problems	4	You feel tense or downhearted and low most of the time
		5	You feel tense or downhearted and low all of the time
	Social Functioning		*Vitality*
1	Your health limits your social activities none of the time	1	You have a lot of energy all of the time
2	Your health limits your social activities a little of the time	2	You have a lot of energy most of the time

3	Your health limits your social activities some of the time	3	You have a lot of energy some of the time
4	Your health limits your social activities most of the time	4	You have a lot of energy a little of the time
5	Your health limits your social activities all of the time	5	You have a lot of energy none of the time

Source: Reprinted from *Journal of Health Economics* 21(2), Brazier J, Roberts J, Deverill M. The estimation a preference-based single index measure for health from the SF-36, pp. 271–292, copyright © 2002, with permission from Elsevier.

Note: The SF-36 items used to construct the SF-6D are as follows: physical functioning Items 1, 2, and 10; role limitation due to physical problems Item 3; role limitation due to emotional problems Item 2; social functioning Item 2; both bodily pain items; mental health Items 1 (alternate version) and 4; and vitality Item 2.

producing fewer than two values, or producing values without any variation. This left 611 respondents in the data set providing 3,518 observed SG valuations across the 249 health states. Mean health state values ranged from .21 to .99 and the standard deviations ranged from .2 to .45.

Estimation of Scoring Algorithm

A valuation algorithm has been estimated from the SG valuation data to value all health states defined by the SF-6D. This was a complex data set that was skewed, bimodal, truncated, and clustered (by respondent). A range of alternative modeling specifications were examined and the best performing one selected on the basis of predictive ability. The best model in terms of prediction was an ordinary least squares model of the mean health state values, where each dimension level was entered as a dummy variable along with a crude interaction term (that equaled 1 when any dimension was at a severe level). The original scoring algorithm for the SF-6D (36) was based on this model, and a similar model was used to estimate an algorithm for the SF-6D (12). These algorithms are presented in Table 2. The main difference between the algorithms and the models is that inconsistent coefficients from the model were merged.

The SF-36 and SF-12 are copyrighted and can be obtained from the Medical Outcomes Trust and QualityMetric. Programs for applying the SF-6D to SF-36 or SF-12 data sets are available from the University of Sheffield free of charge for noncommercial applications. Programs are available in Excel, SPSS, and SAS.

Comparison With Other Preference-Based Measures

Differences in mean scores have often been found to be little more than .05 between SF-6D and two widely used measures, the EQ-5D and HUI3. This mean statistic masks considerable differences in the distribution of scores. There is a substantial disagreement between the scores at the individual level (see Figure 1). The ranges differ markedly, with the range for the EQ-5D, for example, covering –.4 to 1.0, as compared with .3 to 1.0 for the SF-6D. Negative values on the EQ-5D are associated with values on the SF-6D as high as .75. There is a larger cluster of data points to the right of the line of agreement that are all associated with patients being at the most severe level in one or more of the EQ-5D dimensions. In the UK EQ-5D scoring algorithm, these patients have an additional decrement known as the "N3" term.

There is evidence that the SF-6D suffers from floor effects in its descriptive systems. Patients with more severe problems in physical functioning, role limitations, and social functioning tend to be on or near the bottom level. This means that

Table 2 Scoring for the SF-36 and SF-12 versions of the SF-6D

SF -6D (SF-36)		*SF-6D (SF-12)*	
C	1.000	c	1.000
PF23	–0.035		
PF4	–0.044		
PF5	–0.056	PF3	–0.045
PF6	–0.117		
RL234	–0.053		
		RL234	–0.063
SF2	–0.057	SF2	–0.063
SF3	–0.059	SF3	–0.066
SF4	–0.072	SF4	–0.081
SF5	–0.087	SF5	–0.093
PAIN23	–0.042		
PAIN4	–0.065	PAIN3	–0.042
PAIN5	–0.102	PAIN4	–0.077
PAIN6	–0.171	PAIN5	–0.137
MH23	–0.042	MH23	–0.059
MH4	–0.100	MH4	–0.113
MH5	–0.118	MH5	–0.134
VIT234	–0.071	VIT234	–0.078
VIT5	–0.092	VIT5	–0.106
MOST	–0.061	MOST	–0.077

Source: Brazier J, Roberts J. The estimation of a preference-based measure of health from the SF-12. *Medical Care* 2004, 42:851–59. Reprinted with permission.

the SF-6D is less able to detect changes in such patients. In contrast, the EQ-5D suffers from ceiling effects, since a large number of patients are in the best health state (i.e., Level 1 for each dimension), and so is less able to detect changes in patients with milder problems.

A review comparing the sensitivity of the SF-6D with other generic preference-based measures found a mixed picture. The SF-6D was found to be more sensitive in the general population, medical rehabilitation, hearing aid provision, and leg reconstruction but less sensitive in liver disease and hip replacement. No one preference-based measure is better across all conditions.

New Developments

More Advanced Algorithm Based on a Bayesian Approach

There has been further work on the modeling of the valuation data. The first has been the estimation of an algorithm using a nonparametric Bayesian approach that has been shown to perform better in terms of predictive ability (mean absolute error of .089 as compared with .104 out of sample predictions) and overcomes the bias of the original regression models of underpredicting the worst health states (e.g., it predicts a value of .203 for the worst SF-6D state as compared with .301 using the original algorithm). The overall impact on mean health state values was between .01 and .04 across four data sets. It is recommended that researchers use this algorithm in future work but use the original algorithm for comparability.

Values Based on Rank Data

There are concerns that SG health state values may be contaminated by nonhealth considerations such as loss aversion, and so researchers have been exploring alternative methods that use ordinal data. A model has been estimated from the rank data collected in the original valuation survey using a rank-ordered logit model. The weights generated from rank data were found to be similar to those from SG data. Researchers who do not like SG may want to consider using this algorithm as an alternative.

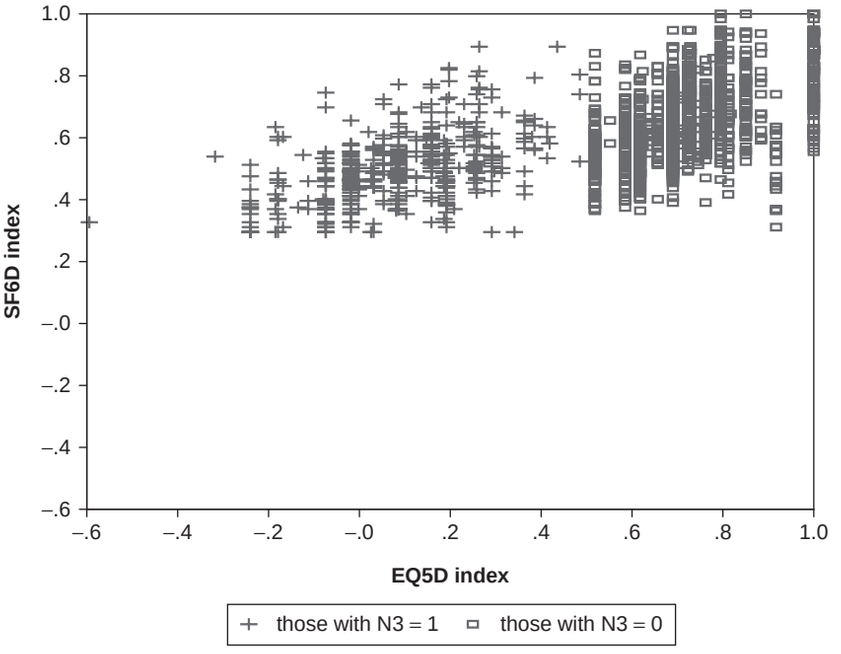


Figure 1 SF-6D as compared with EQ-5D across seven patient groups

Source: Brazier JE, Tsuchiya A, Roberts J, Busschbach J. A comparison of the EQ-5D and the SF-6D across seven patient groups. *Health Economics* 2004, 13(90):873–884. Reprinted with permission.

Predicting SF-6D Index Using Mean SF-36 Dimension Scores

A problem until recently has been that the estimation of the SF-6D requires access to individual level data. However, reviews of published evidence for populating economic models often uncovers studies where it is not possible to obtain individual level data and only mean SF-36 health dimension scores are available. To make the most of such evidence, an algorithm has been estimated for predicting mean SF-6D scores from mean SF-36 dimension scores. The models are reasonably accurate in predicting mean SF-6D values across subgroups or differences in change over time, with more than 90% to within ±.05. All algorithms mentioned above can be obtained from the University of Sheffield's Health Economics and Decision Science Web site.

Valuation Surveys in Other Countries

There is emerging evidence that SG health state values differ between countries. There have been valuation surveys completed in Japan and Hong Kong each with 600 or more respondents. Surveys are currently also being undertaken in Australia, Brazil, Portugal, and Singapore. Researchers interested in using the SF-6D in these countries should consult the University of Sheffield's Health Economics and Decision Science Web site for information on the relevant country-specific research group.

The Future

The SF-6D preference-based measure of health has been widely used to assist in undertaking economic evaluation in healthcare (with over 200 citations in 2008). It meets the requirements of many reimbursement authorities interested in cost-effectiveness evidence. However, there is scope for improving methods and improving our understanding of methods, including the following: Undertake comparisons with other generic preference-based measures in more data sets; explore ways to lower the floor of the SF-6D by introducing additional levels to some dimensions; revalue SF-6D in more countries and formally undertake comparisons across countries; examine methods for reducing the burden from revaluation studies by drawing on

existing data sets using Bayesian methods; and revalue SF-6D using TTO and other methods to enhance comparability across instruments.

John E. Brazier

See also Cost-Effectiveness Analysis; Health Status Measurement, Floor and Ceiling Effects; Health Status Measurement Standards; Quality-Adjusted Life Years (QALYs)

Further Readings

Ara, R., & Brazier, J. (in press). Predicting SF-6D preference-based utilities using the SF-36 health summary scores: Approximating health related utilities when patient level data is not available. *Value in Health*.

Brazier, J. E., & Roberts, J. (2004). Estimating a preference-based index from the SF-12. *Medical Care, 42*(9), 851–859.

Brazier, J. E., Roberts, J., & Deverill, M. (2002). The estimation a preference-based single index measure of health from the SF-36. *Journal of Health Economics, 21,* 271–292.

Brazier, J. E., Tsuchiya, A., Roberts, J., & Busschbach, J. (2004). A comparison of the EQ-5D and the SF-6D across seven patient groups. *Health Economics, 13*(90), 873–884.

Kharroubi, S. A., Brazier, J. E., Roberts, J., & O'Hagan, A. (2007). Modelling SF-6D health state preference data using a nonparametric Bayesian method. *Journal of Health Economics, 26,* 597–612.

Lam, C. L. K., Brazier, J., & McGhee, S. M. (in press). Valuation of the SF-6D health states is feasible, acceptable, reliable and valid in a Chinese population. *Value in Health*.

Longworth, L., & Bryan, S. (2003). An empirical comparison of EQ-5D and SF-6D in liver transplantation patients. *Health Economics, 12,* 1061–1067.

McCabe, C., Brazier, J., Gilks, P., Tsuchiya, A., Roberts, J., O'Hagan, A., et al. (2006). Using rank data to estimate health state utility models. *Journal of Health Economics, 15*(5), 418–431.

Moock, J., & Kohlmann, T. (2004). Comparing preference-based quality of life measures: Results from rehabilitation patients with musculoskeletal, cardiovascular or psychosomatic disorders. *Quality of Life Research, 17*(3), 485–495.

University of Sheffield, Health Economics and Decision Science: http://www.shef.ac.uk/scharr/sections/heds/mvh/sf-6d

Walter, S., & Brazier, J. E. (2003). What is the relationship between the minimally important difference and health state utility values? The case of the SF-6D. *Health and Quality of Life Outcomes, 1,* 4.

SF-36 and SF-12 Health Surveys

The SF-36® Health Survey is a questionnaire with 36 standardized questions (items) used to assess generic health outcomes, which are often referred to as patient-reported outcomes (PRO) or health-related quality of life (HRQOL). The SF-36 measures eight domains of health outcomes: physical functioning, role-physical, bodily pain, general health, vitality, social functioning, role-emotional, and mental health. Version 2 of the SF-36 (SF-36v2® Health Survey) measures the same domains but contains improvements in wording and response choices to cover a wider range of health. The SF-12® Health Survey consists of a subset of 12 items from the SF-36 covering the same eight domains. The SF-36 and SF-12 are part of the "SF family" of patient-reported outcomes measures for adults, which also includes the SF-8™ Health Survey and DYNHA® Generic Health Assessment (a dynamic or "computerized adaptive" instrument). These generic tools are cross-calibrated and scored on the same norm-based metric to maximize their comparability.

The SF-36 and SF-36v2 Health Surveys yield an SF-profile of scores including eight domain scale scores (each summarizing the information from the items within a single domain of health), as well as two component summary scores formed from the eight domain scale scores: a physical component summary (PCS) and a mental component summary (MCS). A self-reported health transition (HT) rating and a preference-based health utility index (SF-6D) can also be scored from the SF-36 and SF-36v2. Version 1 of the SF-12 Health Survey provides PCS and MCS scores and a preference-based health utility index (SF-6D). In addition, Version 2 of the SF-12 (SF-12v2® Health Survey) provides the eight domain scale scores.

The SF-36 and SF-12 Health Surveys are useful for medical decision making by providing

systematic assessment of health outcomes from the patient's point of view. Potential uses include comparing general and specific populations, comparing the relative burden of diseases, differentiating the health benefits produced by a wide range of different treatments, screening individual patients, and predicting healthcare costs, mortality, and other important outcomes. The instruments are frequently used in outcomes studies: The SF-36 has been cited in more than 9,800 publications, including approximately 1,600 published randomized clinical trials, while the SF-12 has been cited in more than 1,000 publications. Among the advantages of the SF-36 and SF-12 are the availability of guidelines for administration, check of data quality, data analysis and interpretation, standardized software for scale scoring and data quality monitoring, and translations into 109 country/language versions. The SF-36 and SF-12 have been successfully administered to persons 14 years and older using self-administration by paper and pencil, the Internet, telephone, interactive voice response (IVR), and personal digital assistant (PDA), as well as interviewer-administered forms. The surveys are available in standard (4-week recall) or acute (1-week recall) forms.

Background

The conceptual and methodological framework for the SF-36 and SF-12 was developed in two large-scale studies of health services: the Health Insurance Experiment (HIE) and Medical Outcomes Study (MOS). These studies measured a broad array of functional status and well-being concepts and demonstrated that scales constructed from self-administered surveys can be reliable and valid tools, yielding high-quality data for assessing changes in health status in the general population and in people with chronic conditions, including the elderly. The HIE and the MOS succeeded in achieving comprehensive assessments of patient-reported health outcomes but still incurred a considerable response burden. The SF-36 was constructed to retain the benefits of a comprehensive, valid, and reliable health assessment but with greater practicality in terms of reduced response burden and ease of use.

The 8 health domains represented in the SF-profile were selected from 40 domains that were included in the MOS. Chosen health domains represented those most frequently measured in widely used health surveys and believed to be most affected by disease and health conditions. The SF-36 was first made available in "developmental" form in 1988 and released in final original form in 1990 by its principal developer, John E. Ware Jr.

In 1991, the International Quality of Life Assessment (IQOLA) Project began with the goal of developing validated translations of a single health status questionnaire that could be used in multinational clinical studies and other international studies of health. The SF-36 was chosen as the health status measure to be translated, adapted, and tested internationally in the IQOLA project. By 1993, 14 countries were represented in the IQOLA Project. Interest in developing translations of the SF-36 continued; as of 2008, there are 109 country/language translations.

SF-36v2 Health Survey

Although the original SF-36 proved to be useful for many purposes, 10 years of experience revealed the potential for improvements. A need to improve item wording and response choices identified through the IQOLA Project, as well as a need to update normative data, led to the development of SF-36v2, which was made available in 1998.

Version 2 includes the following improvements: (a) improved instructions and item wording; (b) improved layout of questions and answers; (c) increased comparability in relation to translations and cultural adaptations, and minimized ambiguity and bias in wording; (d) five-level response options in place of dichotomous choices for items in the role-physical and role-emotional scales; and (e) simplified response options for the mental health and vitality scales. Without increasing the number of questions, improvements make the survey easier to understand and complete, and substantially increase the reliability and validity of scores over a wider range, thereby reducing the extent of floor and ceiling effects in the role performance scales.

SF-12 and SF-12v2 Health Surveys

The SF-12 Health Survey was developed to offer a shorter alternate version of the SF-36, measuring

the same eight domains of health. Version 1 of the SF-12 succeeded in achieving PCS and MCS scores that are comparable to the PCS and MCS scores of the SF-36, but reporting of the scale score profile is not recommended for Version 1 of the SF-12. Improvements to the SF-12v2 allow for scoring the eight-scale score profile, in addition to the two component summary scores and the health utilities index.

Scales and Component Summaries

Physical Functioning

The Physical Functioning (PF) scale reflects the importance of distinct aspects of physical functioning and the necessity of sampling a range of severe and minor physical limitations. The SF-36 includes 10 PF items representing levels and kinds of limitations between the extremes of physical activities, including lifting and carrying groceries; climbing stairs; bending, kneeling, or stooping; and walking moderate distances. One self-care item is included. The SF-12 includes two PF items, focusing on moderate activities and climbing stairs. The PF items capture both the presence and extent of physical limitations using a three-level response continuum.

Role-Physical

The Role-Physical (RP) scale covers an array of physical health-related role limitations in the kind and amount of time spent on work, the difficulties performing work, and the level of accomplishment associated with work or other usual activities. The SF-36 includes four and the SF-12 includes two RP items.

Bodily Pain

The SF-36 Bodily Pain (BP) scale includes two items: one pertaining to the intensity of bodily pain and one measuring the extent of interference with normal work activities due to pain. The SF-12 includes the latter BP item.

General Health

The General Health (GH) scale in the SF-36 consists of five items, including a general rating of health ("excellent" to "poor") and four items addressing the respondent's views and expectations of his or her health. The SF-12 includes the single general health rating item.

Vitality

The Vitality (VT) scale was developed to capture ratings of energy level and fatigue. The scale consists of four items in the SF-36 and one item in the SF-12.

Social Functioning

The Social Functioning (SF) scale measures the effects of health on the quantity and quality of social activities and, specifically, the impact of either physical or emotional problems on social activities. The SF-36 has two SF items, and the SF-12 has one SF item.

Role-Emotional

The Role-Emotional (RE) scale covers mental health-related role limitations assessing time spent on, level of accomplishment associated with, and level of care in performing work or other usual activities. The SF-36 has three RE items, and the SF-12 has two RE items.

Mental Health

In the SF-36, the Mental Health (MH) scale contains five items—including one or more items from each of four major mental health dimensions (anxiety, depression, loss of behavioral/emotional control, and psychological well-being). The SF-12 contains two MH items assessing depression and psychological well-being.

Physical and Mental Component Summary

The aggregate of the scales are referred to as "component" summaries because they were derived and scored using principal components analysis. The component summary scores capture approximately 85% of the information contained in the scale score profile. Although they reflect the two broad components or aspects of health—physical and mental—*all* items are used to score *both* component summary measures.

SF-36 and SF-12 items, scales, and summary measures are scored so that a higher score indicates a better health state.

Health Transition

The SF-36 includes reported health transition (HT), a general health item that asks respondents to rate the amount of change they experienced in their health in general over a 1-year period. This item is not used to score any of the eight multi-item scales or component summary measures, but provides information about perceived changes in health status that occurred during the year prior to the survey administration. The SF-12 does not include the health transition item.

Health Utility

The SF-6D Health Utilities Index can be calculated from both Version 1 and 2 of the SF-36 and the SF-12. This index (scored on a 0 to 1 range) weights together both physical and mental dimensions of health based on utilities assigned to different health states.

Norm-Based Scoring

The SF-36 originally produced eight scales with scores ranging from 0 to 100 and norm-based PCS and MCS scores. The SF-36v2 and SF-12v2 yield norm-based scores for all eight scales and the two component summaries, easing interpretation and score comparability. Norm-based scoring linearly transforms the scales and summary measures to have a mean of 50 and standard deviation of 10 in the 1998 U.S. general population. Thus, scores above and below 50 are above and below the average, respectively, in the 1998 U.S. general population. Also, because the standard deviation is 10, each 1-point difference or change in scores has a direct interpretation; that is, it is one tenth of a standard deviation or an effect size of 0.10.

Scoring Software

Scoring instructions for the eight scales, PCS and MCS, HT, and an optional Response Consistency Index (RCI) for assessment of data quality are published in the user's manual. Standardized scoring of all SF instruments is available through the QualityMetric Health Outcomes Scoring Software. For respondents with missing data, scale scores can be computed when at least one item in a scale is answered and component scores can be computed in most situations when items from at least seven scales are answered. In addition, the scoring software conducts data quality evaluations (i.e., data completeness, responses outside range, response consistency index [RCI], percentage of estimable scale scores, item internal consistency, item discriminant validity, and scale reliability) and allows users of the SF-36 and SF-12 (Versions 1 and 2) to make direct comparisons of scores across data sets that use different versions of the SF surveys.

Reliability and Validity

Reliability, validity, responsiveness, and interpretation of the SF-36 and SF-12 have been evaluated in numerous studies, which are summarized in several user's manuals and thousands of articles. Much of this research focused on the original survey versions, but because item content was retained across forms, most results from evaluations of the Version 1 surveys generalize to Version 2.

Evaluation of reliability includes internal, alternate forms, and test-retest (Version 1) reliability. For SF-36v2, internal consistency (Cronbach's alpha) estimates using data from the 1998 U.S. general population ranged from .83 to .95 across the eight scales and summary component measures (internal consistency reliability estimates for the summary components take into account the reliability of and covariances among the scales); all exceeding the recommended minimum standard (.70) for group-level comparison of scores. Reliability estimates for general population subgroups and different chronic disease populations are also favorable, and higher for component summary estimates than the eight scales. Studies of alternate forms reliability using the DYNHA item banks found reliabilities ranging from .76 to .93.

Evidence of the tool's *construct validity* has been documented in studies involving factor analysis, item-scale correlations (the correlation between each item and the domain scale scores), interscale correlations, and known-groups comparisons. *Criterion validity* has been demonstrated through

the correlations of each scale with the score for its associated DYNHA item bank. Data on the likelihood of future events (e.g., job loss, psychiatric treatment) based on scale score ranges also provide evidence of criterion validity. *Content validity* has been shown through a comparison of the SF-36v2's coverage of health domains to the health domain coverage of other general health surveys. Validity of the tools is fully documented in the user manuals for the SF-36v2 and SF-12v2 and further documented in peer-reviewed articles by the developer and in numerous studies from the research literature.

Interpretation

Interpretation of research results should preferably be based on prespecified hypotheses. For exploratory purposes or in the case of clinical data, interpretation of the SF-36v2 or SF-12v2 begins by determining if the norm-based scores (NBS) for the PCS and MCS measures deviate from what is considered the "average" range for the U.S. general population or a relevant clinical comparison group. This is followed by an examination of the scale scores to make a similar determination. Each of these decisions is based on separate, empirically based individual patient- and group-level guidelines available in the user manuals. A graphical presentation of the profile (see example in Figure 1) should begin with a presentation of the results of the PCS and MCS measures, emphasizing the importance of first considering findings from these more general measures of health status. For reasons of standardization and ease of interpretation, the eight-scale profile should be presented in the following order: PF, RP, BP, GH, VT, SF, RE, MH.

Empirically derived minimally important differences for group comparisons and evaluation of individual scores are provided in the user manuals, along with score cutoffs for determining the likelihood of the presence of a physical or mental disorder and U.S. general population norms for age, gender, age-by-gender, and combined groups for both the standard and acute forms.

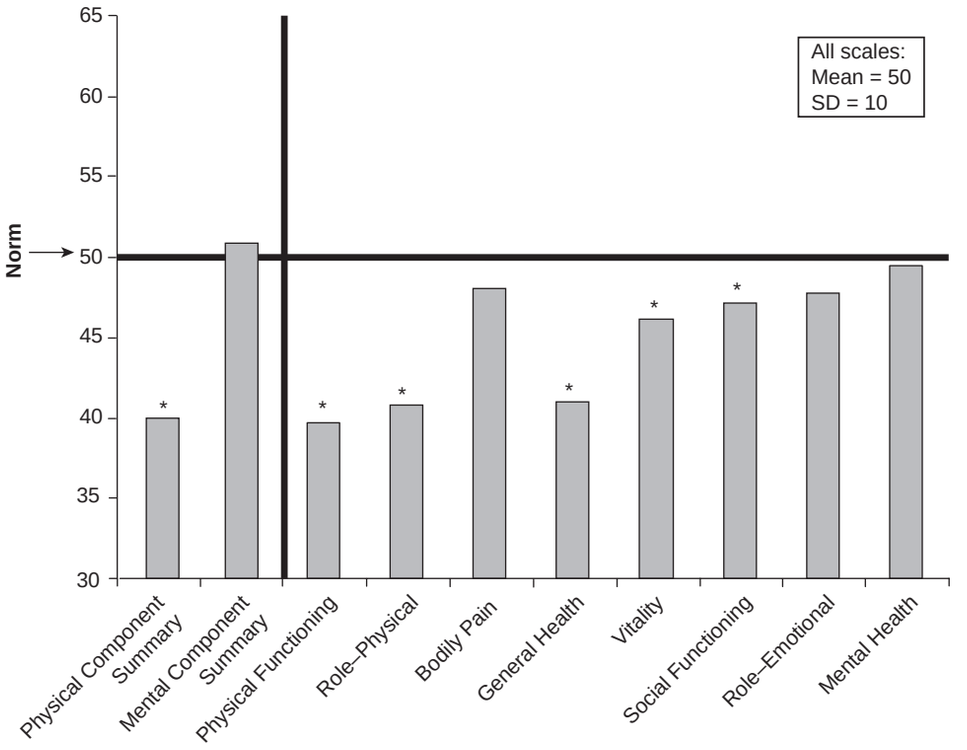


Figure 1 SF-36 health survey profile of norm-based scores (adult asthma sample)

Source: Adapted from Okamoto, L. J., Noonan, M., DeBoisblanc, B. P., & Kellerman, D. J. (1996). Fluticasone propionate improves quality of life in patients with asthma requiring oral corticosteroids. *Annals of Allergy, Asthma and Immunology, 76,* 455–461.

Applications

Applications of the SF-36 (Versions 1 and 2) include

- population monitoring;
- estimating the burden of disease (by standardizing questions, answers, and scoring, reliable and valid comparisons can be made to determine the relative burden of different conditions in several domains of health);
- evaluating treatment effects in clinical trials;
- assessing the cost-effectiveness of products and procedures;
- providing direct-to-consumer information (i.e., educating the public about medical conditions, their symptoms and effects, and potential treatment options; prompting recognition or detection of personal health problems that may benefit from clinical consultation, thereby encouraging more appropriate care seeking, case finding, and physician–patient dialogue; and promoting self-care and compliance with treatment regimens);
- disease management and risk prediction (i.e., the ability to predict health outcomes, hospitalization, future medical expenditures, resource utilization, job loss and work productivity, risk of depression, use of mental healthcare, future health, and mortality);
- enhancing patient–provider relations; and
- clinic-based evaluation and monitoring of individual patients.

The SF-12 (Versions 1 and 2) can be used in population studies for the same purposes, and PCS and MCS scores can be used for risk prediction purposes. However, use of the eight-scale profile for individual patient monitoring is not recommended for the SF-12 or SF-12v2 forms.

A decision to use an SF tool (or any other outcomes tool) should always be preceded by an analysis of whether the tool adequately covers the domains of interest. If not, the SF tool should be supplemented or replaced with an instrument covering these domains. For application in specific clinical groups, the combination of generic surveys (such as the SF-36 and SF-12) and appropriate disease-specific measures is often advantageous. Carrying out a small pilot study using the intended combination of instruments to evaluate readability, acceptability, and response burden is also recommended. In some populations, for example, the elderly, interview administration may be preferable to a paper-and-pencil form. In populations with severe illness or disabilities, some SF domain scales may show minor to moderate floor effects. Again, combination with disease-specific tools can be beneficial.

Resources and Additional Information

Joint copyright for the SF-36® Health Survey, SF-36® Health Survey (Version 2), SF-12® Health Survey, SF-12® Health Survey (Version 2), and SF-8™ Health Survey is held by QualityMetric Incorporated (QM), Medical Outcomes Trust (MOT), and Health Assessment Lab (HAL). Further information about tools from the SF family of instruments is available from the QualityMetric Web site or the sf36.org Web site. The sf36.org Web site is a community forum for users of the SF tools that offers news, events, online discussion, and a searchable database of SF publications. Licensing information for SF tools is also available from QualityMetric (www.qualitymetric.com/products/license). Those conducting unfunded academic research or grant-funded projects may qualify for a discounted license agreement through QM's academic research program, the Office of Grants and Scholarly Research (OGSR).

SF-36®, SF-36v2®, SF-12®, and SF-12v2® are registered trademarks of Medical Outcomes Trust (MOT). DYNHA® is a registered trademark, and SF-8™ and QualityMetric Health Outcomes™ are trademarks of QualityMetric Incorporated.

Jakob B. Bjorner and Diane M. Turner-Bowker

See also Health Outcomes Assessment; Health Status Measurement, Construct Validity; Health Status Measurement, Face and Content Validity; Health Status Measurement, Floor and Ceiling Effects; Health Status Measurement, Generic Versus Condition-Specific Measures; Health Status Measurement, Minimal Clinically Significant Differences, and Anchor Versus Distribution Methods; Health Status Measurement, Reliability and Internal Consistency; Health Status Measurement, Responsiveness and Sensitivity to Change; Health Status Measurement Standards; Scaling; SF-6D

Further Readings

Gandek, B., & Ware, J. E., Jr. (Eds.). (1998). Translating functional health and well-being: International quality of life assessment (IQOLA) project studies of the SF-36 Health Survey. *Journal of Clinical Epidemiology, 51*(11), 891–1214.

QualityMetric Inc.: http://www.qualitymetric.com

SF-36.org: http://www.sf36.org

Ware, J. E., Jr. (2000). SF-36 Health Survey update. *Spine, 25*(24), 3130–3139.

Ware, J. E., Jr., & Kosinski, M. (2001). *SF-36 physical & mental health summary scales: A manual for users of Version 1* (2nd ed.). Lincoln, RI: QualityMetric.

Ware, J. E., Jr., Kosinski, M., Bjorner, J. B., Turner-Bowker, D. M., & Maruish, M. E. (2007). *User's manual for the SF-36™ health survey*. Lincoln, RI: QualityMetric.

Ware, J. E., Jr., Kosinski, M., Turner-Bowker, D. M., & Gandek, B. (2002). *How to score Version 2 of the SF-12® health survey (with a supplement documenting Version 1)*. Lincoln, RI: QualityMetric.

Ware, J. E., Jr., & Sherbourne, C. D. (1992). The MOS 36-Item Short-Form Health Survey (SF-36). I. Conceptual framework and item selection. *Medical Care, 30,* 473–483.

Ware, J. E., Jr., Snow, K. K., Kosinski, M., Gandek, B. (1993). *SF-36® health survey: Manual and interpretation guide*. Boston: Health Institute.

Shared Decision Making

The term *shared decision making* traditionally refers to the communication between a patient and a clinician as they consider a medical decision that involves a choice between two or more clinically reasonable options. These discussions involve exchanging information about medical evidence, about personal values, and about which course of action would be most consistent with patients' preferences once they are informed about the trade-offs among the risks and benefits of each therapy.

This entry outlines the process of shared decision making, the context in which it was originally developed, the roles involved, and the motivations that lead to implementation and research in clinical care. Next, this entry outlines some newly emerging contexts in which the principles of shared decision making have recently been suggested. Based on this background, a historical overview of the range of assumptions and theories underlying the multiple perspectives for considering shared decision making is provided. Finally, the entry highlights various international efforts in the field of shared decision making.

Process

Shared decision making has been defined as an iterative process that educates patients about their healthcare options while facilitating the incorporation of their personal values into medical treatment planning. It may occur in a single conversation between a patient and physician or over multiple visits with a team of healthcare professionals. Patients may choose to include their spouse, family members, friends, clergy, or a legally appointed guardian in their decision-making discussions.

During the process, clinicians may use decision support tools such as patient decision aids. Different from patient education materials, which are used to help patients understand what's involved in undergoing a recommended therapeutic intervention, decision aids are specifically designed to help patients understand and choose between two or more equally relevant options. They do this by providing patients with information about each option's possible benefits, the potential risk of side effects, and the trade-offs between the uncertainties inherent in each option. They also help patients understand that the selection of a particular option is dependent on their informed preferences toward these benefits, risks, and trade-offs. Decision support tools may be as simple as a brochure, as visually engaging as a video, or as complex as an interactive Web site that individually tailors medical information based on a patient's health characteristics. While decision aids by themselves do not constitute full shared decision making, they provide accurate medical information, assistance with the decision-making process, and one type of standardized strategy for introducing the steps of shared decision making into clinical care routines.

Original Context

Shared decision making was developed for medical decisions that were considered "preference sensitive." Therapeutic actions may be categorized

according to the following: (a) the extent to which the scientific evidence about an intervention's clinical effectiveness is clear or not; (b) the ratio between an intervention's potential benefits and harms; and (c) the degree of congruence among patients and clinicians about the desirability or undesirability of those trade-offs. *Effective care* denotes the category of interventions for which there is clear evidence, a high ratio of benefits over harms, and high agreement among clinicians and patients about which is the best therapy. Examples include insulin for diabetes, aspirin and beta blockers for heart attacks, and emergency care for a gunshot wound.

Conversely, the category of *preference-sensitive care* refers to interventions for which the effectiveness evidence is unclear, the risk of harm is considerable, and there is a wide range of opinion among patients about whether the particular intervention is congruent with their beliefs, values, and lifestyle. Examples include some forms of screening (e.g., prenatal genetic tests, the test for prostate cancer), surgical/nonsurgical approaches to symptom management (e.g., surgery vs. physical therapy for osteoarthritis of the knee), aspects of cancer care (e.g., breast-removal vs. breast-conserving surgery for early-stage breast cancer), aspects of end-of-life care, and entry into clinical experiments. In these kinds of preference-sensitive situations, the well-informed patient's attitudes toward the pros and cons of the options under consideration become the deciding factor toward arriving at a choice. Accordingly, the process of shared decision making has been traditionally most strongly advocated in the context of preference-sensitive care.

Roles

It is important to note that the process of honoring a patient's preferences includes revealing and acting on their preferences about participating or not participating in shared decision making in the first place. Patients may comprehend the general concept of choosing between preference-sensitive options for care and state that they prefer to be involved in this kind of decision making. However, as their clinician begins to provide a clear, balanced overview of the stakes involved in choosing between the optional therapies, they may, for various reasons, switch to an informed preference to delegate the responsibility for this particular decision to their clinician. Conversely, well-informed patients may initially prefer to delegate the decision making to others and then switch to taking a more active role in this particular decision as they learn more about the options' pros and cons. Hence, the shared-decision-making model emphasizes clear communication of the nature, content, and scope of the decision while supporting patients' autonomous preferences for participation and choice throughout the decision-making process.

Motivations

There is a range of assumptions and motivations for implementing shared decision making in healthcare.

A Clinical Motivation: Helping Patients With Decisional Conflict

As noted above, the initial motivation for shared decision making stems from the ethical obligation to support patients' autonomous decision making in situations in which no single "best" treatment could be clearly recommended and in which the patient indicates a desire to be involved in arriving at an informed, preference-based choice. However, as a patient begins to engage with the choice, he or she may experience *decisional conflict*.

Decisional conflict is a psychological state experienced when an individual simultaneously wants to accept and reject an uncertain course of action. A functional degree of decisional conflict may encourage effective information search and processing strategies, generate greater insight into one's preferences, and motivate one to resolve a decision dilemma. However, a high level of decisional conflict may be dysfunctional, in that it can interfere with effective decision making and cause considerable distress, ranging from worry to panic.

Decisional conflict can be exacerbated by a range of factors that contribute to the overall uncertainty in the decision situation. Some of these factors are modifiable, including incomplete or misunderstood information; lack of clarity about one's preferences; inadequate social support or inappropriate social pressure; and inadequate material resources to put one's informed, preference-based decision into effect. Therefore, clinicians may be motivated to help patients

experiencing this kind of distress by partnering with the patient in a process of shared decision making that systematically addresses these modifiable factors.

Thus, shared decision making and related decision support can serve as a form of "knowledge therapy" to ameliorate factors that block decision making and engagement in medical care. Successful decision support can parse a seemingly large decision into manageable steps, help clarify the relative value of competing options, and improve realistic expectations of treatment outcomes. In addition, surrogate or proxy decision makers—those who are asked to make a decision on behalf of their parent, spouse, or child—may benefit from guidance during difficult familial discussions. For those facing multiple decisions in chronic care management, shared decision making may foster long-term decision-making skills. Hence, from a clinical perspective, shared decision making may simultaneously be motivated by and used to address psychological and educational blocks to successful healthcare.

A Quality Assurance Motivation: Fostering Patient-Centered Care

A second motivation centers on fostering *patient-centered care*, defined as healthcare that is organized around the patient's individual preferences and needs. Evidence suggests that increased patient-centered care leads to both increased patient satisfaction and better health outcomes; accordingly, the U.S. Institute of Medicine designated the degree of patient-centered care as one indicator of a hospital's overall quality of healthcare. Consequently, various patient-centered quality measures have been developed. Quality "report cards" (such as the Healthcare Effectiveness Data and Information Set, or HEDIS) are used to compare hospitals' performance on important dimensions of patient-centered care and service. The active implementation of preference-based care planning (i.e., the degree to which treatment plans are centered on the expressed preferences of the patient) is one aspect of patient-centered care gauged by these measures. Therefore, some endorse the effort to deliberately incorporate shared decision making into the pathways of preference-sensitive care as one way to ensure that individualized, preference-based care plans are developed and honored.

There exists considerable debate about the desired outcomes of shared decision making. Since preference-sensitive decisions involve choosing among two or more options, there is no single "right" choice. Therefore, the quality of a choice cannot be assessed in terms of the outcomes of that choice; given the uncertain probabilities inherent in these decisions, a patient could make a well-informed choice but experience a poor clinical outcome, or could experience good clinical outcomes even if he or she made a poorly informed choice. Some argue that an effective shared-decision-making process is one that leads to a high-quality choice that is well-informed, congruent with the patient's preferences, and able to be acted on. Decision quality indexes designed to assess the level to which patients' preference-sensitive choices are well-informed, congruent with values, and acted on are under development. Hospital quality assurance programs may be very interested in using such indices to monitor the process and outcomes of any shared-decision-making program that they introduce into their roster of services.

A Health Services Research Motivation: Modifying Unwarranted Variations in Preference-Sensitive Care

In the United States, the rates at which some healthcare services (i.e., numbers of surgeries, screening tests, or medications prescribed for a particular condition) are offered vary from location to location. In areas where more disease exists, increased services are expected; however, some of these variations cannot be explained by differences in the underlying health of the populations, their access to healthcare, or their age, gender, race, ethnicity, level of education, or income. When the healthcare service under consideration is known to be safe and effective, these variations represent disparities in appropriate healthcare that must be addressed.

However, for preference-sensitive healthcare decisions, these variations may be either warranted or unwarranted. If an observed variation accurately reflects the population's informed preferences, the observed variations could be considered warranted. If, however, observed variations actually occur

because the patient population is unaware of the preference-sensitive nature of the decision, has an inaccurate understanding of what's at stake in the decision, is experiencing undue social pressures, has not been offered the opportunity to provide their informed preferences, or has had its expressed preferences disregarded, then the observed variations could be considered unwarranted.

These unwarranted variations, in turn, raise several social, philosophical, and economic issues. Since shared decision making systematically focuses on providing patients with understandable, balanced, up-to-date information and on clarifying patients' preferences, health services researchers suggest that this process may be one strategic intervention for identifying and reducing unwarranted variations in preference-sensitive healthcare.

A Fundamental Science Motivation: Studying Patients' Decision-Making Processes

Decision scientists in healthcare maintain interest in a wide range of fundamental and applied questions about patients' healthcare decision making. The process of shared decision making could serve as a platform for comprehensive research programs addressing these questions. For example, fundamental investigators study the determinants and distributions of patients' preferences for whether to participate in shared decision making; the anxiety or decisional conflict they experience; their information-processing pathways; their comprehension of probabilities; their formulation and reporting of current and anticipated preferential attitudes; or the ways by which they arrive at and act on a particular choice. On the other hand, applied researchers may study the effects of different practice models for implementing shared decision making in clinical practice; the role of different decision aid media (e.g., video vs. Web site); the decision support skills needed by various healthcare practitioners; and the clinician-patient communication patterns that emerge during shared decision making. From either perspective, there are overarching issues about different sociodemographic subgroups of patients (who vary by gender, education, and cultural background) and the different kinds of preference-sensitive decision situations (e.g., decisions about screening, treatment, palliative care, and clinical trial entry) that they face.

Newly Emerging Contexts

As noted above, many interventions in healthcare rest on strong evidence about effectiveness, a high ratio of benefits over harms, and high levels of agreement among clinicians and patients about which treatment is the best therapy. These therapies are considered standards of care, and the clinician seeks the patient's consent to accept these clearly recommended interventions. Recently, some have advocated a role for shared decision making in this arena of effective care, for two reasons. First, some propose that the philosophy and principles underlying shared decision making are also relevant when a patient is asked to consider a recommended, effective therapy. From this perspective, shared decision making could be the process used to help the patient make an informed choice to accept or reject a recommended treatment, with the traditional concept of *informed consent* being replaced by the concept of *informed choice*.

Second, for patients with multiple chronic conditions, some aspects of care are clearly indicated by high-quality clinical evidence, and other aspects rest squarely in the preference-sensitive arena. One could argue that shared decision making is an important strategy in this kind of context, because it's important (a) to integrate patient education materials (designed to help patients understand their practitioners' clearly indicated recommendations for care) with decision support (designed to help patients make values-based informed choices among relevant preference-sensitive care options); (b) to clarify, communicate, and establish patients' priorities for recommended/optional care; and (c) subsequently, to help with the coordinated implementation of these individualized healthcare management plans.

Historical Overview: Assumptions and Theories

Descriptions of shared-decision-making ideas date back to legal debates in the 1950s about physicians' responsibility to obtain informed consent from patients before administering treatment. Through the 1960s and 1970s, theories from sociology, ethics, and quality improvement began to influence these discussions. In 1981, the physician

accommodation model was proposed to provide a framework for analyzing sources of conflict and opportunities for improvement in medicine without the apparent biases of either the traditional paternalistic (physician directed) and the then emerging consumerist (patient directed) medical models. The concept of accommodation emphasized seeking moral certainty in medicine through increased respect for patient autonomy, mutually understood roles, and values communication. It later formed the foundation on which the United States' Preventive Services Task Force and the President's Commission for the Study of Ethical Problems in Medicine published reports advocating a new model of patient empowerment in medical decision making.

Over the following decades, the term evolved to incorporate the perspectives of a variety of disciplines, ranging from social psychology (patient empowerment), communication theory (knowledge therapy), the decision sciences (decision support/patients' decision aids), clinical practice policy (evidence-based medicine), quality improvement (patient-oriented communication), economics (utility assessment), and bioethics/law (informed patient choice). Thus, shared decision making could, in its largest sense, be referring to three ideas: (a) the concept of an ethical obligation to involve patients in making decisions about their healthcare, (b) a model for analyzing physician-patient communication, and (c) the process of physicians and patients engaging in an interaction that involves the mutual sharing of both medical expertise and personal values.

International Applications

Currently, investigative and clinical work in shared decision making is under way in about 10 countries in North America, Europe, Australia, and Asia. This work has been differentially driven by the different philosophical, disciplinary, theoretical, and motivational forces outlined above.

Canada, the United States, and the United Kingdom have developed research and educational programs in shared decision making and have launched demonstration projects to test different practice models for local or broad-based implementation. In North America., private companies and research institutes initially developed decision support tools; government-funded projects are now evaluating cost-effective strategies for delivering these tools in a diverse, multipayer system. Scientific curiosity and information disclosure laws have led to the establishment of shared-decision-making centers, educational training programs, and an inventory of available patients' decision aids. Institutes in the United Kingdom have developed a series of patient decision support tools and are maximizing the use of information technology and financial incentives to widen the scope of implementation.

In Europe, German Ministry of Health initiatives have specifically linked increased patient participation with hospital quality and now support a national research effort to develop physician training programs and patient education programs and the assessment of specific decision aid tools. Patients' legal right to health information provides the stimulus for increasing shared-decision-making policies in France. The Netherlands supports patient empowerment in decision making, along with movements toward increased regulated competition in healthcare. Italian and Australian efforts toward implementing shared-decision-making practice models are in the early stages; the focus is on developing educational programs, condition-specific tools, and public awareness campaigns. As these countries move forward in developing, establishing, and sustaining shared-decision-making programs, the International Patient Decision Aids Standards collaboration continues to compile guidelines for theory- and evidence-based, feasible decision support tools and strategies.

With increasing public interest, an evolving theoretical base, validated decision support tools, and cumulative effectiveness evidence, the idea of shared decision making is gaining greater attention. In addition, debates are unfolding about shared decision making and professional education, legal requirements, financial incentives, reimbursement plans, quality improvement initiatives, and the larger field of the decision sciences.

Aubri S. Rose and Hilary A. Llewellyn-Thomas

See also Decisional Conflict; Decision Quality; Evidence-Based Medicine; Informed Consent; Patient Decision Aids; Utility Assessment Techniques

Further Readings

Charles, C., Gafni, A., & Whelan, T. (1999). Decision-making in the physician-patient encounter: Revisiting the shared treatment decision-making model. *Social Science & Medicine, 49,* 651–661.

Janis, I. L., & Mann, L. (1977). *Decision making.* New York: The Free Press.

Kasper, J. F., Mulley, A. G., & Wennberg, J. E. (1992). Developing shared decision making programs to improve the quality of health care. *Quality Review Bulletin, 1,* 183–190.

O'Connor, A. M., Stacey, D., Rovner, D., Holmes-Rovner, M., Tetroe, J., Llewellyn-Thomas, H., et al. (2001). Decision aids for people facing health treatment or screening decisions. *Cochrane Database of Systematic Reviews, 3,* CD001431.

O'Connor, A. M., Wennberg, J. E., Légaré, F., Llewellyn-Thomas, H. A., Moulton, B., Sepucha, K., et al. (2007). Towards the tipping point: Accelerating the diffusion of decision aids and informed patient choice as a standard of practice. *Health Affairs, 26,* 716–725.

President's Commission for the Study of Ethical Problems in Medicine and Biomedical and Behavioral Research. (1982). *Making health care decisions: The ethical and legal implications of informed consent in the patient-practitioner relationship*. Washington, DC: Author.

Rothert, M., & Talarcyzk, G. J. (1987). Patient compliance and the decision making process of clinicians and patients. *Journal of Compliance in Health Care, 2*, 55–71.

Siegler, M. (1981). Searching for moral certainty in medicine: A proposal for a new model of the doctor-patient encounter. *Bulletin of the New York Academy of Medicine, 57,* 56–69.

Szasz, T. S., & Hollender, M. H. (1956). A contribution to the philosophy of medicine: The basic models of the doctor-patient relationship. *Archives of Internal Medicine, 97,* 585–592.

Tversky, A., & Kahneman, D. (1981). The framing of decisions and the psychology of choice. *Science, 211,* 453–458.

Whitney, S. N., McGuire, J. D., & McCullough, L. B. (2003). A typology of shared decision making, informed consent and simple consent. *Annals of Internal Medicine 140*, 54–59.

Sickness Impact Profile

The Sickness Impact Profile (SIP) is one of the first generic health status measures made available for use in clinical and research settings. It contains 136 items that describe functional limitations in 12 categories: ambulation, mobility, body care and movement, communication, alertness behavior, emotional behavior, social interaction, sleep and rest, eating, work, home management, and recreation and pastimes. Each item is written in the first person, in the present tense, and describes everyday activities of daily living. Respondents endorse only those items that describe their current level of functioning. Users can calculate a total score, two domain scores (physical and psychosocial), or individual scores for each of the 12 categories. Higher scores are reflective of a greater degree of dysfunction. A 68-item short form (SIP68) has been developed; evidence suggests that its psychometric properties are comparable to those of the 136-item SIP.

In the context of medical decision making, the SIP can be used to quantify and compare the physical and psychosocial burden of various medical conditions and treatment modalities. Conclusions based on this type of information can be used by healthcare consumers, providers, and advocates to make informed decisions about medical interventions, funding priorities, and service allocation. The items of the SIP focus on observable behavior; therefore, changes in responses may be evident even if there is no corresponding change in the underlying disease process. The SIP has been used with a variety of medical populations to assess sickness-related dysfunction. This entry reviews the development and validation of the SIP and the SIP68 and the psychometric properties of each and provides an overview of the manner in which the SIP has been used in research in the field of medical decision making.

Development

The authors of the SIP sought to create a behaviorally based measure of sickness-related dysfunction that could be used for evaluation, program planning, policy formation, and cost-effectiveness analysis. Development of the SIP began in 1972 with an effort to collect statements describing impairment in various aspects of functioning from patients, caregivers, healthy individuals, and healthcare professionals. Additional statements were gathered by literature review. A series of field

experiments were conducted to reduce the number of statements, group the items by category, and assign weighted values to each item for scoring purposes. All items were subjected to rigorous methodological evaluation to ensure adequate reliability and validity and sensitivity to change and to assess the comparability of alternative administration procedures. Efforts were made to ensure that individuals representing various levels of illness and all sociodemographic groups were included. The SIP can be self-administered or interviewer administered. It takes approximately 20 to 30 minutes to administer and 5 to 10 minutes to complete the scoring procedures. Scores range from 0 to 100 and are calculated by tallying the weighted values associated with each endorsed item, dividing by the total possible score, and multiplying by 100. The SIP is available for use in several languages/cultures, including British, Chicano-Spanish, Danish, Dutch, French, German, Italian, Norwegian, and Swedish.

The SIP68 was published in 1994 in an effort to overcome one of the major criticisms of the SIP, its length. Other short forms have been developed but are typically specific to a condition of interest (e.g., arthritis, back pain, stroke). Using principal components analysis, researchers determined which SIP items contributed most to respondents' scores. The 68 items that emerged from this analysis comprise six categories: somatic autonomy, mobility control, psychological autonomy and communication, social behavior, emotional stability, and mobility range. These categories can be collapsed into three dimensions: physical, psychological, and social. Scoring of the SIP68 differs from that of the SIP in that the items are not differentially weighted. Total, subscale, and dimension scores are calculated by summing the endorsed items. Total scores range from 0 to 68; ranges for subscale and dimension scores depend on the number of items per scale. Because of divergent scoring procedures, the two questionnaires cannot be directly compared.

Psychometric Properties

The psychometric properties of the SIP and the SIP68 have been thoroughly investigated. Test-retest reliability, or temporal stability, for the overall SIP, its dimensions, and its categories is considered to be good to very good (r = .45 to .60). Estimates of internal consistency, the degree to which items are related to one another, for each of the scores listed above are also strong (α = .60 to .95). Estimates of test-retest reliability and internal consistency for the SIP68 are in the same range as those reported for the SIP. Overall, reliability estimates for the individual categories are somewhat weaker than those found for the total score and dimension scores.

The validity of the SIP has been examined in a number of ways. Criterion validity, the degree to which an instrument correlates with an accepted gold standard, has been demonstrated by expected associations with other well-known and often used functional status measures. Construct validity can be demonstrated by evaluating the pattern of relationships between a newly developed measure and existing measures that it should or should not be related to according to prior theory. Correlations between the SIP total score and other existing measures are generally fairly low; however, expected patterns of relationships emerge when examining the relationship between the two dimensions of the SIP and related measures of physical and psychosocial functioning. Only a few studies could be identified that have attempted to replicate the proposed internal factor structure of the SIP. This preliminary evidence suggests that only portions of the proposed factor could be reproduced; therefore, additional research should be conducted to provide further validation for the underlying constructs.

Construct validity of the SIP68 was evaluated by examining the pattern of relationships among the six subscales. Results demonstrated good construct validity in that subscales conceptually similar to one another were more highly correlated than subscales containing dissimilar item content. As evidence of criterion validity, scores on the SIP68 were compared with measures of self-care ability, life satisfaction, and level of spinal cord injury. Expected patterns of relationships were reproduced to a satisfactory degree. The SIP68 was not validated in a U.S. sample until 2003. Although it was shown to be a reliable and valid abbreviated version, the proposed factor structure and item loadings were not confirmed by the factor analysis conducted in this sample.

Because the SIP is often used to evaluate functional outcomes in longitudinal research, it is

critical for the instrument to be able to detect clinically significant change over time. The responsiveness of the SIP and SIP68 has been evaluated; however, further research is needed in this area. It has been noted that the broad nature of the SIP makes it difficult to score at the maximum level, which may make it more likely to show deterioration than to detect functional improvement. Overall, results have indicated that the SIP and SIP68 are equally responsive to identifying significant change in health-related functional status. The populations typically assessed with the SIP are generally diagnosed with some form of chronic medical condition; therefore, small changes are often observed.

Limitations

Although there are clearly strengths of the SIP, such as strong psychometric qualities and a large body of supporting literature, it is not without weaknesses. Very little research has been conducted to verify the proposed factor structure of either version of the SIP; additional research in this area would strengthen the existing evidence for construct validity of this measure. Additionally, researchers have noted that scoring procedures can lead to inconsistent, and in some cases, illogical results. It is theoretically possible for an individual with a minor impairment to earn a score reflective of greater dysfunction than someone with a more profound impairment. Item content and order are another concern. Some items, particularly those in the mobility category, are mutually exclusive, making it impossible to attain the maximum category score. For example, a respondent cannot logically endorse both "*I do not walk at all*" and "*I walk more slowly*." These two items also speak to the ordering of the items, which can become redundant and potentially frustrating for patients who are asked to respond to logically discrepant questions about mobility. The work category has been criticized for not accurately representing those who do not work for non-health-related reasons (e.g., unemployed, retired). To rectify these problems, researchers have evaluated alternative administration and scoring procedures designed to enhance clarity of items and interpretability of scores and to reduce respondent burden.

Use in Medical Decision Making

When faced with a medical decision, one of the factors likely to be considered by patients, physicians, and policy makers alike is the impact of the intervention on functional status. The SIP has been used to evaluate this type of health outcome in countless studies over the past 30 years. It has proven to be a useful evaluation tool among patients with a wide range of acute and chronic health conditions such as cardiovascular disorders, cancer, neurologic conditions (spinal cord injury, dementia), pain conditions (arthritis, back pain), end-stage renal disease, and diabetes. This instrument can be used to quantify the psychosocial and physical burden of various healthcare conditions as well as the impact of medical intervention on these conditions. Although limitations have been noted regarding the content and scoring of the SIP, recent efforts have been made to develop modified methods of scoring and administration. Taking these factors into consideration, the SIP and its short form should be considered as one of the primary options for assessing functional health status in any population.

Erin Winters Ulloa

See also Health Outcomes Assessment; Health Status Measurement, Generic Versus Condition-Specific Measures; SF-36 and SF-12 Health Surveys

Further Readings

Bergner, M., Bobbitt, R. A., Carter, W. B., & Gilson, B. S. (1981). The Sickness Impact Profile: Development and final revision of a health status measure. *Medical Care, 19*, 787–805.

deBruin, A. F., deWitte, L. P., Stevens, F., & Diederiks, J. P. M. (1992). Sickness Impact Profile: The state of the art of a generic functional status measure. *Social Science & Medicine, 35,* 1003–1014.

deBruin, A. F., Diederiks, J. P. M., deWitte, L. P., Stevens, F. C. J., & Philipsen, H. (1994). The development of a short generic version of the Sickness Impact Profile. *Journal of Clinical Epidemiology, 47,* 407–418.

deBruin, A. F., Diederiks, J. P. M., deWitte, L. P., Stevens, F. C. J., & Philipsen, H. (1997). Assessing the responsiveness of a functional status measure: The Sickness Impact Profile versus the SIP68. *Journal of Clinical Epidemiology, 50,* 529–540.

Nanda, U., McLendon, P. M., Andresen, E. M., & Armbrecht, E. (2003). The SIP68: An abbreviated sickness impact profile for disability outcomes research. *Quality of Life Research, 12,* 583–595.

Pollard, B., & Johnston, M. (2001). Problems with the Sickness Impact Profile: A theoretically based analysis and a proposal for a new method of implementation and scoring. *Social Science and Medicine, 52,* 921–934.

SMARTS AND SMARTER

SMARTS (Simple Multi-Attribute Rating Technique Using Swings) and SMARTER (Simple Multi-Attribute Rating Technique Exploiting Ranks) are two prescriptive techniques for making choices under certainty between options evaluated on multiple attributes proposed by Edwards and Barron in 1994 to replace the original SMART proposed by Edwards in 1977. Each assumes a weighted, additive, multi-attribute utility model, and each seeks to simplify the operations necessary to estimate the multi-attribute utility of each option under consideration. The techniques differ primarily in the procedure for weighting the importance of attributes.

Consider a headache sufferer making a choice between three pain relievers, each of which has a different level of provided relief, duration of relief, and potential for side effects. For example, Pain reliever A provides excellent relief for 2 hours with rare side effects, Pain reliever B provides good relief for 4 hours with rare side effects, and Pain reliever C provides limited relief for 12 hours with no side effects. How should the patient choose between these options?

SMART

In the SMART technique, the decision maker directly assesses the values of the choice options on each attribute rather than performing a (large) series of choices between hypothetical alternatives from which attribute values are inferred, which was characteristic of earlier approaches. Edwards and Barron refer to this simplification as "the strategy of heroic approximation." It results in a substantially shorter assessment procedure. After identifying the decision purpose, decision makers, value structure for the decision, and choice options, the analyst constructs an option-by-attribute matrix and directly assigns single-attribute utility values to options in this matrix. An example of such a matrix appears in Table 1. Subjective values are assigned to each attribute by the decision maker on a scale from 0 (*worst*) to 100 (*best*).

Dominated options, which outperform another option on all attributes, are then removed. Mathematical tests of the value structure may be performed to confirm that it meets assumptions required of additive models (e.g., tests for conditional monotonicity). If an additive model is to be assumed, attributes are then weighted to reflect their relative importance. In SMART, these weights are derived by asking the decision maker to judge the ratio of the importance of each attribute to all others. Finally, the multi-attribute utility is computed by summing the product of attribute weight and attribute value for each attribute for each option and selecting the option that maximizes the

Table 1 An option-by-attribute matrix for pain relievers

Option	*Attribute*					
	Level of Relief	*Level of Relief (Value)*	*Duration of Relief (hr)*	*Duration of Relief (Value)*	*Side Effects*	*Side Effects (Value)*
A	Excellent	100	2	0	Rare	0
B	Good	50	4	30	Rare	0
C	Limited	0	12	100	None	100

sum. For example, if the attributes of level of relief, duration of relief, and side effects had relative weights of .5, .3, and .2, respectively, the multiattribute utility of pain reliever A would be $.5 \times 100 + .3 \times 0 + .2 \times 0 = 50$. Similarly, the multiattribute utilities of B and C would be 34 and 50, respectively. The decision maker should be indifferent between A and C and prefer either to B.

SMARTS

The SMARTS technique corrects an error in the process of assigning attribute weights that was present in SMART. This is easily illustrated in our example by considering the side effects attribute. While side effects might seem to warrant a large relative weight in the abstract, in this particular decision, the range of side effect values is from "none" to "rare." That is, the meaning of a 100 point change in the side effect attribute is not very large in this context.

In the SMARTS approach, importance weights of attributes are determined by a two-step swing weighting process. In the first step, the rank orders of the attributes are determined by asking the decision maker to consider a hypothetical alternative with the lowest value on all attributes. The decision maker then indicates which single attribute she or he would prefer to change from the lowest value (normalized to 0) to the highest value (normalized to 100). For example, would the decision maker faced with a pain reliever that has limited relief for 2 hours with rare side effects prefer to improve the level of relief to "excellent," the duration to "12 hours," or the side effects to "none"? The selected attribute is noted as the most important attribute. The procedure is repeated with the remaining attributes to establish a rank order. Then, using the rank-ordered attributes, weights are determined by asking the decision maker to make a series of magnitude estimations comparing the impact of a 100 point (lowest to highest) swing in the most important attribute to the impact of a 100 point swing in less important attributes (or by making a series of indifference judgments). In this example, the weight of a swing in side effects from "rare" to "none" might be relatively low. With weights of .70, .25, and .05 for level of relief, duration of relief, and side effects, respectively, Pain reliever A would be preferred to B and B to C.

SMARTER

The SMARTER technique proceeds like SMARTS up through and including the first step of the attribute weighting procedure, which results in a rank order of attributes by importance. At this point, the weights of the attributes are simply computed directly from the ranks using a procedure called rank order centroid (ROC) weights (not to be confused with receiver operating characteristic curves). ROC weights are those weights that minimize the error of estimation of the weights with only rank information. For example, given two ordered attributes, the space of possible weights in which the first attribute is more highly weighted than the second has centroids at weights of .75 and .25. That is, if all possible combinations of weights are considered equally possible, the average (and hence least squares error minimizing) weights would be .75 and .25. The formula for the weight of the kth attribute in a set of K attributes is given by Edwards and Barron as

$$w_k = \frac{1}{K}\sum_{i=k}^{K}\frac{1}{i}.$$

They also provide a table of ROC weights for different numbers of attributes, which is reproduced in part as Table 2. Columns show different numbers of attributes of evaluation; the weights listed in the table cells under each column are assigned to the highest ranked, next highest ranked, and so on, attribute. In our example, with three attributes, the ROC weights for level of relief, duration of relief, and side effects (assuming a descending order of importance) would be .61, .28, and .11; Pain reliever A would be preferred over either B or C, which would be nearly identical in utility. Recent simulations by Ahn and Park suggest that ROC weights regularly outperform other methods of estimating decision weights from ordinal information alone.

Because SMARTER uses only approximates of the weights that would be assigned through a decision-specific procedure (like SMARTS), it may result in less optimal decisions. Studies reported by Edwards and Barron, however, suggest that value loss is both rare and, when it occurs, tends to be small. Because SMARTER is considerably less time-consuming for decision makers, however, and is a substantial improvement over making decisions

Table 2 SMARTER attribute weights

Rank	*Number of Attributes*							
	2	3	4	5	6	7	8	9
1	.750	.611	.521	.457	.408	.370	.340	.314
2	.250	.278	.271	.257	.242	.228	.215	.203
3		.111	.146	.157	.158	.156	.152	.148
4			.063	.090	.103	.109	.111	.111
5				.040	.061	.073	.079	.083
6					.028	.044	.054	.061
7						.020	.034	.042
8							.016	.026
9								.012

Source: Edwards, W., & Barron, F. H. (1994). SMARTS and SMARTER: Improved simple methods for multiattribute utility measurement. *Organizational Behavior and Human Decision Processes 60,* 306–325.

with no systematic multi-attribute approach, it may have wide utility in cases where time or other factors do not permit a full multi-attribute utility assessment.

Alan Schwartz

See also Multi-Attribute Utility Theory

Further Readings

Ahn, B. S., & Park, K. S. (2008). Comparing methods for multiattribute decision making with ordinal weights. *Computers & Operations Research, 35*(5), 1660–1670.

Barron, F. H., & Barrett, B. E. (1996). Decision quality using ranked attribute weights. *Management Science, 42*(11), 1515–1523.

Edwards, W. (1977). How to use multiattribute utility measurement for social decision making. *IEEE Transactions on Systems, Man, Cybernetics, SMC-7,* 326–340.

Edwards, W., & Barron, F. H. (1994). SMARTS and SMARTER: Improved simple methods for multiattribute utility measurement. *Organizational Behavior and Human Decision Processes, 60,* 306–325.

Social Factors

Sociologists bring a perspective and unique methods that broaden our understanding of clinical decision making and complement the valuable work of other disciplines. This entry briefly discusses four propositions that are designed to improve our understanding of clinical decision making and increase the relevance of decision making studies for social policies intended to improve public health.

Macro Determinants of Clinical Decision Making

Sociologist Talcott Parsons provided a theoretical perspective on the doctor-patient relationship (D-P relationship), reflecting the situation in the United States around the middle of the 20th century; his view dominated thinking in health services research for the remainder of the century. The Parsonian view of the D-P relationship is depicted in Figure 1, and several features should be highlighted:

- The spotlight of analysis was generally on the doctor, who enjoyed high social status and had a

dominant role as the repository of valuable medical knowledge and expertise.
- Only two actors were involved in the interaction, with the doctor acting professionally and being altruistically motivated to serve only the patient's interests.
- The patient occupied a subordinate and reciprocal role and was expected to trust the doctor's judgment and follow (his) clinical recommendations (*credat emptor* was the prevailing ethos).

Many health services researchers and decision theorists still employ this idealistic perspective, even though the D-P relationship today (within which clinical decisions occur) bears little resemblance to earlier formulations. The world of healthcare has shifted beneath decision theorists' feet, producing results of little policy relevance and suggesting educational efforts that are unlikely to produce desired changes in clinical practice. Some indication of the magnitude of the transformation of U.S. healthcare is evident in the words used to describe the once special D-P relationship—the doctor has become "a provider," the patient is now a "client," and the relationship is now considered "an encounter."

Some of the major new influences affecting clinical decision making within the new client-provider encounter (C-P Encounter) are illustrated in Figure 2 and include the following:

- Physicians are increasingly forced into specialization (generalists are in short supply) and most are full-time salaried employees in large and increasingly concentrated organizations.
- Corporatized physician employees are required to go along (with clinical guidelines and pay-for-performance schemes) if they are to get along (receive promotions and salary improvements).
- The spotlight is now on a knowledge-empowered patient/client who occupies center stage and is the ultimate object of all revenue in a profit-driven healthcare system.
- Insurance companies dictate what exactly any clinician can actually decide for any given case (test ordering, referrals, prescriptions, and follow-up).
- Pharmaceutical companies advertise directly to consumers and suggest that they should ask their providers for specific medications. "Doctor knows best" is no longer the prevailing viewpoint.
- Widely publicized reports of financial kickbacks and clinical malpractice, as well as recognition that doctors may now serve several masters, appear to have eroded trust in the profession of medicine (as with car repairs, caveat emptor is the emerging ethos). The popular media no longer portray doctors as cultural heroes—compare *Marcus Welby, MD* with today's *House* or *Green Wing*.

Unlike the one-on-one, closed-system relationship of the past, clinical decision making is now required to occur on an ever more crowded stage (Figure 2). The spotlight is now on the patient as an object of revenue (client-centered care for an objectified condition); the patient is digitally empowered and activated by private (mainly pharmaceutical) interests; while supposedly making decisions solely in the interest of the patient, doctors are now required to also serve their corporate employers (there is no guaranteed coincidence of interest); the possible range of clinical actions and costs is dictated by a patient's health insurance (assuming that they have such); the state (government) is now essentially an onlooker, unwilling to protect the prerogatives of doctors and concerned to reduce the burden of ever-increasing healthcare costs. Surveys reveal high levels of physician dissatisfaction with their workplace and complaints about administrative encroachments on clinical autonomy.

Much research and thinking on clinical decision making appears to overlook the macrosociological influences that now shape everyday clinical decision making. Clinical decision making increasingly occurs on a stage where doctors are no longer the leading actors. A more sociological approach to clinical decision making recognizes that fundamental changes have occurred in U.S. healthcare over the past several decades and that large organizations and institutions have now assumed leading roles and shape any decisions a provider is required to make.

Circularity of Bayesian Reasoning

Sociologists developed the concept of a "self-fulfilling prophecy" to describe the process by

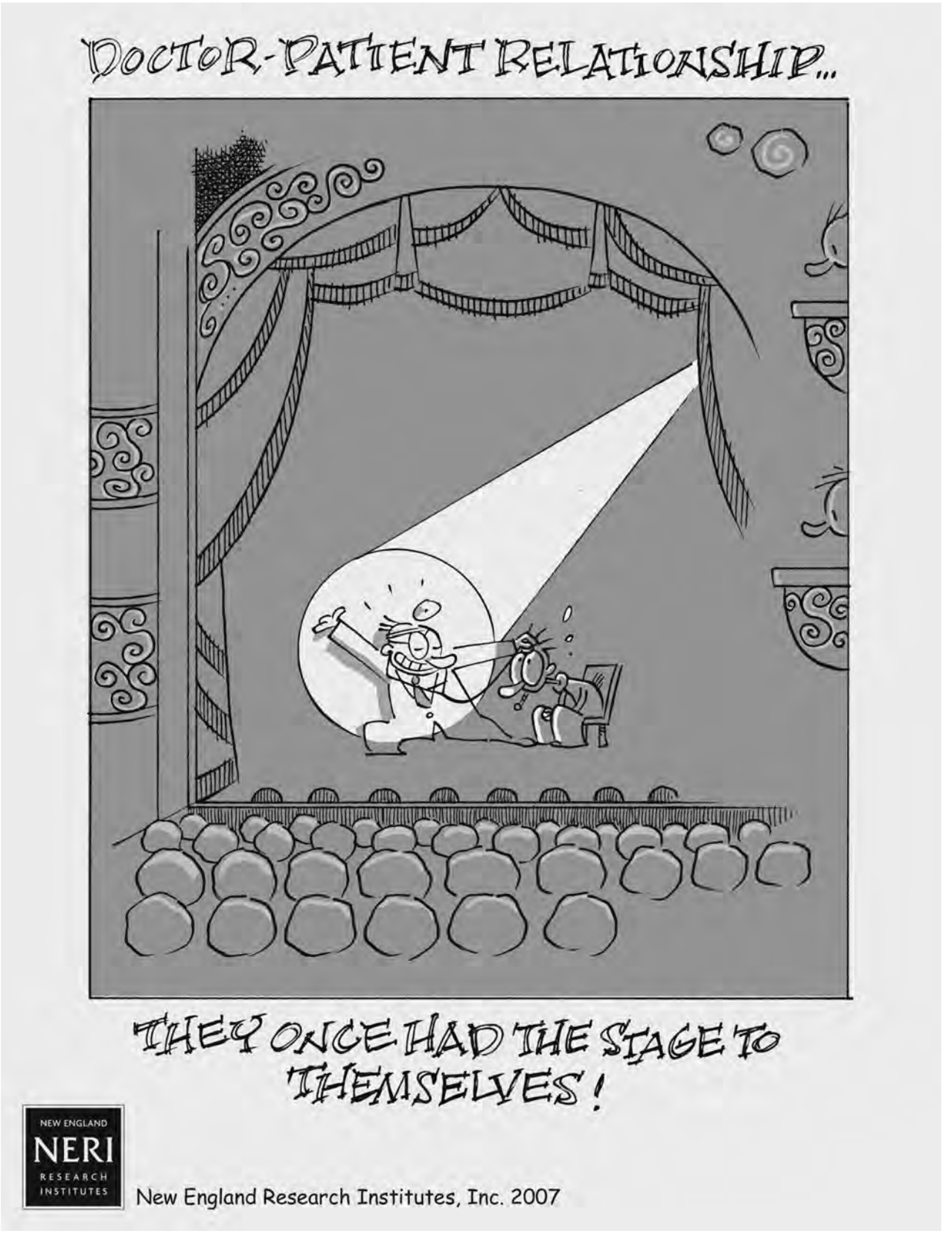


Figure 1 Doctor-patient relationships . . . they once had the stage to themselves

Source: New England Research Institutes, Inc. Reprinted with permission.

Figure 2 Patient-centered 21st-century medicine . . . the stage is now so crowded

Source: New England Research Institutes, Inc. Reprinted with permission.

which phenomena that are perceived to be real eventually become real in their consequences. With respect to clinical decision making, if a physician believes that particular patients are more likely to have *X*, then such patients are more likely to be diagnosed with and treated for *X*, irrespective of the actual signs and symptoms presented. This self-fulfilling process can be demonstrated using the prevailing Bayesian approach to clinical decision making. Bayes's rule can be stated as follows:

$$p(\theta|y) = p(\theta, y)/p(y) = p(\theta)p(y|\theta)/p(y) \propto p(\theta)p(y|\theta), \quad (1)$$

where

θ is some disease,

y is the clinical presentation of a patient, and

$p(\cdot)$ is a probability density function.

In other words, the posterior probability of the disease θ, given the clinical presentation of the patient, $p(\theta|y)$, is proportional ($\propto$) to the prior probability of the disease $p(\theta)$ times the probability of the clinical presentation given the disease, $p(y|\theta)$. Some might wish to take into consideration patient characteristics z. Equation 1 can then be written as

$$\begin{aligned} p(\theta|y, z) &= p(\theta, y|z)/p(y|z) \\ &= p(\theta|z)\ p(y|\theta, z)/p(y|z) \\ &\propto p(\theta|z)p(y|\theta, z). \end{aligned} \quad (2)$$

Suppose we make an important assumption concerning $p(y|\theta, z)$ and $p(y|z)$: that they *only vary by clinical signs and symptoms*; that is, $p(y|\theta, z) = p(y|\theta)$ and $p(y|z) = p(y)$ for all z. In other words, the clinical presentation (y) given the disease θ ($y|\theta$), or by itself (y), does not vary by nonclinical patient characteristics such as age, gender, and socioeconomic status (possible components of z).

$$\begin{aligned} p(\theta|y, z) &= p(\theta|z)p(y|\theta, z)/p(y|z) \\ &= p(\theta|z)p(y|\theta)/p(y). \end{aligned} \quad (3)$$

Let us now consider the theoretical consequences if the clinical presentation (y) is invariant or consistent. Suppose we have two patients, each with values of the vector z, z_1, and z_2. Any variability in the posterior diagnosis in each patient resulting from the same clinical presentation (y) given the disease θ is

$$\begin{aligned} p(\theta|y, z_1)/p(\theta|y, z_2) &= (p(\theta|z_1)p(y|\theta, z_1)/p(y|z_1))/ \\ &\quad (p(\theta|z_2)p(y|\theta, z_2)/p(y|z_2)) \\ &= (p(\theta|z_1)p(y|\theta)/p(y))/ \\ &\quad (p(\theta|z_2)\ p(y|\theta)/p(y)) \\ &= p(\theta|z_1)/p(\theta|z_2). \end{aligned} \quad (4)$$

In other words, if the clinical presentation is invariant, it should, theoretically, add no useful information to the diagnostic process, and any diagnostic variability observed should only reflect the prior $p(\theta|z)$. This is consistent with Bayesian logic and demonstrates the circularity of Bayesian reasoning: Namely, the variability in the posterior distribution, $p(\theta|y, z)$, is the same as in the prior distribution, $p(\theta|z)$.

Now, if we, experimentally, standardize the clinical presentation so that it does not vary between patient encounters, we would, according to the Bayesian approach, expect to observe only the prior distribution, $p(\theta|z)$, of the disease.

McKinlay and colleagues ran three factorial experiments, using vignettes of patients presenting symptoms suggestive of coronary heart disease (CHD) or diabetes, in which it is possible to estimate the unconfounded effects of patient gender, age, race/ethnicity, and socioeconomic status. For these experiments, the clinical presentation ($y|\theta$) is *identical* regardless of the nonclinical patient characteristics. The authors observed the posterior probability of the disease given the clinical presentation (which is theoretically, according to the above Bayesian argument, proportional to the unobserved prior probability of the disease). The experiment was run twice for the CHD vignette; and although 95% of physicians gave a CHD diagnosis, the certainty of the CHD diagnosis *did* vary by patient characteristics (Table 1), reflecting the physician's unobserved prior probabilities. For a separate experiment focusing on diabetes, the probability of a diabetes diagnosis also varied by patient characteristics (Table 1).

It would appear from these results that the physicians in the experiments are (if they are being good Bayesians) using priors based on currently

Table 1 Results from three factorial experiments—patients presenting with symptoms suggestive of coronary heart disease (CHD) or diabetes

	CHD—Average Certainty (0–100)				*Diabetes—Correct Diagnosis (%)*	
		p		*p*		*p*
Location	MA		NC/SC		NJ/NY/PA	
Sample size	128		256		192	
Patient gender		.0476		.0117		.1528
Male	62.3		61.7		65.6	
Female	53.5		53.0		56.2	
Patient age[a]		.7887		<.0001		.6320
Younger	57.3		51.9		62.5	
Older	58.5		62.8		59.4	
Patient race/ethnicity		.0028		.4963		.0092
Black	51.2		58.3		73.4	
Hispanic	—		—		60.9	
White	60.3		56.4		48.4	
Patient SES[b]		.2717		.2842		.2651
Lower	55.5		55.9		64.6	
Upper	60.3		58.8		57.3	

Sources: New England Research Institutes. (2007). *CHD clinical decisions in older patients* (AG16747). Watertown, MA: Author. New England Research Institutes. (2007). *Cognitive basis of CHD disparities* (HL079174). Watertown, MA: Author. New England Research Institutes. (2007). *Diabetes: Race and ethnic disparities in diabetes* (DK066425). Watertown, MA: Author.

a. For the CHD vignette, the patient ages were 55 or 75. For the Diabetes vignette, the patient ages were 35 or 65.

b. The patient socioeconomic status (SES) is depicted by current/former occupation. For the CHD vignette, the occupations were janitor or schoolteacher. For the Diabetes vignette, the occupations were janitor or lawyer.

available epidemiologic data—where heart disease is less common in women than men and diabetes is less common in whites than minorities. However, the same epidemiological data show that heart disease is more common in blacks than in whites, which is not consistent with these experiments. And the question remains, is it race/ethnicity or is it socioeconomic status (since in the United States these constructs are confounded, as compared with the experimental situation, in which they are not confounded). However, these epidemiologic rates are themselves the result of the circular reasoning presented here. If physicians do not test for condition *X*, they cannot tell a patient that they have been diagnosed with *X*. The question then becomes, "Is Bayesian reasoning a suitable model for clinical decision making?" When presenting symptoms to a physician, patients should expect to

be diagnosed by their signs and symptoms, not by who they are (gender, age, race/ethnicity, or socioeconomic status).

Contribution of Clinical Decision Making to the Generation and Amplification of Disparities

Beyond the circularity inherent in Bayesian reasoning, physicians' clinical decision making also contributes to the generation and amplification of health disparities. Official epidemiologic base rates are assumed to reflect the "real" underlying prevalence of biologic phenomena; and indeed, the accuracy of the Bayesian processing described above relies on this assumption. As part of a long research tradition, however, sociologists have shown how such rates are socially constructed so that the process of applying labels to cases is contingent on socially negotiated activities. For example, the sociologist Emile Durkheim explained the low official suicide rates in Ireland when he discovered that coroners were less likely to designate suicide as a cause of death in a Catholic country, resulting in aggregate rates that were lower than neighboring countries. Similarly, epidemiologic base rates are the sum of "cases" as they are labeled during individual doctor-patient encounters.

This process is depicted in Figure 3. At the center of the picture is a physician-policeman who is directing traffic consisting of presenting patients who need to be sorted according to their disease category. The physician directs approaching patients to follow one of three routes: heart disease, diabetes, or emotional disorders. As the cats in the picture comment, this sorting process is not based just on "what they have" but on "who they are," with heavy white men being sorted into the heart disease category, the black men sorted into the diabetes group, and the heavy white woman receiving an emotional disorders diagnosis. Once patients are sorted into these categories, they become "cases" and are counted toward the official epidemiologic rates for each type of disease. Based on the Bayesian model outlined above, these rates are reified as they are used as a basis for further sorting of new patients. Therefore, to the extent that this process of sorting patients into diagnostic categories is biased, physician decision making contributes to and amplifies existing health disparities.

This point is readily illustrated with data from McKinlay and colleagues' factorial experiment studies of clinical decision making (Table 1). The results show that physicians' decisions are often inconsistent with observed prevalence rates. These biases are critical for the process of sorting patients into disease categories. For example, if physicians are cognitively predisposed toward assuming that younger women are at low risk for coronary heart disease, then they are less likely to entertain that diagnosis and test for the condition. If there is no diagnostic label or test confirmation, then it does not exist as a "case" to be counted in prevalence rates. As a result, rates reify preexisting *cognitive assumptions* rather than reflecting the *true prevalence* of the condition. Similarly for diabetes: If physicians are mistakenly inclined to assume that white patients are at lower risk than blacks, they are less likely to test for and identify the condition in those patients, independent of the presenting symptoms.

Factorial experiments conducted for over a decade have systematically manipulated the characteristics of the patients (including gender, age, race/ethnicity, and socioeconomic status); physicians (gender and level of experience); the healthcare system (the United States, the United Kingdom, and Germany); and cognitive processing (priming physicians to consider a specific diagnosis) to assess how diagnostic and treatment decisions vary even when the presentation of symptoms was identical. These studies have replicated the existence of such biases in clinical decision making not only across study generations but also across conditions depicted in the vignettes (CHD, depression, diabetes, breast cancer). Considered in combination with Bayesian reliance on epidemiologic base rates for establishing prior probabilities of disease, the biases observed in clinical decision making have far-reaching implications and point to targets for policy interventions to minimize the amplification of disparities.

The Importance of Diverse Methodologies

A corollary to the blind men studying different parts of the elephant from their respective disciplinary perspectives is the tendency for each blind man to also use a favored research method. In the same way that different disciplines tend to ask different

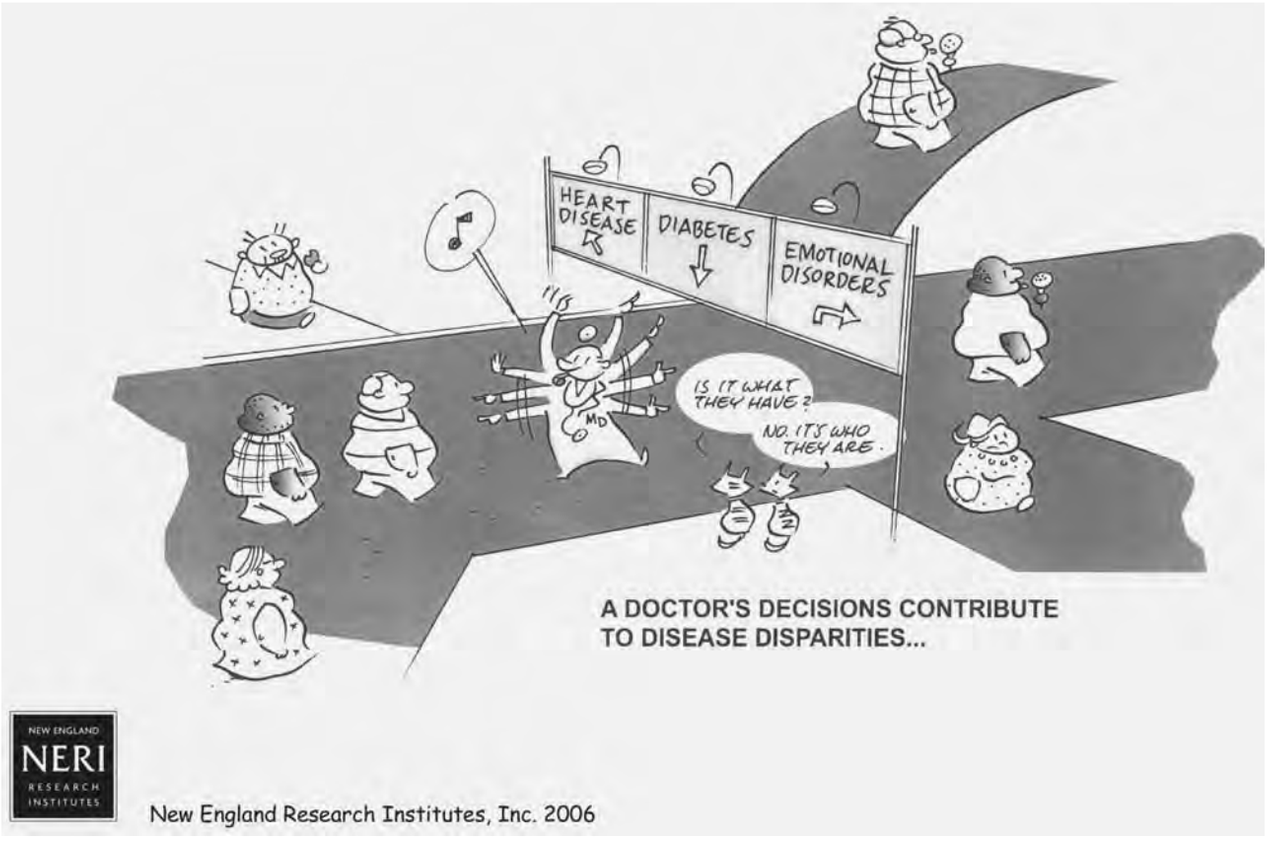


Figure 3 A doctor's decisions contribute to disease disparities

Source: New England Research Institutes, Inc. Reprinted with permission.

types of research questions, they also tend to employ methods especially well suited to their interests. As a result, methodological differences tend to reinforce disciplinary separations. Overwhelmingly in clinical decision making research, these are quantitative approaches and focus on aggregate associations.

These studies tend to describe which decisions doctors make using survey methods and convenience samples obtained in large medical centers with adequate numbers of physicians. These studies obviously have limited generalizability. Furthermore, even the most sophisticated multivariate analyses of this type of data (often collected for administrative purposes) cannot disentangle the influence of associated characteristics (e.g., patient race and socioeconomic status). This problem is compounded by uncontrolled variation in patients' presentations, a complication associated with the common use of medical records data. This lack of internal validity compromises the process of concretely identifying sources of variation in clinical decision making. The factorial experimentation studies discussed in this entry illustrate an alternative approach that allows for unconfounded estimates of the influence of particular factors, standardized presentation of patient symptoms, and generalizability beyond convenience samples. To improve our knowledge in this field, there need to be increased methodological diversity and incisive quantitative approaches that overcome these challenges.

Still missing from much of the work to date, however, is knowledge about the cognitive processing behind observed aggregate associations: Why do physicians make certain decisions, and how do they process available information? To answer these types of questions, clinical decision-making research must move beyond quantitative approaches to also incorporate qualitative methods. While qualitative approaches are sometimes cavalierly dismissed as "soft" compared with statistical approaches, they are often more concerned

with validity and reliability than their quantitative counterparts and therefore make a unique and needed contribution. Examination of actual behavior, or directly eliciting physicians' explanations of how they process information, fills gaps in our knowledge that are otherwise left to speculation. In-depth interviewing, ethnography, think-aloud protocols, and conversation analysis have rich traditions not only in sociology but also in anthropology, linguistics, and social psychology. These approaches allow for the collection and analysis of detailed information grounded directly in the perspectives of physicians themselves ("emic" perspective) rather than those of the researchers who study them ("etic" perspective). Such approaches can be used alone or in combination with quantitative work. The factorial experiments integrate an open-ended think-aloud segment into a more structured interview questionnaire, allowing physicians to explain in depth, and in their own words, how they arrived at their clinical assessments of the patients in the vignettes. This type of methodological diversity can be critical for expanding the set of research questions under consideration to include the "whys" of clinical decision making.

Karen E. Lutfey, Carol L. Link, Lisa D. Marceau, and John B. McKinlay

See also **Social Judgment Theory**

Further Readings

American Diabetes Association: http://www.diabetes.org/diabetes-statistics/prevalence.jsp

American Heart Association: http://www.americanheart.org/presenter.jhtml?identifier=4478

Durkheim, É. (1997). *Suicide*. New York: The Free Press. (Original work published 1897)

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2004). *Bayesian data analysis* (2nd ed.). Boca Raton, FL: Chapman & Hall.

McKinlay, J. B., & Link, C. L. (2007). Measuring the urologic iceberg: Design and implementation of the Boston Area Community Health (BACH) Survey. *European Urology, 52*(2), 389–396.

McKinlay, J., Link, C., Arber, S., Marceau, L., O'Donnell, A., & Adams, A. (2006). How do doctors in different countries manage the same patient? Results of a factorial experiment. *Health Services Research, 41*(6), 2182–2200.

McKinlay, J., Potter, D., & Feldman, H. (1996). Non-medical influences on medical decision-making. *Social Science & Medicine, 42*, 769–776.

Parsons, T. (1951). *The social system*. New York: The Free Press.

Thomas, W. I., & Znaniecki, F. (1919). *The polish peasant in Europe and America: Monograph of an immigrant group*. Boston: Gorham Press.

Social Judgment Theory

Social judgment theory, developed by the psychologist Kenneth R. Hammond, addresses the implications of our reliance on interrelated multiple fallible cues for making diagnostic and predictive judgments. It addresses the processes of learning to make judgments under uncertainty, learning about the judgments of others, conflict arising from judgmental differences, and how task properties affect judgment processes. It also proposes methods for improving judgment and for addressing problems caused by the fallibility of judgment and our inability to access and describe our own judgment processes or those of others. Those methods are based on a technique called *judgment analysis*. In medical decision making, social judgment theory and its associated methods address problems arising from difficult diagnostic and prognostic judgments and describe the implications of the judgmental processes of physicians, nurses, other healthcare providers, and patients who must make difficult judgments. Diagnosing otitis media is an example of such a difficult judgment. An application of social judgment theory would address the diagnostic process by identifying the cues to this judgment (e.g., bulging or redness of the tympanic membrane), examining the use of those cues in making diagnoses, and determining whether the most valid cues are the ones that the physician relies on most heavily. Uncertainty about the diagnosis, given the cues, and the unreliability of judgments are also addressed.

Overview

Social judgment theory is an extension of Egon Brunswik's probabilistic functionalism. Hammond extended Brunswik's theory, which is primarily

concerned with the perceptual processes of individuals, into the area of diagnostic and predictive judgments of individuals as well as groups. Social judgment theory describes the implications of fallible judgment for people working together as well as for social policy.

Following Brunswik, social judgment theory emphasizes the importance of the task in shaping judgment. The task is both the context for judgment and the context for learning to make judgments. Careful study of the task and understanding how task properties affected judgment is critical for explaining and improving judgmental performance.

People must make judgments under conditions of uncertainty and ambiguity about causes. This results in quasi-rational judgment processes, that is, processes that involve both intuition and analysis. Intuitive processes are not accessible to us. We cannot accurately describe our own judgment processes, and because of causal ambiguity, it is difficult to discern the reasons for others' judgments. As a result, (a) it is difficult to learn to make accurate judgments, (b) it is difficult to learn to understand the reasons for the judgments of others, and (c) this can create misunderstandings that can lead to conflict. The remedies proposed by social judgment theory for these problems involve the use of judgment aids that make the judgment process explicit.

Judgment Analysis

Judgment analysis is a method for making a person's judgment strategy explicit. It is used both as a method for studying judgment and to implement the judgment aids recommended by social judgment theory. Judgment analysis begins with judgments about each case in a set of cases. Each case is described by the values of several variables, or cues. Cases can be real (e.g., patients judged in a clinical setting) or hypothetical. If the cases are hypothetical, they and the judgment made must be representative. Representative design means that the conditions that the researcher wants to generalize to must be specified, and those conditions must be adequately represented in the experimental task so that the desired generalizations can be supported.

The judgments are regressed on the cues, resulting in a statistical model describing the relation between the cues and the judgment. It has been found that linear multiple regression provides a reasonably good fit for the judgments of an individual in a variety of contexts. Judgments of an individual are described by identifying the relative weights for the cues, the shape of the function relating each cue and the judgment (e.g., linear or nonlinear), and the principle by which multiple cues are organized into a judgment. Typically, the organizing principle is linear, but other organizing principles are possible. In social judgment theory, the quality of a model describing judgment is evaluated by its usefulness, not by whether it accurately reproduces actual mental operations.

Learning and Cognitive Feedback

Multiple-cue probability learning has been studied extensively. In the typical paradigm, subjects are shown a series of cases, each consisting of the values of several cues. They are asked to make a judgment on a numerical scale and are then shown the correct answer, which is only probabilistically related to the cues. Learning takes place if, over a number of trials, the correlation between their judgments and the correct answer increases. This is called the *outcome feedback paradigm*. Research has shown that learning from outcome feedback is slow and that it is not difficult to create a task that cannot be learned, even with hundreds or thousands of trials. For example, if there are more than a few cues, one or more cues are nonlinearly related to the correct answer, or if there is high task uncertainty (a weak relation between the cues and the correct answer), learning from outcome feedback is difficult or impossible. Social judgment theory argues that such conditions are not uncommon.

Cognitive feedback has been proposed as an aid to learning and has been studied in medical contexts. Cognitive feedback requires computer software to analyze judgments after a block of trials that includes sufficient cases to calculate a multiple regression model. Typically, relative weights and function forms from judgment analysis are shown to the learner and compared with the optimal weights and function forms for the task. Research has shown that cognitive feedback facilitates learning for tasks that are very difficult to learn with outcome feedback.

There are few if any settings where cognitive feedback is routinely used in practice. Although its value has been demonstrated in the laboratory, there appear to be limitations to its value in applied settings.

Interpersonal Learning

There are situations where one person needs to learn about the judgments of another. These might be training situations, where a student is attempting to learn from a teacher. Or they might be cooperative work situations, where two or more people must reach a common judgment but sometimes disagree.

Learning to understand the judgments of another person is difficult for the same reasons that learning a judgment task is difficult. However, we might expect that learning about another person would be easier if that person could tell us about his or her judgment strategy. Unfortunately, to the extent that there is an intuitive element in judgment, people can neither understand their own judgments nor explain them adequately to others. The remedy proposed by social judgment theory is to analyze the judgments of both parties, thus making them explicit and accessible. This has been demonstrated in laboratory studies, and this process for helping people learn about the judgments of others is the basis for the technique for conflict management described next.

Cognitive Conflict

It is obvious that two or more people sharing common goals and working together in a cooperative setting can make different judgments. Because of causal ambiguity and uncertainty and the inaccessibility of intuitive judgment processes, the reason for those differing judgments can be difficult to determine. Nevertheless, two parties who disagree are likely to seek reasons for that disagreement. The reasons that come to mind are likely to involve competing interests or venality. In other words, the common explanations for disagreement are that people have different interests or that they are deceptive or have nefarious motives. A third explanation, one that may never occur to people who don't understand the nature of judgment, is that both people are honest and goodwilled but have just learned different judgment strategies. But since their judgments involve intuition, they cannot understand or explain their differences. The proposed remedy for this involves use of judgment analysis to make the judgment strategies of both parties explicit so that they can constructively discuss their differences about the use of cues. This shifts attention from personalities and self-interest to the problem at hand, and the expected result is cooperation and conflict reduction.

This result has been demonstrated in laboratory studies, and the method has been found useful in facilitated decision conferences.

Cognitive Continuum Theory

As indicated above, a central problem for social judgment theory is understanding how task properties influence judgment processes. Cognitive continuum theory addresses this problem. In short, cognitive continuum theory argues that both judgment processes and task properties lie at points on a continuum from intuitive to analytic. The idea that intuition and analysis define the ends of a continuum, rather than the opposites of a dichotomy, is a key difference between social judgment theory and other theories that posit a dual system. Cognitive continuum theory specifies those task properties that tend to induce analysis and those that tend to induce intuition. The theory argues that the locations of cognitive activity and task properties on the intuitive-analytic continuum will tend to match and that that match is necessary for best performance.

There have been only a few empirical tests of cognitive continuum theory, and the results have been mixed. It remains one of the few attempts to systematically address the relation between task properties and cognition.

Thomas R. Stewart

Note: There is no relation between social judgment theory as described in this entry and the social judgment theory developed by Muzafer Sherif and Carl Hovland.

See also Lens Model

Further Readings

Balzer, W. K., Doherty, M. E., & O'Connor, R., Jr. (1989). Effects of cognitive feedback on performance. *Psychological Bulletin, 106*(3), 410–433.

Cooksey, R. (1996). *Judgment analysis: Theory, methods, and applications*. New York: Academic Press.

Doherty, M. E., & Kurz, E. M. (1996). Social judgement theory. *Thinking and Reasoning, 2*(2/3), 109–140.

Hammond, K. R. (1965). New directions in research on conflict resolution. *Journal of Social Issues, 21*(3), 44–66.

Hammond, K. R. (1971). Computer graphics as an aid to learning. *Science, 172*, 903–908.

Hammond, K. R. (1996). *Human judgment and social policy: Irreducible uncertainty, inevitable error, unavoidable injustice*. New York: Oxford University Press.

Hammond, K. R., Stewart, T. R., Brehmer, B., & Steinmann, D. O. (1975). Social judgment theory. In M. F. Kaplan & S. Schwartz (Eds.), *Human judgment and decision processes* (pp. 271–307). New York: Academic Press.

Hammond, K. R., Wilkins, M. M., & Todd, F. J. (1966). A research paradigm for the study of interpersonal learning. *Psychological Bulletin, 65*, 221–232.

Tape, T. G., Kripal, J., & Wigton, R. S. (1992). Comparing methods of learning clinical prediction from case simulations. *Medical Decision Making, 12*, 213–221.

Wigton, R. S. (1996). Social judgement theory and medical reasoning. *Thinking and Reasoning, 2*(2/3), 175–190.

Split Choice

A split choice arises if different members of an (apparently) homogeneous population select different options when confronted with the same decision scenario. The splitting of a treatment choice for clinically indistinguishable patients has implications for the economic evaluation of an intervention if the determinants of individual uptake are correlated with the valuation of the health benefit of the intervention. *Split-choice bias* is a term coined to describe the consequences of ignoring these implications.

Split Choice in Clinical Practice

In a typical clinical setting, it may happen that one treatment has clear advantages over another, with side effects that are negligible in comparison with the potential gains. Most patients will accept such an intervention regardless of the strength of their preference for the likely health outcomes. On the other hand, there are many scenarios where some patients will decline an intervention that others will accept, simply because they place a lower relative value on its consequences. For some, the overwhelming concern may be a wish to avoid certain side effects, while others may simply attach less importance to the potential benefits of the treatment.

Examples include radical treatment for potentially fatal conditions (e.g., Stage I prostate cancer), prophylactic interventions in healthy individuals (e.g., mastectomy for women at high genetic risk of breast cancer), and prenatal testing for genetic abnormality. Many screening tests also fall into this category. In such cases, patients may make different treatment choices even though they have identical clinical prospects, with identical probabilities attached to the outcomes of their treatment. In other words, the clinical population splits into preference subgroups when confronted with the treatment choice. This can also apply when the effective treatment choice is deferred until after any side effects have come into play. Thus, a therapy with reversible side effects may be discontinued by some patients even though their objective clinical experience has been no worse than that of some who choose to continue.

Economic Evaluation and Split-Choice Bias

Economic evaluations for healthcare providers often quantify health benefits using preference-based comparisons of different health states. For example, the calculation of quality-adjusted life years (QALYs) is made by taking an average, over all health states, of the time spent in a state multiplied by a preference weight attached to that state. The preference weights are elicited as utility values between 0 (*death*) and 1 (*the best conceivable state of health*) from an appropriate population. In practice, a health state will generate different utility assessments from different individuals. Therefore an operational preference weight is obtained as a population-average utility.

This approach is unproblematic if all patients make the same decision when confronted with a treatment choice. The QALY value of a treatment

obtained from the average population utility will equate exactly to the QALY value that would be obtained if individual utilities were used instead: Subjects with higher and lower values will cancel each other out, since all receive the same treatment.

In a split-choice scenario, the health benefit of a treatment can be experienced only by those who choose to accept it. It follows that the QALY valuation of a treatment ought to use a set of preference weights for the health state outcomes derived solely from the subpopulation of acceptors. This is an important stipulation, since it is plausible that treatment acceptors will attach higher utility values to the outcomes of a treatment than would those who choose to decline it. Indeed, a fully rational model for patient decision making would predict that those accepting treatment would do so just because of the higher expected utility that they associate with its consequences. Thus, a potential for bias arises in any economic evaluation in which preference weights for treatment outcomes are obtained from a population that includes individuals who would actually decline the treatment. Such split-choice bias can operate in one direction only, namely, to dilute the apparent effectiveness and hence the cost-effectiveness of the treatment in the group of patients who would choose to accept it. Moreover, the extent of the bias can be considerable and is sensitive to the degree to which individual decision making follows rational precepts. For example, it has been demonstrated that the QALY value could be underestimated by a factor of up to one half for treatments accepted by 70% of patients, with even larger biases possible at lower acceptance rates. Biases as large as this would lead to grossly distorted cost-effectiveness ratios.

Avoiding Split-Choice Bias

In practice, published cost-utility analyses often disregard the potential for split-choice bias. For example, several recent analyses have tackled the subject of prostate cancer screening using population-based QALY calculations. Similarly, a number of models of screening for bowel cancer have failed to distinguish between those who accept and decline treatment.

Whether split-choice bias is an issue for an economic evaluation will depend on the aims of that evaluation. If the evaluation is done simply to decide which of two active treatments to fund (typically a choice between a standard and a novel treatment), conventional QALY calculations based on population-average preference weights may often be used with impunity. The issue of split-choice bias does not arise if all future patients will experience the same treatment. The more problematic cases arise where the aim of the evaluation is to assess treatments that will never be more than an option in a particular clinical scenario, such as those mentioned above. Then the requirement to avoid bias in a cost-utility analysis suggests that relevant health state preference weights should be elicited alongside respondents' attitudes to the treatment choice at issue. The incremental QALY calculations must then be carried out within the subgroup of patients who have indicated that they would accept the optional treatment. A major difficulty with this proposal is that it cannot be carried out using published tables of preference weights, such as those associated with the EQ-5D instrument. Instead, a purposive sample of preference weights must be constructed for each decision scenario that is evaluated. This is inconvenient and also unreliable, in the sense that the precision of each new evaluation will be limited by the size of the purposive sample of responses that can be mustered. An alternative, though untried, suggestion is to model the relationship between treatment acceptance and individual utility scores across a range of clinical scenarios and use this relationship to generate an average QALY valuation among treatment acceptors.

Alan Girling and Richard Lilford

See also Cost-Utility Analysis; EuroQoL (EQ-5D); Expected Utility Theory; Quality-Adjusted Life Years (QALYs); Utility Assessment Techniques

Further Readings

Bekker, H., Thornton, J. G., Airey, C. M., Connelly, J. B., Hewison, J., Robinson, M. B., et al. (1999). Informed decision making: An annotated bibliography and systematic review. *Health Technology Assessment, 3*(1), 1–156.

Fransen, M., & Edmonds, J. (1999). Reliability and validity of the EuroQol in patients with osteoarthritis of the knee. *Rheumatology (Oxford), 38,* 807–813.

Lilford, R. J., Girling, A. J., Braunholtz, D., Gillett, W., Gordon, J., Brown, C. A., et al. (2007). Cost-utility analysis when not everyone wants the treatment: Modeling split-choice bias. *Medical Decision Making, 27,* 21–26.

Lilford, R. J., Girling, A. J., Stevens, A., Almasri, A., Mohammed, M. A., & Braunholtz, D. (2006). Adjusting for treatment refusal in rationing decisions. *British Medical Journal, 332,* 542–544.

Pauker, S. G., & Kassirer, J. P. (1997). Contentious screening decisions: Does the choice matter? *New England Journal of Medicine, 336,* 1243–1244.

Singer, L. G., Gould, M. K., Tomlinson, G., & Theodore, J. (2005). Determinants of health utility in lung and heart-lung transplant recipients. *American Journal of Transplantation, 5,* 103–109.

Statistical Notations

Statistics are pervasive in scientific journals as well as imbedded in business reports, academic textbooks, and even the popular media. An understanding of basic statistical terminology can be critical toward both disseminating and comprehending methodology and facilitating the appropriate interpretation of data. Ideally, there would exist a clear consensus for the utilization of a specific notation for statistical terms; however, even between various scientific communities (e.g., statistics, epidemiology, medical decision making, econometrics), there exist some variations in the notation for certain statistics. For instance, the acronym SEM may refer to highly diverse terms, such as the standard error of the mean, structural equation modeling, or scanning electron microscopy, within particular scientific disciplines. As such, a basic understanding of the context of any research is always helpful toward properly interpreting statistical terminology. Despite some variability in the usage of terms, there remain certain statistics that can generally be identified and characterized across disciplines. The objective of this entry is to outline some of the commonly used notations, symbols, and acronyms to serve as a reference for consumers of statistical data.

Beyond specific notations, there are general rules of thumb regarding the interpretation of statistical terminology, based on the manner in which they are presented. For instance, the use of the Greek alphabet is typically indicative of population parameters, which in practice are unobservable theoretical values (e.g., σ^2 is the variance from a population). In contrast, Latin terms are more commonly indicative of sample statistics (e.g., s^2 as the sample variance). The hat symbol (e.g., $\hat{\boldsymbol{Y}}$) is used to denote estimated or predicted values (in this case for a random variable Y), which are commonly expressed in regression notation. By convention, boldface terminology represents vectors or matrices (e.g., the matrix $\boldsymbol{X}$ may contain values x_{ij}, where i indicates the row and j indicates the column location of the matrix element). In general, uppercase values denote random variables, and lowercase values represent specific values. For example, $Pr(X = x)$ represents the probability that a random variable X is equal to a given value x. The bar symbol over a lowercase Latin value generally indicates a mean value ($\bar{x}$, spoken "x-bar"), and a period (.) or plus (+) subscript after a variable both generally refer to the summation across level(s) ($x_{1\cdot}$ or x_{1+} indicates the sum of all observations across all x_{1j} values). The following are an alphabetized listing of often used notations and acronyms (arranged alphabetically by the primary descriptor) along with a brief (nontechnical) descriptor that may serve as a guide to those interested in the science of medical decision making.

Statistical Notations

ADL—*Activities of daily living:* A measure of patient level of function—typically used as an outcome measure or risk adjustment in research.

α—*Alpha:* Most commonly denotes the significance level of a hypothesis test; in addition, it is used in regression notation as an intercept parameter (e.g., $y = \alpha + \beta x$).

$\boldsymbol{H_1}$ **or** $\boldsymbol{H_a}$—*Alternative hypothesis:* The hypothesis that the null hypothesis is tested against—also referred to as the research hypothesis.

ANOVA—*Analysis of variance:* An analysis designed to detect variation in a set of responses as a function of independent variable(s). The simple one-way generalization tests for differences in means between two or more groups.

ANCOVA—*Analysis of covariance:* Generally refers to an extension of an ANOVA model with inclusion of continuous covariates.

AUC—*Area under the curve:* Typically used as a measure of the predictive value of a given diagnostic for a categorical response variable (also applicable for pharmacokinetic studies); technically represents the integral of a function in a given domain.

ARMA—*Autoregressive moving-average model:* A form of time series model that incorporates both a moving average component and the correlation of observations that are more proximate.

BLUE—*Best linear unbiased estimator:* A linear estimator of a given parameter; the "best" descriptor indicates that it has the lowest variability as compared with any other estimator.

BLUP—*Best linear unbiased predictor:* Predicted value of a response variable from a BLUE.

β—*Beta:* Typically denotes the probability of a Type II error (1 – power); also used in the context of regression models as an effect parameter (e.g., $y = \alpha + \beta x$).

χ^2_v —*Chi-square:* Commonly used distribution to test hypotheses (e.g., association of two variables in a contingency table); may be denoted with the applicable degrees of freedom (ν).

R^2—*Coefficient of determination:* The square of the correlation statistic (r); ranges from 0 to 1 and indicates the amount of variability of a response variable explained by one or more explanatory variables.

CV—*Coefficient of variation:* A measure of dispersion for a data set, equal to 100 × standard deviation/mean.

$C(n, r)$ or ${}_nC_r$—*Combination:* Commonly associated with probability for unordered groups or for sampling with replacement of elements; $C(n, r)$ is equal to $n!/(n - r)! \times r!$ for a given n and r with $0 \le r \le n$.

C_{max}—*Concentration maximum:* Used primarily in pharmacokinetic studies.

|—*Conditional statement:* Spoken as "given"; for example, $E[Y|X]$ is an expression indicating the expected value of a random variable Y given (conditioned on) X.

CI—*Confidence interval:* A range of values estimated by observations to contain a given parameter; the uncertainty level (5% for a 95% CI) is indicative of the process to create the interval rather than the probability that a parameter is contained in a particular interval.

r—*Correlation coefficient (Pearson):* Sample measure of strength and direction of the linear relationship between two continuous variables.

r_s—*Correlation coefficient (Spearman):* Nonparametric sample measure of strength of association between variables using only the respective ranks of observations.

ρ—*Correlation coefficient:* Population parameter estimated by the sample statistic r.

Cdf—*Cumulative distribution function:* Characterizes a random variable; technically equivalent to the $\Pr(X \le x)$ for a given random variable X and typically denoted as $F(x)$.

df—*Degrees of freedom (ν):* Most generally interpreted as the amount of information available to test a given hypothesis, but it has multiple uses, including defining the forms of standard statistical distributions (e.g., χ^2_v, t_ν, $F_{\nu 1,\nu 2}$).

Δ—*Delta:* Expresses the change in a given value or parameter.

~—*Distributed as:* Typically refers to how a random variable is distributed; for example, $X \sim N(0, 1)$ indicates that the random variable X is distributed as a normal distribution with $\mu = 0$ and $\sigma = 1$.

ε—*Error:* Theoretical error parameter, commonly noted for the form of a regression model (e.g., $Y_i = \alpha + \beta X_i + \varepsilon_i$).

$E[Y]$—*Expected value:* The mean of a random variable—here denoted for a random variable Y.

!—*Factorial: n*! is equal to $n \times (n-1) \times (n-2) \times \cdots \times 1$.

***F* or $F_{\upsilon 1,\upsilon 2}$**—*F distribution:* The *F* test is commonly used for an ANOVA model and may be represented with numerator and denominator degrees of freedom.

GEE—*Generalized estimating equation:* Most commonly used as a model for longitudinal data with nonnormal error distributions.

GLM—*Generalized linear model:* A general class of statistical models; the term *linear* refers to the response variable expressed as a linear combination of explanatory variables.

GIS—*Geographic index system:* Software indicating the geographic location or proximity of units of analysis and also commonly used to track the spread of disease outbreak.

HR (AHR)—*Hazard ratio (adjusted hazard ratio):* The effect of an explanatory variable on the hazard (or risk level) for a particular event—commonly associated with survival analysis.

∫—*Integral:* Expression used to express the "area" of a function over a given range; $\int_b^a f(x)dx$ indicates the integral of a function $f(x)$ over the range a to b.

ITT—*Intention to treat:* Most commonly refers to the type of analysis in a clinical trial in which subjects are evaluated based on original group assignment, as compared with analyses that use information about subject treatment group changes over the course of a trial.

∩—*Intercept:* Probability notation indicating the inclusion of common elements in two or more groups; for example, $A \cap B$ indicates that all elements that are common to both sample spaces *A* and *B* are considered.

IQR—*Interquartile range:* Defined as the difference between the 75th and 25th percentiles of a data set; indicative of the dispersion of the "common" values in a data set.

ICC—*Intraclass correlation:* Most commonly refers to the proportion of variation that can be attributed to "between-subjects" relative to the total variation in a model.

κ—*Kappa coefficient:* Typically used to assess the level of agreement between raters; ranges between 0 and 1, with higher levels associated with more agreement.

LSD—*Least significant difference:* Measure used to estimate significant differences of group mean levels accounting for multiple comparisons and "experiment-wise" error.

MCMC—*Markov chain Monte Carlo:* Sampling methodology often using the simulation of observations to produce predictions of observations and evaluate distributional assumptions; commonly used in Bayesian statistics.

mle—*Maximum likelihood estimation:* Methodology used to estimate parameters that result in estimators with certain useful statistical properties.

MSE—*Mean squared error:* Measure of the sampling variability within treatments; commonly reported with ANOVA models.

MOM—*Method of moments:* Methodology used to derive point estimators; these estimators do not always have the most desirable properties (e.g., less efficient, biased) and as such are less commonly used currently with modern computational methods.

MAR—*Missing at random:* Refers to an assumption used for missing values in a data set. Depending on the nature of missing values, whether missing values are informative (nonrandom) can affect the method of analysis; another level is missing completely at random (MCAR), which is a broader assumption.

MANOVA—*Multivariate analysis of variance:* An extension of an ANOVA model with multiple response variables; commonly used for repeated measures data.

ns—*Not significant:* Used as shorthand to denote that a particular estimate is not statistically significant based on a given Type I error probability (e.g., RR = 1.05, 95% CI: 0.98–1.03, $p = ns$).

H_0—*Null hypothesis:* The hypothesis of a study that typically indicates "no effect" or the status quo; what an alternative hypothesis is tested against.

OR (**AOR**)—*Odds ratio (Adjusted odds ratio):* Measure of the odds of a "success" relative to a "failure" in one group versus another; commonly reported from contingency tables or logistic regression models and may be adjusted for other factors.

OLS—*Ordinary least squares:* Linear regression model assuming a particular distributional form, including normally distributed independent random variables with common variance.

$P(n, r)$ **or** ${}_nP_r$—*Permutation:* Commonly associated with probability or sampling without replacement of elements; $P(n, r)$ is equal to $n!/(n - r)!$ for a given n and r with $0 \leq r \leq n$.

φ—*Phi coefficient:* Measure of association of two independent variables from a 2×2 contingency table; ranges between -1 and 1.

μ—*Population mean:* Theoretical arithmetic mean for a variable from a population.

π—*Population proportion:* Parameter indicating the proportion of individuals in a population with a certain trait.

N—*Population size:* Number of observations in a population.

σ—*Population standard deviation:* Indication of the variability or dispersion of a variable from a population; σ^2 is the population variance.

$P(A)$ or $\Pr(A)$—*Probability:* Notation indicating the probability of an event (A) occurring.

pdf—*Probability density function:* Characterizes a random variable; technically satisfies the expression $F(x) = \int_{-\infty}^{x} f(t)dt$ for all x.

p or *p* **value**—*Probability of a Type I error:* The probability of a given result assuming the null hypothesis; often a p value $<.05$ is considered statistically significant; p may also be used to indicate a sample proportion.

QUALY/QOL—*Quality of life:* Used in studies to provide a mathematical index for patients in various health states; ranges from 0 to 1 (from *dead* to *alive with an ideal quality of life*).

Q-Q plot—*Quantile-quantile plot:* Used to provide a representation of the goodness of fit for a data set for a theoretical distribution; cases in which ordered observed values closely align with quantiles of a distribution suggest an adequate fit.

RBD—*Randomized block design:* An experimental design in which units receive a randomly assigned treatment within each block. For example, plants located in three different rows (blocks) randomly receive one of three types of fertilizer.

RCT—*Randomized controlled trial:* A clinical trial using a control group and a randomization procedure to allocate subjects to study arms.

ROC—*Receiver operating characteristic curve:* Plot of the sensitivity against 1 – specificity, typically used to represent the overall predictive value of a diagnostic and particular values at which a diagnostic is most predictive.

RR—*Relative risk:* The incidence rate of an event in one condition divided by the incidence rate of an event for observations in another condition.

REML—*Residual (or restricted) maximum likelihood estimation:* Method to estimate parameters based on a restricted likelihood (rather than the actual likelihood); common in mixed models and has the advantage of producing unbiased estimates of variance components.

$\bar{x}$—*Sample mean:* Arithmetic mean for a sample of observations; a measure of central tendency for a continuous variable.

n—*Sample size:* The number of observations in a sample.

SD **or** *s*—*Sample standard deviation:* Measure of the variability or dispersion of a variable from a sample data set; is in the same units as the original observation; s^2 is the sample variance.

SE* or *SEM*—*Standard error (of the mean): The standard deviation of the sampling distribution of a statistic; the standard error of the sample mean is estimated by $s/-\sqrt{n}$.

SMR—*Standardized mortality ratio:* Ratio of the number of deaths for a given population and the number of deaths expected based on certain characteristics (e.g., age and gender); also can be generalized for events other than death.

Σ—*Summation:* Notation used to represent the addition of an indicated variable (e.g., Σx is used to represent the sum of all x values); technically should include an indication of the range of observations that are to be summed; $\Sigma_{i=1}^{4} x_i$ indicates that x is to be summed from x_1 to x_4.

t* or *t*—*t distribution: Also known as the student's t distribution; most commonly used as a test for a sample mean or a comparison of sample means; as ν increases, the distribution approaches the normal distribution.

T_{max}—*Time of maximum concentration:* Time at which maximum concentration occurs in pharmacokinetic studies.

′ or ^{T}—*Transpose:* Notation used to indicate the rotated alignment of a matrix; for example, for a 2×3 matrix **C**, $\mathbf{C}' \times \mathbf{C}$ indicates the product of $\mathbf{C}'$ (which is 3×2 when transposed) and **C**.

∪—*Union:* Probability notation indicating the inclusion of all elements in two groups; for example, $A \cup B$ indicates that all elements in the sample space of either A or B are considered.

VIF—*Variance inflation factor:* Measure of the impact of other explanatory variables on the variance of another variable in the context of a regression model.

z—*z distribution:* Refers to standard normal distribution with mean equal to 0 and variance equal to 1.

Jesse D. Schold

See also Basic Common Statistical Tests: Chi-Square Test, *t* Test, Nonparametric Test; Statistical Testing: Overview

Further Readings

Agresti, A. (1990). *Categorical data analysis*. New York: Wiley.
Altman, D. G. (1991). *Practical statistics for medical research*. London: Chapman & Hall.
Cassela, G., & Berger, R. L. (1990). *Statistical inference*. Belmont, CA: Duxbury Press.
Everitt, B. S. (1998). *Dictionary of statistics*. Cambridge, UK: Cambridge University Press.
Steel, R. G. D., Torrie, J. H., & Dickey, D. A. (1997). *Principles and procedures of statistics: A biometrical approach* (3rd ed.). New York: McGraw-Hill.
Streiner, D. L., & Norman, G. R. (2003). *Health measurement scales*. Oxford, UK: Oxford University Press.
Wackerly, D., Mendenhall, W., III, & Scheaffer, R. (1996). *Mathematical statistics with applications* (5th ed.). Belmont, CA: Duxbury Press.

Statistical Testing: Overview

Statistical testing is common in clinical settings as a method for drawing inferences for an unknown population value based on a sample of subjects. A clinician may wish, for instance, to test the hypothesis that a new surgical technique reduces the probability of adverse postoperative outcome or that, on average, a new drug reduces blood glucose in diabetics. Many types of statistical tests exist, and selection of an appropriate test is guided by the type of data collected, the statistics for which inferences are desired, the number of groups under study, and the sample size.

General Testing Procedure

Implementation of a statistical test begins with the specification of distinct hypotheses; a *null hypothesis* (denoted by H_0) is assumed to be true, and the test is performed to evaluate the evidence against H_0 in favor of an *alternative hypothesis* (denoted by H_A). Typically, the hypothesis that the researcher may wish to show is specified as H_A. For example, H_0 for the study on the new surgical technique could be stated as "The probabilities of adverse outcome for patients assigned to the new treatment and patients assigned to the standard of care are equal." H_A would correspondingly be stated as

"Adverse outcome probabilities are unequal for the two treatment techniques."

The investigator collects a sample of data from the population of interest, and a *test statistic* is derived using the sample responses. The test statistic is defined as a quantity that summarizes the sample in such a way that a decision to accept H_0 (in preference to H_A) or to reject H_0 (in favor of H_A) can be made based on all possible values of the quantity. The set of values corresponding to the decision to accept H_0 is called the *acceptance region*, while the set of values corresponding to the decision to reject H_0 is called the *rejection region*.

The test statistic has a certain probability distribution—the *sampling distribution*—under the assumption that H_0 is true (i.e., the probability distribution of test statistics arising from many repeated samples). The shape of the sampling distribution depends on the type of test being implemented.

As the relevant population quantity is unknown (thereby necessitating testing), the potential exists for the test to produce an incorrect conclusion. The conclusion to reject H_0 when indeed H_0 is true is known as a *Type I error* or *false positive*. On the other hand, accepting H_0 when H_A is true is called a *Type II error* or *false negative*. The probability of committing a Type I error (often referred to as the *significance level*) is commonly denoted by α, while the probability of committing a Type II error is commonly denoted by β. The significance level corresponds to the range of the rejection region and is specified by the experimenter prior to testing. Many journals require $\alpha = .05$ for testing.

Statistical tests can also be implemented in terms of *p values*, defined as the probability of observing a test statistic as extreme or more extreme than that which would be observed if the experiment were to be repeated many times—or, in other words, the observed significance level of the test. The decision to reject H_0 is made if the p value is less than α, and the proximity of the p value to 0 is a measure of the strength of the evidence against H_0.

Alternative hypotheses can be specified as *two sided* or *one sided*. Suppose an investigator was interested in comparing the average blood pressure between patients randomized to an experimental medication (μ_t) and patients randomized to control (μ_c). A two-sided set of hypotheses in this instance would be H_0: $\mu_1 = \mu_c$; H_A: $\mu_t \neq \mu_c$. A one-sided set of hypotheses, in this case, could be stated as H_0: $\mu_1 \geq \mu_c$; H_A: $\mu_t < \mu_c$. While one-sided tests may require fewer patients for the same statistical power, they should only be implemented under the rare circumstance that the investigator was certain that the treatment would not result, for instance, in a worse/higher blood pressure (as compared with the placebo).

Choice of an Appropriate Test

Choice of an appropriate test requires an understanding of the type of outcome measured, the experimental design, and the assumptions required for implementation of each test. Adequate coverage of all assumptions for each test requires much detail and is beyond the scope of this entry; however, as a general rule, parametric tests are recommended for normally distributed, continuous outcomes or when the sample size is large (say $n > 30$), while nonparametric tests are recommended for nonnormal continuous outcomes or when the sample size is small. Also, most tests (excluding paired tests for continuous outcomes and tests for independence of categorical/binary measures) require the assumption that observations are collected independent of one another.

Common Types of Statistical Tests

The overviews in this section assume the p value approach to hypothesis testing; only formulas for p values are presented (rejection regions are not). Also, all stated hypotheses are two sided.

Tests for Categorical Outcomes

Many clinical trials involve categorical outcomes, as treatments often focus on preventing the diagnosis (yes/no or binary outcome) of a certain disease. Other studies may seek to compare the incidence of disease for a specific population to that known for a different population.

One-Sample Proportion Test

The null hypothesis for this test is that the proportion of patients with a certain outcome is equal to a prespecified null value (H_0: $p = p_0$),

while the alternative is H_A: $p \neq p_0$. After the data are collected (and sample proportion $\hat{p}$ has been found), the test statistic can be computed as

$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1 \quad p_0)}{n}}},$$

with a corresponding p value of $2\Pr\{Z \geq |z|\}$, where Z is a standard normal random variable (with mean of 0 and variance of 1).

Example. Suppose the investigator wishes to test the null hypothesis that the incidence of surgical complications in her population is equal to the incidence—say 25%—among a similar patient population that had been previously published. She collects data on 80 patients and finds that 12 experienced a complication (amounting to 15%). The z statistic for this test is calculated as

$$z = \frac{.15 - .25}{\sqrt{\frac{.25(1 \quad .25)}{80}}} = -2.07.$$

The p value associated with this test statistic is .039, and (using $\alpha = .05$) she concludes that the incidence among her patients is significantly lower than that which was previously published.

Pearson's Chi-Square Test

This test is most often used in randomized clinical trials involving a binary treatment indicator (usually treatment vs. control) and a binary outcome (usually diagnosis or absence of disease after treatment). The null hypothesis in this case is that the probability of outcome for patients randomized to the treatment (p_t) equals the probability of outcome for patients randomized to the control (p_t)—symbolically, (H_0: $p_t = p_c$), and the alternative hypothesis is H_A: $p_t \neq p_0$. Assuming that n_t and n_c patients are given the treatment and control (respectively), a *two-way table* can be constructed after the experiment is performed (see Table 1).

Notationally, n_{*1} total patients experienced the outcome n_{11} in the control arm and n_{21} in the treatment arm, while n_{*2} total patients experienced the outcome n_{12} in the control arm and n_{22} in the treatment arm (total sample size of n_{**}).

Table 1 Two-by-two contingency table used in the implementation of the chi-square test

	No Outcome ($j = 1$)	*Outcome* ($j = 2$)	*Totals*
Control ($i = 1$)	n_{11}	n_{12}	$n_{*1} = n_c$
Treatment ($i = 2$)	n_{21}	n_{22}	$n_{*2} = n_t$
Totals	n_{*1}	n_{*2}	n_{**}

The chi-squared test statistic is then a measure of the deviation between the observed frequencies n_{ij} and the expected frequencies $e_{ij} = n_{i*}(n_{*j})/n_{**}$:

$$X^2 = \sum_{i=1}^{2}\sum_{j=1}^{2} \frac{(n_{ij} - e_{ij})^2}{e_{ij}}.$$

This formula can be extended to include I treatments and J possible outcomes. The test statistic X^2 follows a χ^2 distribution with $(I - 1)(J - 1)$ degrees of freedom, and the p value is $\Pr\left(\chi^2_{(I \quad 1)(J \quad 1)} \geq X^2\right)$.

Example. An investigator wishes to determine whether or not administration of a new anti-inflammatory drug is effective in reducing the incidence of postsurgical infection in a population of cardiac surgery patients. He randomizes 1,500 patients to receive the new drug and another 1,500 to receive a placebo, of which 130 and 194 patients, respectively, developed an infection. The observed counts n_{ij} and expected counts e_{ij} are shown in Table 2.

The X^2 statistic is thus

$$X^2 = \frac{(1306 - 1338)^2}{1338} + \frac{(1370 - 1338)^2}{1338} + \frac{(194 - 162)^2}{162} + \frac{(130 - 162)^2}{162} = 14.2.$$

The probability of finding an X^2 statistic (with $(2 - 1)(2 - 1) = 1$ degree of freedom) greater than or equal to his observed X^2 of 14.2 is less than .001; therefore, he concludes that the anti-inflammatory drug is effective in reducing the incidence of postoperative infection.

Table 2 Contingency table for anti-inflammatory study

	No Infection	*Infection*	*Totals*
Placebo	n_{11} = 1,306 e_{11} = 1,338	n_{12} = 194 e_{12} = 162	n_{*1} = 1,500
Anti-Inflammatory	n_{21} = 1,370 e_{21} = 1,338	n_{22} = 130 e_{22} = 162	n_{*2} = 1,500
Totals	n_{*1} = 2,676	n_{*2} = 324	n_{**} = 3,000

Parametric Tests for Continuous Outcomes

Normally distributed outcomes occur frequently in clinical practice. Some examples of normal measures include weight, height, and blood pressure. Inference for normal outcomes focuses on the population mean μ. Though the parametric tests in this section are best interpreted for normal outcomes, inference for means of nonnormal outcomes is valid as long as the sample is of adequate size. Additionally, these tests may be considered for nonnormal outcomes after an appropriate normalizing transformation is applied (such as the log or square-root transformations).

One-Sample t Test

The one-sample *t* test is ideal for situations where the investigator is interested in comparing the mean of an outcome (collected on a single group of patients) to a prespecified null value μ_0. The null hypothesis is that the population mean (μ) is equal to a certain value (H_0: $\mu = \mu_0$). The alternative hypothesis is that the mean is not equal to the null value (H_A: $\mu \neq \mu_0$). The test statistic is

$$t = \frac{x - \mu_0}{s/\sqrt{n}},$$

where $\bar{x}$ is the sample mean and *s* is the sample standard deviation. This statistic follows a *t* distribution with $n - 1$ degrees of freedom (*df*). The *p* value for this test is $2\Pr\{t_{n-1} \geq |t|\}$.

Two-Sample t Test (Paired Samples)

Two typical clinical designs where paired observations arise are (1) measuring an outcome before and after a treatment regimen for each patient and (2) measuring one posttreatment outcome on patients who were matched on baseline characteristics (e.g., half-randomized to treatment, half-randomized to control). Statistical test of mean difference for paired samples amounts to a one-sample *t* test performed on the differences (e.g., posttreatment—baseline, treatment—placebo), usually with $\mu_0 = 0$ and *n* equal to the number of observed differences. Hypotheses are stated in terms of the mean difference (μ_d), as in H_0: $\mu_d = 0$ and H_A: $\mu_d \neq 0$.

Example. Dr. Huang has developed a new diet program for clinically obese patients and has reason to believe that it may be effective in weight loss. She has 20 patients begin and complete her new diet program, weighing each of them before beginning and after completion. The statistical hypothesis she tests is that the mean difference in weight is 0. The mean (*SD*) difference (after enrollment minus before enrollment) in weight was −5(12) pounds, giving a test statistic of

$$t = \frac{x - \mu_0}{s/\sqrt{n}} = \frac{(-5) - 0}{12/\sqrt{20}} = -1.86.$$

The *p* value for this test is $2\Pr\{t_{n-1} \geq |t|\} = 2\Pr\{t_{20-1} \geq -1.86|\} = .78$, and she therefore cannot reject her null hypothesis.

Two-Sample t Test (Independent Samples)

This test is the most common analysis approach for the typical two-arm randomized clinical trial investigating the difference in means of an outcome between a treatment group (μ_t) and a control group (μ_c). The null hypothesis is that the difference in means between the two groups is equal to a null difference δ, which is typically 0 (H_0: $\bar{x}_t - \bar{x}_c = \delta$. The alternative hypothesis is H_A: $\bar{x}_t - \bar{x}_c \neq \delta$. Assuming unknown and potentially unequal population variances, the test statistic is approximated using a *t* statistic:

$$t \approx t' = \frac{(x_t - x_c) - \delta}{\sqrt{\frac{s_t^2}{n_t} + \frac{s_c^2}{n_c}}},$$

where (x_t, x_c) are sample means, (s_t, s_c) are sample standard deviations, and n_t, n_c are sample sizes from the (treatment, control) groups. This test

statistic follows a t distribution with df approximated using the Welch-Satterthwaite method:

$$df \approx df^* = \frac{\left(\frac{s_t^2}{n_t} + \frac{s_c^2}{n_c}\right)^2}{\frac{s_t^4}{n_t^2(n_t \quad 1)} + \frac{s_c^4}{n_c^2(n_c \quad 1)}}.$$

The p value for this test is then $2\Pr\{t_{df^*} \geq |t|\}$.

Example. Mean (*SD*) case duration of 75 cardiac patients randomized to preoperative treatment with beta blockers was 213 (65) minutes, while mean (*SD*) case duration for 75 patients randomized to a preoperative placebo was 204 (61) minutes. The t statistic for the null hypothesis of no difference in mean case duration is

$$t' = \frac{(213 - 204) - 0}{\sqrt{\frac{65^2}{75} + \frac{61^2}{75}}} = .87,$$

and the p value is $2\Pr\{t_{df^* \approx 147} \geq |.87|\} = .39$ using the Welch-Satterthwaite method to approximate df. This study does not give evidence that beta blockers cause a change in average case duration.

Nonparametric Tests for Continuous Outcomes

The area of nonparametric testing has provided an array of statistical tests that address a diverse collection of hypotheses. Adequate coverage of these tests requires a full-length text (e.g., Siegel & Castellan, 1988); this section focuses on tests that are common in the clinical setting. For example, a popular clinical implementation is when the investigator is interested in comparing the shape of the distribution of a nonnormal outcome (e.g., body mass index, survival time after treatment, certain serum measurements) between the treatment and control arms of a clinical trial.

Many nonparametric tests are based on the ranks of the responses rather than the actual values. Let $R(x)$ denote the rank function for variable x, where the smallest x is assigned a score of 1 and the largest x is assigned a score of n (any tied x values are typically assigned the average score of the affected ranks). Also, let $I(c)$ denote the indicator function for the Boolean clause c (equal to 0 if c is false and 1 if c is true).

Wilcoxon Signed-Rank Test

The signed-rank test is generally implemented as an alternative to the paired-samples t test, as it requires only the assumption that the calculated differences follow a symmetric (as opposed to normal) distribution. Define the differences (either baseline and posttreatment measurements pre_i and post_i or matched treatment and control measurements treat_i and control_i, respectively) as $d_i = \text{post}_i - \text{pre}_i$ or $d_i = \text{treat}_i - \text{control}_i$. Then the null hypothesis is that the median of the differences equals a certain null value, which is usually 0 (H_0: $\tilde{d} = \tilde{d}_0$), and the alternative is H_A: $\tilde{d} \neq \tilde{d}_0$.

Implementation of this test begins by removing all w observations where $d_i = \tilde{d}_0$. Then, define the ranks as $r_i = R(|d_i|)$, $i = 1, 2, \ldots, n'$, where $n' = n - w$. The test statistic T is $\min(T^+, T^-)$, where

$$T^+ = \sum_{i=1}^{n'} r_i I(d_i > \tilde{d}_0),$$

and

$$T \quad = \sum_{i=1}^{n'} r_i I(d_i < \tilde{d}_0) = \frac{n'(n'+1)}{2} - T^+.$$

With adequate sample size (say $n' > 15$), the p value can be approximated:

$$p = 2\Pr\left(z > \frac{|T - n'(n'+1)/4|}{\sqrt{n'(n'+1)(2n'+1)/24}}\right),$$

where z follows a standard normal distribution.

Example. Dr. Martinez is interested in studying the effects of a new medication on resting systolic blood pressure (SBP). He recruits 17 cardiac patients, measures resting SBP before and after treatment, and tests the null hypothesis that the median difference in (after minus before) is equal to 0 mmHg. The observed differences (in ascending order) and ranks r_i are shown in Table 3.

Based on these observations, the median difference in SBP is −7 mmHg. Since there is one observation where $d_i = \tilde{d}_0$, $n' = 16$; T^+ is calculated as the sum of the ranks of the positive d_i, which is 36, and $T^- = (16(17)/2) - T^+ = 100$. Therefore, T is 36. The p value is

$$p = 2\Pr\left(z > \frac{|36 - 16(17)/4|}{\sqrt{16(17)(37)/24}}\right) = 2\Pr(z > 1.563) = .12.$$

Dr. Martinez cannot conclude that median difference in SBP is different from 0 mmHg.

Table 3 Differences (after treatment minus before treatment) in systolic blood pressure, as well as ranks, for 17 cardiac patients

d_i	–42	–40	–27	–15	–12	–12	–9	–8	–7	–7	–5	–1	0	6	10	18	22
r_i	16	15	14	11	9.5	9.5	7	6	4.5	4.5	2	1	—	3	8	12	13

Wilcoxon/Mann-Whitney Rank Sum Test (U Test)

Since it doesn't require any distributional assumptions, the U test is a powerful alternative to the independent-samples t test. H_0 is that the outcome of interest (x_i) follows a common statistical distribution between the treatment and control arms of a clinical trial, while H_A is that (x_i) for a randomly selected patient in the treatment arm is different from (x_i) for a randomly selected patient in the control arm. Let $r_i = R(x_i)$ and $T_i = I$(patient i in treatment arm), $i = 1, 2\ldots, n$. Then the number of patients in the treatment and control arms, respectively, are $n_t = \sum_{i=1}^{n} T_i$ and $n_c = \sum_{i=1}^{n} (1 - T_i)$ The test statistic is

$$U = \min\left[\left(\sum_{i=1}^{n} r_i T_i - \frac{n_t(n_t+1)}{2}\right), \left(\sum_{i=1}^{n} r_i(1-T) - \frac{n_c(n_c+1)}{2}\right)\right].$$

The p value for this test can also be approximated using the standard normal distribution if $(n_c > 15)$:

$$p = 2\Pr\left(z > \frac{|U - n_t n_c/2|}{\sqrt{n_t n_c(n_t + n_c + 1)/12}}\right).$$

Example. Thirty patients were randomized to receive either an experimental drug ($n_t = 15$) or a placebo ($n_c = 15$) for the treatment of epileptic seizures. The number of seizures in a 2-week period (x_i) for each patient, as well their corresponding ranks r_i, are summarized in Table 4.

The test statistic U is

$$\min\left[\left([4+4+6.5+\cdots+28+28] - \frac{15(16)}{2}\right), \left([1+2+4+\cdots+28+30] - \frac{15(16)}{2}\right)\right] = 102.5,$$

Table 4 Data, ranks, and treatment indicator variable for an epileptic seizure study

Experimental Drug															
i	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
x_i	5	5	6	7	7	7	8	9	10	10	11	11	11	13	13
r_i	4	4	6.5	10	10	10	14	17	20	20	24	24	24	28	28
T_i	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Placebo															
i	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
x_i	3	4	5	6	7	7	8	9	9	9	10	11	11	13	14
r_i	1	2	4	6.5	10	10	14	17	17	17	20	24	24	28	30
T_i	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

and the associated p value is

$$p = 2\Pr\left(z > \frac{|102.5 - 15(15)/2|}{\sqrt{15(15)(31)/12}}\right) = 2\Pr(z > .41) = .68.$$

The investigator cannot claim that the distributions of the number of seizures in a 2-week period for patients given the new medication and patients given a placebo are different.

Jarrod E. Dalton

See also Analysis of Variance (ANOVA); Confidence Intervals; Sample Size and Power

Further Readings

Miller, I., & Miller, M. (1999). *John E. Freund's mathematical statistics* (6th ed.). Upper Saddle River, NJ: Prentice Hall.

SAS Institute. (2003). *SAS OnlineDoc® 9.1*. Cary, NC: Author.

Siegel, S., & Castellan, N. J. (1988). *Nonparametric statistics for the behavioral sciences*. New York: McGraw-Hill.

Steady-State Models

Dynamic modeling is important to many areas of medical decision making. Cancer screening, infectious disease transmission, demographic modeling of healthcare and pension costs, economic growth, and budget forecasting provide a few arenas in which dynamic modeling plays a critical role. Differential or difference equations provide a framework for many of these problems.

In many cases, analysts can define differential equations that cannot be solved analytically at each moment in time. Fortunately, the long-term behavior of these systems can be explored or characterized through steady-state analysis.

This piece presents the basic character and limitations of steady-state analysis. It begins with some general background. It then discusses a specific steady-state analysis: the random-mixing model of infectious disease transmission in a population of injection drug users (IDUs).

Background

Suppose one derives some differential equation governed by Equation 1 below:

$$\frac{dX}{dt} = f(X) \tag{1}$$

The time rate of change in X is given by some function $f(X)$. Given our interest in steady-state behavior, we have assumed that the differential Equation 1 does not explicitly depend on time. However, Equation 1 has greater generality than is first apparent.

For example, when $\boldsymbol{X}$ is the vector $[x(t), dx/dt, d^2x/dt^2, \mathrm{K}]$, one can model higher-order equations. In this framework, the equation

$$\frac{d^2x}{dt^2} = -4x \tag{2}$$

can be represented as

$$\frac{d}{dt}\left[x(t), \frac{dx(t)}{dt}\right] \quad \frac{d}{dt}[x_1(t), x_2(t)] \quad f(X) \quad [x_2, \quad 4x_1]. \tag{3}$$

Here, we use the expressions $x_1(t)$ and $x_2(t)$ to indicate two different functions that depend on time t and that evolve over time in accordance with Equation 3.

Steady-state values satisfy the relationship $f(\boldsymbol{X}) = 0$. When some $\boldsymbol{X}$ satisfies this relationship in a deterministic model, once the system reaches this value, it will stay there. In the case of Equation 3, $x(t) = dx/dt = 0$ is the unique steady-state solution. A system that starts at this value will stay at this value indefinitely.

Many physical systems converge on a steady-state solution if one waits a sufficiently long period of time. Yet as the above example indicates, there is no reason any physical or economic system necessarily tends to the steady state. Readers may recognize Equation 2 as embodying Hooke's law. A mass attached to a spring experiences a force proportional to its displacement from its rest position. It is readily demonstrated that any function of the form

$$X(t) = A\sin(2t) + B\cos(2t) \tag{4}$$

will satisfy Equation 2, with A and B chosen to match the mass's initial position and velocity.

Unless $A = B = 0$, this system will oscillate indefinitely, without converging to any steady-state position. Equation 5 below is even worse:

$$\frac{d^2x}{dt^2} = +4x. \tag{5}$$

Unless $x = dx/dt = 0$, Equation 5 describes a system that diverges to $\pm\infty$. Undergraduate texts in differential equations provide further information about the existence, uniqueness, and stability of steady-state solutions.

Strengths and Weaknesses: Infectious Disease Transmission Among Injection Drug Users

The remainder of this entry illustrates the value of steady-state analysis by exploring the "random mixing" model of infectious disease spread. It then applies this framework to a particular application: HIV and hepatitis C transmission among IDUs. This critical public health challenge provides a simple framework to exhibit both the strengths and the limitations of steady-state analysis.

More elaborate discussion is provided elsewhere (e.g., Pollack, 2001, 2002). This model seeks to capture two interrelated processes: entry and exit from the population of active IDUs and the process of mixing and infection spread among active injectors. It also seeks to explore how the introduction of substance abuse treatment can alter both the overall population of drug users and infectious disease prevalence.

This framework illustrates that real systems can have multiple steady states. Consider the differential Equation 6 below. This is the simplest random mixing model used in the study of infectious disease spread.

$$\frac{dx}{dt} = \kappa\lambda x(1-x) - \delta x. \tag{6}$$

Here, $x(t)$ corresponds to the proportion of a given population that is infected with the disease. Equation 6 defines the simplest random-mixing model of infectious disease spread. The motivating idea is that each drug user shares needles with some randomly selected member of the drug-using population. Individuals do this at rate λ per unit time. If an uninfected person shares a needle with an infected person, there is some probability κ that the infection will be spread. Each person, infected or uninfected, leaves the population at the same exit rate δ per unit time.

The term $\kappa\lambda x(1-x)$ captures infectious disease incidence: the rate per unit time that infected and uninfected people encounter each other and an infection occurs. When prevalence is very low ($x \approx 0$), few new infections occur. Susceptible people are unlikely to encounter others who could spread the infections. When prevalence is very high ($x \approx 1$), few new infections occur, because infected people encounter few susceptible people likely to be infected. The term $-\delta x$ captures the removal of infected individuals from the population, for example, due to disease mortality.

Equation 6 and simple variants can be solved exactly, yielding a classic logistic curve. However, the mathematical details quickly become intricate. Steady-state analysis often yields the critical policy insights within a much simpler framework.

Setting $dx/dt = 0$, we find the steady state by factoring out $x(t)$ and solving

$$x[\kappa\lambda(1-x)\ \theta = 0. \tag{7}$$

There are two steady-state solutions to Equation 7. The value $x_1 = 0$ always provides one steady-state solution. When $\kappa\lambda > \delta$, a second steady-state solution is also pertinent: the case $x_2 = 1 - \delta/(\kappa\lambda)$. When $x(0)$ takes on any positive value, $x(t)$ will converge over time to the value x_2.

Notice that when $\kappa\lambda < \delta$, steady-state prevalence goes to 0. This corresponds to the case where the disease will die out, and the population maintains *herd immunity*. When herd immunity is maintained, the epidemiological environment removes infected individuals from the population faster than new infections occur. So steady-state prevalence is 0.

This quantity, $\kappa\lambda/\delta$, plays a fundamental role in epidemiological modeling. It is frequently labeled the reproduction number R_0. The reproduction number has a rather simple intuitive meaning. Imagine that a single infected drug user is placed within a completely susceptible population. That individual will share needles some λ occasions per unit time, infecting some fraction κ of her partners. She will be in the active drug-using population for some period $(1/\delta)$.

Thus, the expected number of people she infects is the quantity $R_0 = (\kappa\lambda/\delta)$. For an outbreak to occur, she must at least replace herself with one other infected person. So if $R_0 < 1$, no outbreak will occur. More complex models yield more complex reproduction numbers. The result remains the same. When $R_0 < 1$, steady-state prevalence I^* falls to 0 in steady state. When local outbreaks occur, some individuals may become infected. Yet these local outbreaks will not cause a population-wide epidemic.

Random-mixing models have been applied and extended to a large class of biological and public health concerns, such as immunization. Interested readers are referred to the classic text by Anderson and May.

Evolution of the Drug-Using Population

Before considering any infectious disease epidemiology, we start by considering the overall population of $N(t)$ active IDUs. Every day, θ uninfected individuals enter the drug-using population. Active IDUs also leave the population and at different rates depending on their receipt of treatment services. We assume an exit rate of some ω per person per day among individuals receiving substance abuse treatment and some δ per person per day within the remaining population of IDUs. If there is excess demand for treatment services, and there are $M(t)$ clients receiving services at time t, we have Equation 8:

$$\frac{dN}{dt} = \theta - [N(t) - M(t)]\delta - M(t)\omega. \qquad (8)$$

Here, $M(t)$ is a policy variable. Given a specific time path $M(t)$, one could solve for $N(t)$ explicitly. If $M(t)$ is simply some constant M, this equation is readily solved analytically for all time t.

It is even easier to solve Equation 8 over the long run, when $N(t)$ asymptotes to some steady-state N^*. Note that this equation is only valid when there is excess demand for treatment. That is, $N(t) > M(t)$.

When steady state is reached, the time rate of change in N approaches 0. This implies that

$$0 = \theta - [N^* - M]\delta - M\omega. \qquad (9)$$

Doing the algebra,

$$N^* = M + \frac{\theta - M\omega}{\delta}, \quad \text{when } \theta \geq M\omega. \qquad (10)$$

Equation 10 is readily interpreted. There are θ users entering the population every day. There are $M\omega$ users exiting the population from substance abuse treatment and $(N^* - M)\delta$ users exiting the population from the out-of-treatment group. Because we are in steady state, the inflow and the outflow must offset each other. So $\theta = M\omega + (N^* - M)\delta$, which leads to Equation 10.

Equation 10 also informs us when it is valid. We assumed above that $N^* \geq M$. So this analysis is only operative when $(\theta - M\omega) > 0$. Per unit time, the number of treatment recipients who exit the drug-using population can be no greater than the number of new arrivals.

Now, what happens when $\theta < M\omega$? In this case, regardless of the initial condition, if we wait long enough, $N(t) < M$ for all $t > t^*$. After this point, Equation 8 becomes

$$\frac{dN}{dt} = \theta - M\omega. \qquad (8')$$

Since the right-hand side is a negative constant, $N(t)$ simply declines linearly to $N^* = 0$.

$$N^* = 0, \text{ when } \theta < M\omega. \qquad (10')$$

Again, the reasoning is straightforward. The population of active drug users drops to 0 because more people are being cured, per unit time, by treatment than are arriving into the population.

Infectious Disease Transmission in the Absence of Treatment

We now add infectious disease transmission to the model. We first do so in the absence of treatment ($M = 0$). We then examine how treatment transforms the problem.

We now turn to the dynamics of infectious disease spread. At any time t, some number $I(t)$ of active drug users are infected with some disease. $I(t)$ is comparable with $x(t)$ above, multiplied by N, the population size of injection drug users. Thus, the remaining $N(t) - I(t)$ drug users remain uninfected and at risk.

Let us consider the evolution of $I(t)$. Assuming that there is no drug treatment, there are $N(t) - I(t)$ susceptible individuals. Each susceptible individual shares needles λ times per week. In each of these encounters, random mixing implies that the sharing "partner" has a probability $I(t)/N(t)$ of being infected (i.e., the proportion of infected people within the drug-using population). When a susceptible person shares with an infected person, there is some probability κ that the infection will spread. Moreover, every week, some $\delta I(t)$ infected drug users leave the population.

Pulling these together, this implies that

$$\frac{dI}{dt} = -\delta I + \kappa\lambda(N(t) - I(t))\left[\frac{I(t)}{N(t)}\right]. \tag{11}$$

What happens in steady state? $I(t)$ approaches some asymptote I^*, and $N(t)$ approaches some other asymptote N^*. So we have

$$0 = -\delta I^* + \kappa\lambda(N^* - I^*)\left[\frac{I^*}{N^*}\right]$$

or

$$I^* = \kappa\lambda(N^* - I^*)\left[\frac{I^*}{N^*}\right]\left(\frac{1}{\delta}\right). \tag{12}$$

The left-hand side of Equation 12 is I^*, *steady-state prevalence* of the disease. On the right-hand side, $1/\delta$ can be interpreted as the average duration an individual persists in the population postinfection. The remainder of the right-hand side represents *steady-state incidence*, the number of new infections per unit time. In steady state, Equation 12 includes a classic result of infectious disease epidemiology: *Prevalence equals incidence multiplied by duration.*

As in the previous section, we solve for steady state:

$$I^*\left[\kappa\lambda\left[\frac{N^* - I^*}{N^*}\right] - \delta\right] = 0. \tag{13}$$

If steady-state prevalence is positive, we can divide both sides by I^*. Since we assume no treatment, $N^* = \theta/\delta$. We can then find that

$$I^* = \left[1 - \frac{\delta}{\kappa\lambda}\right]N^* = \left[1 - \frac{\delta}{\kappa\lambda}\right]\frac{\theta}{\delta}. \tag{14}$$

The bracketed term represents the proportion of the active drug-using population that carries the infection. Since prevalence equals incidence times duration, the steady-state number of new infections per unit time is

$$\iota = \delta I^* = \theta\left[1 - \frac{\delta}{\kappa\lambda}\right]. \tag{15}$$

A Combined Model

Now we combine the above models. For simplicity, we assume that $M(t)$ is some constant M and that the proportion of infected people in treatment matches the overall population. We'll also assume that treatment is perfectly effective in that no treated people spread the disease or become infected with it. This implies that

$$\frac{dI}{dt} = -\delta I\left[\frac{N - M}{N}\right] - \omega I\left[\frac{M}{N}\right] + \kappa\lambda\left[\frac{N - M}{N}\right](N - I)\left[\frac{I}{N}\right], \tag{16}$$

$$\frac{dN}{dt} = \delta(N - M) - \omega M. \tag{17}$$

For simplicity, assume that $N^* > 0$. There are a positive number of drug users in steady state. Solving these out, we have

$$N^* = M + \frac{\theta - M\omega}{\delta}, \tag{18}$$

$$\frac{I^*}{N^*} = 1 - \left[\frac{\delta}{\kappa\lambda}\right]\left[\frac{\delta(N^* - M) + \omega M}{\delta(N^* - M)}\right] = 1 - \left[\frac{\delta}{\kappa\lambda}\right]\left[\frac{\theta}{\theta - M\omega}\right]. \tag{19}$$

These equations allow us to answer useful questions. For example: How many treatment slots are required to achieve herd immunity? We explore this question by setting the right-hand side of Equation 19 equal to 0. In epidemiological terminology, the reproduction number in this model declines from R_0 to

$$R_1 = \left[\frac{\kappa\lambda}{\delta}\right]\left[\frac{\theta - M\omega}{\theta}\right] = R_0\left[\frac{\theta - M\omega}{\theta}\right]. \quad (20)$$

We require $R_1 < 1$ for herd immunity. Some algebra yields that M must be large enough for

$$M\omega > \theta\left[1 - \frac{\delta}{\kappa\lambda}\right]. \quad (21)$$

To achieve herd immunity, we must provide a sufficient number of slots to ensure that the exit rate from substance abuse treatment at least matches the rate of new infections in the absence of intervention.

Similar results hold for immunization. Under the same general assumptions, one can create herd immunity by vaccinating the same proportion of the population that would acquire the infection, absent the intervention.

Additional Analyses

When steady-state analysis captures the key features of a dynamic system, it provides a simple and useful approach to policy analysis. For example, suppose that society values the prevention of each new infection at some $\$D$, and suppose that the cost of each substance abuse treatment "slot" is $\$C$. The social planner then picks the number of treatment slots M to minimize

$$\min_{(M)} L = CM + D\delta I^*. \quad (22)$$

The optimal provision of treatment may or may not be sufficient for herd immunity.

Limitations

Despite their utility, steady-state models face several important limitations. One central issue concerns the time a system takes to approach steady state from a given initial state. This is sometimes operationalized as the time required for $I(t)$ to move from 10% to 90% of its steady-state value. When this process takes many years or decades, steady-state models are clearly limited in their ability to inform current policies.

Within the random-mixing model, the time required to reach steady state is short when the reproduction number R_0 is high. Convergence times can be much longer when the reproduction number is low. Policy analysis suggests that some agents, such as hepatitis C, spread quite rapidly in drug-using populations. Steady-state analysis based on the random-mixing model works quite well within these populations. In the case of HIV, the epidemic process works more slowly. One should therefore base policy analysis on the full time history of the epidemic system.

As noted before, systems can have no pertinent steady state. Alternatively, multiple steady states may exist. Each system must be analyzed in light of these possibilities. Recent work by Caulkins and collaborators provides some of the most sophisticated pertinent applications.

Harold A. Pollack

See also Markov Processes

Further Readings

Anderson, R. L., & May, R. M. (1992). *Infectious diseases of humans: Dynamics and control*. Oxford, UK: Oxford University Press.

Daley, D. J., & Gani, J. (1999). *Epidemic modeling: An introduction* (Cambridge Series in Mathematical Biology, Vol. 15, C. Cannings, F. Hoppensteadt, & L. Segal, Eds.). Cambridge, UK: Cambridge University Press.

Feichtinger, G., Caulkins, J. P., Tragler, G., Behrens, D. A., & Graß, D. (2008). *Optimal control of nonlinear processes: With applications in drugs, corruption, and terror*. Heidelberg, Germany: Springer.

Hochstadt, H. (1964). *Differential equations: A modern approach*. New York: Dover.

Pollack, H. A. (2001). Can we protect drug users from hepatitis C? *Journal of Policy Analysis and Management, 20*(2), 358–364.

Pollack, H. A. (2001). Cost-effectiveness of harm reduction in preventing hepatitis C among injection drug users. *Medical Decision Making, 21*, 357–365.

Pollack, H. A. (2002). Methadone treatment as HIV prevention: Cost-effectiveness analysis. In E. H. Kaplan & R. Brookmeyer (Eds.), *Quantitative evaluation of HIV prevention programs* (pp. 118–142). New Haven, CT: Yale University Press.

Stigma Susceptibility

In the daily medical context, stigma is ubiquitous. On the one hand, health professionals are susceptible to having inappropriate attitudes or expectations concerning patients and their families. On the other hand, stigmatized individuals are susceptible to receiving inadequate treatment or experience disadvantages by the behavior of health institutions. The question arises whether the concept of "being ill" itself is already stigmatized. This entry begins by providing a short definition of stigma followed by a description of the stigmatization process. Thereafter, various examples are given, applying the concept of stigma to decisions made in the medical context. Finally, some suggestions are made to give the reader an idea of how to avoid decisions driven by a stigmatized approach.

The Concept of Stigma

Since the concept of stigma is multidisciplinary, it has been applied to a vast amount of events. Within the contributions of many disciplines, different theoretical approaches were used putting different emphasis on its conceptualization. Bearing this ambiguity in mind, the following attempt to describe stigma is done particularly with regard to its application in the medical context.

Stigma is differentiated in several perspectives. One important distinction is made between public stigma and self-stigma. The first implies stereotypical perceptions by the public. For instance, the health insurance system and other health providers or professionals all inherit specific attributes when dealing with a stigmatized person. Self-stigma, however, subsumes the behavior of the stigmatized individual himself or herself. Therefore, it may affect the well-being, healthcare choices, and even life goals of the person involved. Self-stigmatization generally results from a previously experienced public stigma.

An additional perspective can be derived by an early attempt of Erving Goffman to differentiate observable marks of stigma, the discrimination between discredited and discreditable stigma. Discredited stigma refers to perceivable marks of the individual in question. Hence, the stigmatized individual is labeled by physiognomy or behavior, not having the opportunity to hide these marks from the public (e.g., a blind person is easily identified by his or her white cane). In contrast, a discreditable stigma is characteristic for individuals who have the possibility to hide their condition in front of others. For instance, this could be the case for patients who suffer from mental illnesses, cancer in an early stage, or HIV.

The Process of Stigmatization

Several sociocognitive processes contribute to stigmatization. At first, social and physical cues are subsumed into a category or a label. Every dimension on which people vary can be selected during this process of categorization (e.g., gender, age, race, social class, physical health). For the purpose of information reduction, individuals are mostly categorized by only a single or a few dimensions—although they belong to many. Once a person is categorized, it appears that all social interactions are pervaded by this category (e.g., the blind lawyer). Labeling is a similar process, but based on categories which involve rather vague membership criteria (e.g., mental illness). Labels can only be obtained from three sources: (1) the information given by others (e.g., doctors, nurses), (2) the information given by the stigmatized person, and (3) observed associations (a person coming out of a psychiatrist's office).

Second, to determine whether or not a person belongs to a certain category based on concealable criteria is the stereotype of this category. Stereotypes are knowledge structures that provide particularly useful categorization information about marked social groups (e.g., people with schizophrenia). They enable the general public to learn about collectively agreed characters subsumed under specific groups of persons. On the other hand, simplification has its price: Stereotypes may lead to unjustified overgeneralization and misjudgment.

The third process describes the endorsement of negative stereotypes: prejudice. In addition to the stereotypic beliefs, prejudice involves an evaluative component which is mostly negative. Finally, the rather affective response of prejudice leads to discrimination, the fourth process. During the discriminatory process, the stigmatized person is marked as a member of an out-group—through this, a clear separation takes place. Therefore,

stigma can lead to negative action against the outgroup (e.g., the dismissal of employees with HIV) or exclusively positive action for the in-group (e.g., the refusal of applicants with HIV to protect the staff). Most likely, this treatment implies a severe loss of status for members of a stigmatized group.

This whole process, however, socializes the stigmatized person into the attributed role and can lead to a self-fulfilling prophecy: Once stamped by a stigma, one may act according to the expected behavior and continue to fit the ascribed label. For instance, in a study about the stigma of psychotherapy, clients tended to behave in a socially unfavorable manner when interacting with persons who knew about their need for psychotherapeutic support.

Stigma in the Medical Context

A vast number of diseases are occupied with stigma. It would go beyond the scope of this entry to give a complete overview of all possible occasions applying stigma in the medical context. Therefore, only some examples will give an idea about the interaction of illness, stigma, patients, and health professionals. In general, there exists an inverse relationship between public stigma and healthcare seeking. Those people who tend to blame others for a stigmatized illness often avoid healthcare utilization when suddenly struck by the same illness.

In the case of mental illness, the diagnosis itself pervades all subsequent behavior of the person. Patients suffering from mental illness experience many variations of stigmatization, from problems regarding the workplace to a systematic exclusion of the society. As a consequence, feelings of guilt and lowered status reflect a self-ascribed stigma and may support its perpetuation. In addition, evidence indicates that people with mental illness generally receive fewer medical services than those who do not have mental illness. On the other side, family members are sometimes stigmatized by health professionals of having caused their relative's illness. But mental health stigma does not only affect patients, it also overlaps with the demoralization of professionals. Practitioners in mental health services often experience stigma and feel underappreciated by their clients and society in general (think of the term *headshrinker*). Hence, medical graduate students often avoid training in mental health disciplines.

However, there are also many other diseases bearing a stigma. Just imagine what comes to mind when thinking about AIDS. Stereotypes about gay people, drug addicts, or unsafe sex may lead to an attribution of the blame to the sick person. Indeed, the stigma of AIDS remains a significant barrier to HIV prevention and treatment, from the barriers to getting tested to obtaining optimal HIV care or even safe-sex practices. Moreover, HIV stigma was found to be related to depression, poor adherence, and the degrading disclosure of serostatus. In a similar vein, patients suffering from lung cancer report problems coping with the stigma of their disease. Those who had never smoked before or even stopped smoking years ago conveyed the impression that other people blamed them for their illness. Thus, they feared access to healthcare services and asking for support from other people. Essential behavior to seek help was affected by the stigma of the disease.

Stigma in the medical context is not only related to illnesses. Medical treatment and access can vary among different groups of people according to their experienced prejudices and stereotypes. For example, gender as a discredited category also leads to differentiated treatment. Although women are found to spend more time at a physician's consultation, men receive more specific therapies under the same diagnosis and a higher number of follow-ups. Another source of discredited stigma is ethnicity. For example, previous unfair treatment in healthcare and experience of hostility toward blacks leads to the preference of same-race doctors among blacks. These findings suggest that the ethnic background influences different approaches to healthcare. A study focusing on mental illness showed that women of ethnic minorities sought mental healthcare services to a lesser extent than U.S.-born white women—due to their stigma-related concerns about mental healthcare.

Stigma Awareness

Stigma in the medical context can be regarded as omnipresent. Due to its appearance in so many health issues, it seems that illness itself is already a stigma. Exposed to the power of a health institution, a person may feel treated as if he or she is not of the same status as the medical personnel. Being marked as a patient implies not only a loss

of status but also the impossibility to escape this label while staying in the health institution. Once diagnosed, decisions of caretakers tend to be unshiftable, as they typically try to confirm previous assumptions. An example can be given by an early study, where pseudopatients gained admission to several hospitals pretending that they would hear voices. After the health personnel had assigned the diagnosis of schizophrenia, the volunteers acted perfectly normal and tried to convince the staff that they were sane. However, the pseudopatients' behavior was consequently perceived as insane, confirming the primarily assumed illness. For instance, when a pseudopatient was pacing along the corridor, the medical staff interpreted this as chronic tension. Labeling and stigma have a strong impact on the processing of information and can hardly be diminished—one unfortunate volunteer was only discharged after 52 days.

Taken together, stigma should never be underestimated in the medical context. Healthcare would benefit from changing attitudes toward stigmatized illnesses to avoid wrong medical decisions. This could happen through informative programs, contact with patients during education for health professionals, and especially the everyday attempt to realize and overcome active stigmas (at least) within the medical professional's mind.

Stephanie Müller and Rocio Garcia-Retamero

See also Confirmation Bias; Diagnostic Process, Making a Diagnosis; Errors in Clinical Reasoning; Uncertainty in Medical Decisions

Further Readings

Chapple, A., Ziebland, S., & McPherson, A. (2004). Stigma, shame, and blame experienced by patients with lung cancer: Qualitative study. *British Medical Journal, 8,* 1470–1473.

Cooper, A., Corrigan, P. W., & Watson, A. C. (2003). Mental illness stigma and care seeking. *Journal of Nervous and Mental Disease, 191,* 339–341.

Corrigam, P. W. (2005). *On the stigma of mental illness: Practical strategies for research and social change*. Washington, DC: APA.

Feder-Alford, E. (2006). Only a piece of meat: One patient's reflections on her eight-day hospital experience. *Qualitative Inquiry, 12,* 596–620.

Goffman, E. (1963). *Stigma: Notes on the management of spoiled identity*. Englewood Cliffs, NJ: Prentice Hall.

Herek, G. M., Capitanio, J. P., & Widaman, K. F. (2003). Stigma, social risk, and health policy: Public attitudes toward HIV surveillance policies and the social construction of illness. *Health Psychology, 22,* 533–540.

Kaplan, A. H., Scheyett, A., & Golin, C. E. (2005). HIV and stigma: Analysis and research program. *Current HIV/AIDS Reports, 2,* 184–188.

Krishnatray, P., Melkote, S. R., & Krishnatray, S. (2006). Providing care to persons with stigmatised illnesses: Implications for participatory communication. *Journal of Health Management, 8,* 51–63.

Link, B. G., & Phelan, J. C. (2001). Conceptualizing stigma. *Annual Review of Sociology, 27,* 363–385.

Llerena, A., Cáceres, M. C., & Peñas-Lledó, E. M. (2002). Schizophrenia stigma among medical and nursing undergraduates. *European Psychiatry, 17,* 298–299.

Stochastic Medical Informatics

Stochastic medical informatics is an approach to reasoning about clinical phenomena that manages the inherent uncertainty and complexity through statistical methods such as random sampling from probability distributions. The representation of information, such as the possible clinical outcomes for patients, and reasoning about it can be highly complex and computationally overwhelming. Simplifying assumptions must be made to manage such information to make useful predictions and rational decisions. Observations of patient populations and possible events can often be described conveniently in terms of conditional probability distributions. Such statistics can then be assembled into rational, decision-analytic or simulation models that can then be subjected to systematic analyses for making predictions and policy decisions.

Decision Making Under Uncertainty

Medical decision makers may want to answer a variety of questions about individual patients or populations. This may require comparing

alternative treatment strategies in terms of their cost and effectiveness or predicting the likely frequency of outcomes that have differing chances of occurrence, such as the side effects of treatments. At the level of the individual patient, such comparisons can assist in making treatment decisions. At the population level, policy decisions can be made that optimize the quality of care and allocate limited resources. Such policy decisions may also include the value of obtaining more accurate information on which to make more optimal subsequent decisions or invest in more effective implementations of policies.

Answers to such questions can be found by modeling and simulating the interactions of probable events in various treatment scenarios that may be subject to different conditional criteria over time.

All statistical models are based on observations of the real world. Uncertainty can arise in a number of ways, from the intrinsic variability of the underlying phenomena being measured through the imprecision in the measurements themselves and how they are assembled into a model. Such measurements are typically treated as "random variables," unknown values that may be approximated by point estimate statistics, such as averages over observed values, or distributions describing the frequency of observations over a range of possible values. For example, the probability that an event occurs can be estimated by a single number between 0 and 1 (0% and 100%), while a simulation for a group of patients may be based on an average age. Alternatively, these estimates can be expressed as distributions over a range of values, reflecting the frequency of observations. Probability distributions can be represented similarly in the form of probability density functions, which may be visualized as graphs for which the area under the curve sums to a total probability of 1 (100%). For example, the probability of an event might be described using a beta function (with parameters $\alpha = 2$, $\beta = 5$), whose probability density function is shown in Figure 1. Alternatively, a simpler model may use a point estimate for this probability, such as the mean (.28) or mode (.2) of this distribution. Point estimates or distributions may also be formulated as functions of other parameters in a model, such as time or age, as implemented through table lookup or function calculation.

Models based entirely on constant measurements are referred to as *deterministic models* and are significantly easier to evaluate and analyze, yielding reproducible results. Models based on the distributions of random variables are referred to as *stochastic models* and provide an additional degree of expressiveness at the expense of increased difficulty in evaluation. Due to their complexity, these models are typically evaluated by averaging the results of repeatedly drawing samples to approximate the underlying distributions, often yielding results that are not exactly reproducible.

Additional uncertainty arises from how such values can be extrapolated or accuracy of predictions made with a particular model. Many of these sources of uncertainty can be handled using sensitivity analysis, considering the effect of varying the values of the underlying estimates.

Model Components

Virtually all the types of statistical models used in medical informatics represent the following components:

Patients, Populations, and Clinical Settings—in terms of their various characteristics and frequencies. These may form the initial health states or parameters of a model. While individual patients may have fixed characteristics, populations must often be described in terms of distributions over the possible values of those characteristics.

Events—which may occur nondeterministically with varying frequency and which may result in transitions between health states over time.

Outcomes—associated with transitions between or terminal health states, typically assessed in terms of utilities that may include life expectancy, often adjusted for quality and measured in quality-adjusted life years (QALYs) or disutility-adjusted life years (DALYs), and costs incurred. Quality measurements of each outcome may reflect individual preferences that may be elicited from patients using a variety of assessment techniques.

Each of these components may be described statistically in terms of random variables having some degree of uncertainty. Furthermore, each of

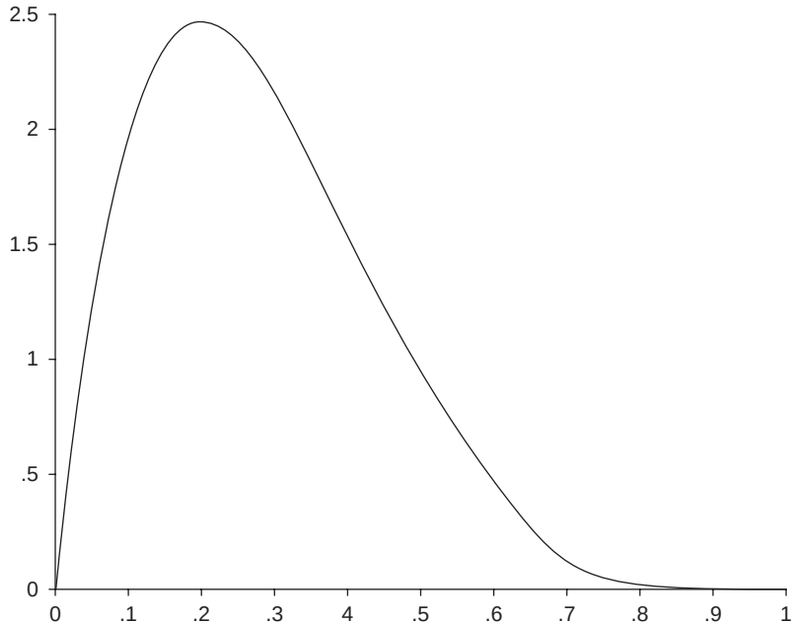


Figure 1 The probability density function for the beta function with $\alpha = 2, \beta = 5$

the probability or utility estimates is often conditional on health states, previous events, or time within a simulation.

Modeling Techniques

A variety of modeling formalisms are used in medical informatics, with varying data requirements and purposes. These include the following.

Decision Trees

Decision trees are models in which alternative decisions and possible events are organized as successive nodes, each with a number of branches, forming a tree structure. Probabilistic events are represented by nodes whose alternate branches each occur with a given probability for the conditional context as it appears in the tree. Leaves of the tree represent outcomes that are associated with utility values. The fundamental evaluation technique of such trees, referred to as Foldback, involves the computation of the weighted average of utilities according to their relative probability. According to decision theory, rational decisions are made to optimize the utilities of their corresponding subtrees.

Influence Diagrams and Bayesian Belief Networks

Influence diagrams and Bayesian belief networks are organized as directed graphs in which each decision, chance, or outcome variable is represented by a single node, and arcs between nodes represent the influence of one variable on another. The value of each influence arc can be conditional on any influences of its source node. Such models are evaluated deterministically by Belief Propagation techniques, which compute the belief that each variable is true. While Bayesian belief networks lack the construct of a decision node, each influence diagram has a corresponding decision tree in which each event or outcome is represented by a node in the tree for each conditional case.

Markov and Semi–Markov Models

Markov and semi–Markov models (Markov chain Monte Carlo methods) are organized around a set of possible health states, each having one or more corresponding utility values for each discrete time interval spent in that state, and probabilities for which transitions to other states occur over time. In a graphical representation, states are represented as nodes with state transitions denoted by arcs between them. Figure 2 depicts a three-state model

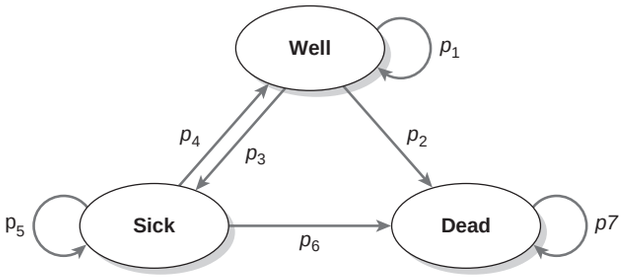


Figure 2 A Markov-state diagram

in which patients may alternate between "Sick" and "Well" states, with transition probabilities p_i until they are absorbed into the "Dead" state.

By assuming that state transitions depend only on the current state and time and are not conditional on any previous states, the Markov assumption, computational complexity can be significantly reduced, sometimes to deterministic solutions. However, this assumption is sometimes relaxed to result in semi–Markov models, which, while organized similarly around states, may include additional "micromodeling" variables on which state transitions may depend. As a result, semi–Markov models require evaluation by simulation, such as first-order Monte Carlo analysis. Additional expressive power and flexibility can be achieved by incorporating Markov, and semi–Markov models can also be incorporated into a decision tree framework, by representing each state transition of the Markov as a decision tree whose terminal nodes are assigned to subsequent states. Alternatively, a graphical representation analogous to influence diagrams or purely statistical representations of Markov models can also be used. Such models can be evaluated using Markov chain Monte Carlo methods such as Gibbs Sampling, in which underlying but unknown joint distributions are approximated by taking samples from univariate conditional distributions.

Markov Decision Processes

Markov decision processes are generalized Markov models that can explicitly involve embedded decisions and are represented as sets of states, actions, transition probabilities, and rewards (or costs) for each state transition. These models can be solved to arrive at optimal decision policies, specifying the best action to take in a given state, to maximize the cumulative utility. Due to their combinatorial complexity, such models must often be evaluated using heuristic search techniques in which some subspace of possible paths, as determined by clinical decision policies, are searched by minimizing or maximizing some guiding criterion.

Discrete Event Simulations

Discrete event simulations are a class of models that also involve the representation of alternative health states but in which state transition events occur at any point in time, according to their corresponding transition probabilities. These models are typically used to compare policy decisions when applied to entire populations that compete for limited resources that are also taken into account by the model. Due to the inherent uncertainty in the occurrence of events, the evaluation of such simulations must also involve statistical sampling techniques, as with complex Markov models, by averaging the results of repeated simulations with new samples drawn at each iteration.

Analysis Techniques

When only point estimates of probabilities are used, many of these types of models can be evaluated deterministically to result in reproducible, optimal answers, using techniques such as Foldback or Belief Propagation. However, when models considering an overwhelming number of possible choices or variables are estimated in terms of probability distributions, deterministic solutions cannot easily be achieved. In such cases, evaluation requires the employment of statistical sampling, stochastic simulation, or search for optimal policies through the potentially vast numbers of possible strategies. Monte Carlo evaluation techniques involve repeatedly reevaluating the models, drawing new samples from the underlying variable distributions each time, and reporting the final outcome results in terms of statistical distributions that include averages, standard deviations, and confidence intervals.

First-Order Monte Carlo Analysis

A first-order Monte Carlo analysis is essentially a simulation that repeatedly considers the

outcomes of one individual at a time. At each chance point in the simulation, the occurrence of possible events is determined by sampling from their corresponding probability distributions and making choices as if by rolling dice. Successive events result in final outcomes for each individual, and then, summary statistics are computed across individuals. This can be visualized in a decision tree by choosing a single path through the tree to an outcome, for each individual, and then averaging the results. The first-order analysis of the Markov model depicted in Figure 2 would involve repeatedly simulating the state of individual hypothetical patients as they make state transitions throughout some "time horizon," a number of cycles representing a period of time or individual's lifetime, and then averaging the expected utility of each patient.

Second-Order Monte Carlo Analysis

A second-order Monte Carlo analysis repeatedly considers the outcomes that occur in entire populations, simultaneously. Each repeated iteration of the simulation, also referred to as a *realization*, involves computing the proportion of the original population for which successive events occur and results in a distribution of the population over terminal health states and their corresponding outcome utilities. As in the first-order analysis, the conditional probability of each event is drawn from a distribution, but all such possible states, each having some proportion of the entire population, are computed simultaneously. This can be visualized in a decision tree by dividing an initial population into subgroups partitioned by successive levels of the tree. Each iteration results in a final distribution across health states, each associated with their respective utilities, and summary statistics can be computed across all the iterations. Gibbs Sampling is an efficient method for performing what is effectively a second-order Monte Carlo analysis, in which the proportions of the population in all possible states are considered simultaneously, typically using a matrix representation, repeatedly drawing new samples for state transitions at each iteration. In the example Markov model of Figure 2, a second-order simulation would involve following the membership of each state as a percentage of an entire population cohort, for some fixed number of cycles or until the entire population is absorbed into the "Dead" state.

Sensitivity and Threshold Analyses

Sensitivity analysis can provide insight into the degree to which the results of a model depend on one or more of its parameters. A one-way sensitivity analysis can be achieved deterministically by repeatedly evaluating a model for each of a range of values for an individual parameter. Threshold analyses involve a search for the specific threshold values of one or more parameters to the model, at which point one strategy becomes more effective or cost-effective than another or reaches a specified willingness-to-pay threshold by repeatedly varying the selected parameters. Multiway sensitivity analyses involve reevaluating the model for different values of two or more variables or performing a threshold analysis for a number of (dependent) variables for each iteration of a sensitivity analysis on a selected (independent) variable. Such searches are typically performed heuristically or systematically over the grid formed by the possible values, but since variables may be continuous, it is in general a nondeterministic optimization problem. However, sensitivity or threshold analyses are further complicated when underlying variables of the model are described in terms of distribution and must then be estimated by using Monte Carlo techniques. Stochastic or probabilistic sensitivity analyses (PSAs) are essentially analyses in which any number of parameters are drawn from suitable distributions of possible values.

Value of Information

Value of information calculations facilitate decision making at a resource allocation level by comparing the scenarios with different degrees of uncertainty in the net benefits of alternative treatments. By considering the possible loss in making wrong decisions, one may compute the expected value of perfect information (EVPI) and expected value of partial perfect information (EVPPI). This can assist decision makers in determining whether or not to invest in further research. One may also compute the expected value of perfect implementation (EVPIM) by considering scenarios in which prescribed policies are followed more diligently

and by considering how much to invest in achieving new compliance goals.

C. Gregory Hagerty and Frank A. Sonnenberg

See also Decision Trees: Evaluation With Monte Carlo; Decision Trees: Sensitivity Analysis, Basic and Probabilistic; Discrete-Event Simulation; Expected Value of Perfect Information; Influence Diagrams; Markov Models, Applications to Medical Decision Making; Markov Processes

Further Readings

Ades, A. E., Lu, G., & Claxton, K. (2004). Expected value of sample information calculations in medical decision modeling. *Medical Decision Making, 24*(3), 207–227.

Briggs, A. H. (2001). Handling uncertainty in economic evaluation. In M. F. Drummond & A. McGuire (Eds.), *Economic evaluation in health care: Merging theory with practice* (pp. 172–214). Oxford, UK: Oxford University Press.

Chapman, G. B., & Sonnenberg, F. A. (Eds.). (2000). *Decision making in health care: Theory, psychology, and applications*. New York: Cambridge University Press.

Gelfand, A. E. (2000). Gibbs sampling. *Journal of the American Statistical Association, 452,* 1300–1304.

King, J. T., Jr., Tsevat, J., Lave, J. R., & Roberts, M. S. (2005). Willingness to pay for a quality-adjusted life year: Implications for societal health care resource allocation. *Medical Decision Making, 25,* 667–677.

Story-Based Decision Making

Story-based decision making submits that individuals attempt to find meaning in their life experiences, and part of this meaning-making process is achieved through the construction of stories based on preexisting experiences and beliefs. When faced with an unfamiliar predicament, multiple stories can be created. The story that best fits the situation is selected, based on coherence, comprehensiveness, plausibility, and so on. As applied to medical decision making, facts and statistics are often presented in isolation and, therefore, may be difficult to understand or relate to an individual's situation. A story might be able to provide a meaningful contextual frame to guide an individual's decision-making process.

Stories as Decision Aids

Traditional, didactic patient education materials on healthcare prevention and treatment intended to inform patient decision making often lack both context and cultural sensitivity. There is evidence that these materials are not as effective with people of color and minority populations, often those at greatest risk. To reach these populations, an alternative approach is to use stories.

People have always learned about themselves and their past through stories; human brains are designed to process stories instinctively. The ease with which stories are understood may be particularly important for patients who feel emotionally overwhelmed and psychologically exhausted yet still have to face complex treatment decisions. An example of a particularly difficult decision is one that involves uncertain, future benefits combined with upfront, guaranteed costs, such as mild to severe side effects and/or the potential for unnecessary, invasive treatments. If patients are unprepared for such immediate costs and/or fail to understand the trade-off they face, they may not adhere to the proscribed treatment regimen or engage in the preventive behavior. However, if patients could hear the stories of those who have gone through this decision-making process before them and come out the other end, they may be more likely to absorb and apply the information to reach a more informed decision.

Stories of Cancer Survival

An example of a medical decision for which stories may be particularly helpful is whether to undergo adjuvant cancer treatment, such as chemotherapy, after completion of one's initial treatment for cancer, often surgical removal of the malignant tumor. This decision is extremely difficult for most patients. First, there may be no benefit to the treatment—surgery could have removed all the cancer cells. Second, if there were any cancer cells left, they would typically be at the molecular level and, consequently, undetectable. So even if the treatment worked, the patient may never even know. Third, duration of survival is only one of

the issues. All adjuvant therapies are toxic, so experiencing side effects is nearly certain. Given such uncertain benefits and guaranteed, upfront costs, there is no "right" decision. Patients' choices will depend on how they perceive the trade-off between experiencing the toxicity of treatment now versus the (unknown) possibility of increased survival in the future.

Clinicians can provide factual information about this trade-off and may even use a decision aid to facilitate deliberations. At a minimum, a decision aid will have two components: (1) visual representations of the risks, benefits, and outcomes for each option and (2) a discussion of patients' values.

An alternative to such traditional approaches is the use of stories. Stories can serve multiple purposes, including informative (e.g., what is the shortest route to the hospital), instrumental (e.g., how to deal with nausea from treatment), and emotional (e.g., coping strategies). But stories also have a fourth distinct purpose: It is only through stories that individuals know they are not alone. It is through stories that individuals are seen and heard by others. Therefore, storytelling may be an especially potent avenue for reaching populations with insufficient numeracy or health literacy skills and/or limited access to or historical mistrust of the medical establishment. Stories may affect such individuals through any of three mechanisms: (1) overcoming barriers, (2) improving information processing, and (3) providing surrogate social connections.

Overcoming Barriers

To affect cancer decision making, one must first overcome patients' resistance to and fear of the diagnosis and treatment of cancer. Stories may be uniquely positioned to do so for three reasons.

First, hearing the story of a patient who successfully completed cancer treatment may enhance patients' self-efficacy, instilling the belief that "if she did it, I can do it, too!" Poor self-efficacy is a common basis for the failure to engage in or maintain a health behavior.

Second, others' stories can help patients develop realistic expectations. This is vital for adjuvant therapy because of the high toxicity of treatment. Through others' stories, patients may be able to envision their own ability to adapt to possible short- and long-term side effects and even be better prepared to persevere if faced with possible negative outcomes.

Third, others' stories may be better at challenging and overcoming misperceptions. Hearing stories from survivors who have completed their treatment provides living proof that cancer patients can survive, thereby challenging potential misperceptions, such as the beliefs that (a) cancer is a death sentence and (b) treatment may be worse than the disease.

Cognitive Information Processing

When exposed to a didactic presentation of cancer facts, such as those in traditional decision aids and patient education materials, patients can (and do) quickly generate counterarguments as to why this information does not apply to them. Consequently, the information may never even be processed.

There are at least four reasons why patients may be less resistant to stories. First, it may be more difficult to generate counterarguments for a story (compared with a list of facts) because recipients' thoughts are devoted to imagining how the story will end. Therefore, the motivation to counterargue is lessened—they do not want to interrupt the flow of the story. Because counterarguments are not generated, the information is absorbed and processed, at least at some level.

Second, such stories represent the lived experience of another and, therefore, may be more difficult to discount. Third, because these stories are about patients' lived experiences, the message may seem more relevant and, thus, more likely to be absorbed and remembered (vs. generic educational material). Finally, given the fear and dread that surrounds the decision to undergo adjuvant therapy to treat cancer, stories may be perceived as less threatening than a document that explicitly lists all the risks and benefits of treatment.

Providing Surrogate Social Connections

A large body of research has found that face-to-face social support can have significant health benefits, both physical and psychological. More recent research has shown that the same benefits can accrue when communicating virtually with

(initially) anonymous others, as in online cancer support groups or chat rooms.

Though reading or hearing a story is not as interactive as direct communication, stories may still be able to provide a type of social support. This is indirectly evidenced by the increasing popularity of autobiographies and biographies that describe the personal experience of illness, with the vast majority focusing on cancer. Though each of these stories is about an individual's unique experience, they all share a need to find meaning in being diagnosed with cancer and to integrate their precancer and postcancer realities.

These stories may approximate a kind of social interaction that provides a sense of social support, especially if recipients identify with the authors and/or perceive them to be in an analogous situation. This type of writing has become prevalent enough that a new field has emerged to study it—pathography.

Hearing a survivor's story has the potential to go beyond the validation of cancer patients' experiences. Stories can help patients transcend the limitations of their individual minds and introduce alternatives that they might never otherwise have considered. However, further research is needed to examine the myriad effects, both positive and negative, that stories may have on the medical-decision-making process.

Julie Goldberg

See also Cognitive Psychology and Processes; Construction of Values; Context Effects; Cultural Issues; Decision Making in Advanced Disease; Decision Psychology; Judgment; Managing Variability and Uncertainty; Patient Decision Aids; Risk-Benefit Trade-Off; Risk Communication; Risk Perception; Treatment Choices

Further Readings

Frank, A. (1995). *The wounded storyteller: Body, illness, and ethics*. Chicago: University of Chicago Press.

Green, M., Strange, J., & Brock, T. (2002). *Narrative impact: Social and cognitive foundations*. Mahwah, NJ: Lawrence Erlbaum.

Greene, K., & Brinn, L. S. (2003). Messages influencing college women's tanning bed use: Statistical versus narrative evidence format and a self-assessment to increase perceived susceptibility. *Journal of Health Communication, 8*(5), 443–461.

Hunsaker Hawkins, A. (1993). *Reconstructing illness: Studies in pathography*. West Lafayette, IN: Purdue University Press.

Kreuter, M. W., Green, M. C., Cappella, J. N., Slater, M. D., Wise, M. E., Storey, D., et al. (2007). Narrative communication in cancer prevention and control: A framework to guide research and application. *Annals of Behavioral Medicine, 33*(3), 221–235.

Shapiro, S., Angus, L., & Davis, C. (1997). Identity and meaning in the experience of cancer: Three narrative themes. *Journal of Health Psychology 2*(4), 539–554.

Slater, M. (2002). Entertainment education and the persuasive impact of narratives. In M. Green, J. Strange, & T. Brock (Eds.), *Narrative impact: Social and cognitive foundations* (pp. 157–181). Mahwah, NJ: Lawrence Erlbaum.

Subjective Expected Utility Theory

Subjective expected utility (SEU) theory is a prescriptive theory of decision making that grew out of economics. The translation of economic concepts to medicine has a number of problems. Although SEU can assist with overcoming some of these problems, the value of SEU is primarily in helping the decision maker to structure the decision. Key concepts in SEU are decision making under risk, utility, and probability. These concepts will be briefly described first.

Key Concepts

Decision Making Under Risk

Decisions are normally choices between alternatives with different probabilities. Even if one chooses an alternative with a sure outcome, one risks rejecting an alternative with some chance of a better outcome. An important understanding is that whether a given decision is viewed as good or bad is not dependent on the outcome but rather on the process. In the world of risk and probability, unfortunately, there is no guarantee of an optimal outcome. Decision makers must make the best decision they can based on the information available to them at the time the decision is made.

Decision theory was derived from economic models, historically constituting gambles. Gambles

provide a paradigm for decision making, in which each alternative has a different value as well as a different probability. Structuring decisions as gambles allows for mathematical analysis of the decision so that a rational or best outcome can be prescribed. The first models were based on expected value (or average value), usually shortened to EV.

Value as a Decision Concept

Mathematical calculation of an optimal decision is easiest when the outcome is defined in terms of money. In that case, both alternatives have the same metric and the metric is quantified. It is taken as a primitive that more money is better than less. However, the amount that a decision maker might be willing to risk is not in a 1:1 relationship to the amount of the alternatives. Prospect theory has demonstrated that there is a decrease in the psychological value of a gain with the increase in overall value of both options. For instance, the subjective difference in the value of a gain of $10 is less when one is betting $2,010 against $2,000 than when one is betting $20 against $10. The opposite is true of loss, although the effect appears to accelerate more rapidly. However, calculation of an optimal choice becomes more problematic when the gamble involves an outcome that is nonmonetary, such as health or a specific physical function.

Probability in Decision Models

Rationality in economic theories of decision making assumes that there are multiple opportunities to make a decision, as is typical in most gambling situations. Over infinite replications of the decision, the best outcome is achieved through choosing the gamble with the largest arithmetic product of value and probability. The assumption underlying this EV model is that people seek to maximize the amount they will gain. Using a gambling paradigm, if one is offered a gamble of .5 probability (a coin flip) of $5 vs. 1.0 probability (a sure thing) of $1, the rational choice would be to take the former gamble because its EV is greater, that is, $5 \times \$5 = \$2.50$, versus $1.0 \times \$1 = \$1.00$ on average. Sometimes the $5 bet will win $5, and sometimes it will win $0, but over a very large number of gambles, on average, it will gain $5 half the time. On the other hand, the $1 gamble will always win $1, so its expected value = $1.00. This works with money, as long as one is able to bet a very large number of times.

When gambling for money, the probability of any outcome can be objectively established. However, probability for many healthcare decisions is not so precisely defined. To complicate matters, generally, medical decisions are made only once, and the patient must live, or sometimes die, with the consequences of the decision made.

Subjective Expected Utility in Healthcare

Subjective Estimates of Probability

In healthcare, it becomes hard to determine an objective probability for specific outcomes. Probability must be estimated based on the experience of the decision maker and/or the evidence from the literature. For instance, a physician who has performed a number of surgeries of a specific type can generally estimate the probability of success under a given set of circumstances. This experience-based probability estimate can be bolstered by published reports from other surgeons. Rather than an objective estimate of probability, this approach provides the best subjective estimate of probability for various outcomes. However, subjective estimates are not precise and are subject to various biases.

Subjective estimates of probability are known to be influenced by various cognitive variables such as saliency and recency of experience. For instance, a person who has just been through a bad outcome is likely to see adverse outcomes as more likely than is a person who has never experienced a severe adverse outcome.

Expected Utility

Expected utility (EU) takes into account the fact that the value of a commodity is subjective and different from one person to another due to differences in circumstance, among other reasons. When choosing a best decision using EU, the value of the competing outcomes to the decision maker is first established. Using EU, it is possible to incorporate nonmonetary decisions. For instance, value can reflect time trade-offs so that the decision maker may choose a short-term gain in health state over a worse state of health for a year. Then the probability

of each can be multiplied by the value placed on an outcome to make a "rational" choice.

In healthcare, the preferences of the patient are considered in determining subjective utilities. Patients' preferences may involve the amount of pain involved in a given procedure as well as quality of life as affected by the procedure. These preferences are not stable but are dependent on numerous other factors that can change over time. For instance, in the terminal stages of cancer treatment, the utility of avoiding pain may trump any slight chance of extending life when cure is not likely. In the early stages, however, the chance of cure may have more utility than avoiding temporary pain. Circumstances of the patients and their experience also affect their subjective utilities. People who have never experienced disability generally will rate the disability as giving more disutility than people who live with a disability.

Physicians also have utilities for different decisions. A doctor may decide to treat a patient for whom the treatment has little objective utility if the patient or his or her family demands treatment. The physician may also weigh the utility of providing a treatment of dubious value against the disutility of being accused of doing nothing.

However, understanding how to make good subjective probability estimates and good estimates of utilities are not the most important aspect of SEU. The real value of SEU is in providing a way to structure decisions so that justifiable (i.e., defensible) decisions can be made. SEU is often used with decision trees for a careful analysis of sequences of decisions. Moreover, SEU can be of value in clinical decision making without constructing an elaborate decision analysis. When the clinician understands the importance of subjective probability estimates and utility estimates, it becomes easier to visualize the structure of the decision process.

How a decision is structured is the cornerstone of effective decision making. For instance, when deciding what antibiotic to order, one should first estimate the likelihood that the disease process is amenable to antibiotic therapy. If the disease is responsive to antibiotics, one must still determine which antibiotic is most appropriate. Assuming that the patient has a bacterial infection rather than a virus or some other disease process, it is useful to determine whether or not the disease organism is sensitive to the antibiotics available. Furthermore, patient-specific variables, such as allergy and the presence of a pregnancy, can influence choice of antibiotic. SEU allows one to identify the outcome possibilities and their utilities to the patient and/or the physician and to evaluate the order in which decisions should be made. In this way, thoughtful, defensible decisions can be made.

SEU allows decisions to be tailored to specific individuals, groups, and situations. It is particularly useful in situations when the value and the probability of each outcome cannot be determined for the specific situation with any accuracy. However, understanding the role of experience and evidence in providing the best guess for probability can help the physician provide a reasoned estimate for key decisions. More important, SEU helps the physician explain the process with an understanding of which decisions are contingent on other decisions so that the process can be communicated effectively.

Application to Medical Decision Making

SEU is an approach to decision making that helps the decision maker structure the problem so that contingent relationships between choices are recognized. SEU uses subjective estimates of both the value of options to stakeholders and the decision maker's subjective estimates of the probability of these options. As such, SEU is useful in healthcare for discussing healthcare options with individuals, families, and small groups. It does not prescribe an optimal economic outcome but rather helps clarify which decisions should be made, discuss options in terms of possible outcomes, and choose the best course of action for the patient, given the information available at that point in time.

James Shanteau and
Alleene M. Ferguson Pingenot

See also Decision Trees: Introduction; Expected Utility Theory; Informed Consent; Prospect Theory; Risk Communication

Further Readings

Edwards, W. (1961). Behavioral decision theory. *Annual Review of Psychology, 12,* 473–498.

Edwards, W., & Newman, J. R. (1982). *Multiattribute evaluation*. Beverly Hills, CA: Sage.

Gigerenzer, G., Swijtink, Z., Porter, T., Daston, L., Beatty, J., & Krüger, L. (1999). *The empire of chance: How probability changed science and everyday life*. Cambridge, UK: Cambridge University Press.

Gilovich, T., Griffin, D., & Kahneman, D. (2002). *Heuristics and biases: The psychology of intuitive judgment*. New York: Cambridge University Press.

Schwartz, S., & Griffin, T. (1986). *Medical thinking: The psychology of medical judgment and decision making*. New York: Springer.

Sox, H. C., Blatt, M. A., Higgins, M. C., & Marton, K. I. (2007). *Medical decision making*. Philadelphia: American College of Physicians Press.

Subjective Probability

Subjective probability is a measure of the degree of belief held for the truth of an answer to a question. It is used in the quantification of uncertainty due to lack of knowledge, also called *epistemic uncertainty*. The word *epistemic* stems from the Greek word for knowledge. It indicates that this uncertainty has its origin in the nature and limits of knowledge. In the case of epistemic uncertainty, there are several answers to a question that are considered as possibly true, while there is only one true answer. The true answer will either be deterministic or probabilistic, depending on the formulation of the question. Both may be subject to epistemic uncertainty. A probabilistic answer uses probability in its classical frequentistic interpretation to quantify uncertainty due to random (or stochastic) variability, also called *aleatory uncertainty*.

Decision making involves asking questions. The answers to most of the questions will be subject to epistemic uncertainty. Quantitatively expressing this uncertainty by subjective probability enables one to employ all concepts, methods, and tools from probability calculus in the quantification of the combined influence of the uncertainties on decision making.

The following paragraphs explain by example the difference between subjective probability and probability in its classical frequentistic interpretation. Then, the connection with medical decision making is briefly pointed out. Rules for calculation are followed by a short introduction into the specification of subjective probability values. The discussion ends with two practical examples and a summary of how subjective probability is used in the uncertainty and sensitivity analysis of results from decision models.

Explanation

Consider the following illustrative example: A die is under the dice box, and it is uncertain which side is up. The question "Which side is up?" has a deterministic answer, namely, the number of eyes on the upper side of the die. This number could be known—one would only have to lift the dice box. There is only one true number, and the uncertainty about the correct number is quantified by subjective probability. Does it make any difference whether the die was cast in the past and covered so that one just needs to lift the dice box or whether the question is "Which side will be up in the next cast?" In both cases, there is only one true but unknown number that answers the question, and subjective probability is used to quantify the uncertainty that prevails until the next cast has been executed. However, the question "Which side is up in any cast of the die?" does not specify the cast and therefore does not have one true number as an answer. Rather, the population of numbers 1, 2, . . ., 6 applies such that for any cast the number can be thought of as randomly chosen from this population. The question can therefore be only answered probabilistically. The probabilistic answer summarizes the random variability among casts in the form of a probability distribution assigning, for instance, probability 1/6 to each of the six possible numbers. Probability is used here in its classical frequentistic interpretation as the limit of relative frequencies and is simply called *probability*. For instance, the probability for the side with number 3 to be up in any cast is the limit approached by the number of times this side was up in n casts divided by n, for an increasing number n of casts.

Clearly, the frequentistic interpretation does not make sense in the case of the "next cast." Performing this next cast many times (in the sense of repeating it exactly) is not possible, and if it were, would always show the same number up. It is, however,

also clear that the subjectivistic interpretation of probability does not make sense in the case of "any cast." There is no single number that would answer the question.

If it is unknown whether the die is fair, then it is doubtful whether 1/6 for each side is correct. There is epistemic uncertainty about the distribution that answers the "any cast" question. This uncertainty about the true probabilistic answer (i.e., the true distribution) is quantified by subjective probability.

The following is a brief summary of the above discussion of differences between two interpretations of probability.

Subjectivistic Interpretation	*Classical Frequentistic Interpretation*
Probability is the degree of belief (held for the truth of an answer to a question)	Probability is the limit of relative frequencies (of a random event)
Expresses state of knowledge	Summarizes random (or stochastic) variability
Quantifies uncertainty due to lack of knowledge (epistemic uncertainty)	Quantifies uncertainty due to random variability (aleatory uncertainty)

Connection to Medical Decision Making

Medical decision making is concerned either with a population of patients or with a specific patient. A decision model dealing with a population uses probability distributions to summarize the variability of characteristics of the patients in the population. The distributions are usually imprecisely known. Their state of knowledge, or epistemic uncertainty, is quantified by subjective probability. Since the decision model is concerned with the variability among a population of patients, the quantification of the variability and the quantification of the state of knowledge, or epistemic uncertainty, of the variability must be carefully kept apart.

In a decision model dealing with a specific patient, all parameters and input values are patient-specific. They are therefore fixed yet imprecisely known values. Their uncertainty is epistemic, and subjective probability is used throughout the model for state-of-knowledge quantification.

Rules for Calculation

Subjective probabilities have to comply with the same rules as probabilities in their frequentistic interpretation. The main rules are as follows:

R1: Subjective probabilities cannot be negative, nor can they be larger than 1.

R2: If A_1 is a possibly true answer to Question A and B_1 is a possibly true answer to Question B, then the subjective probability for both to be true is $sp(A_1B_1) = sp(A_1)sp(B_1|A_1) = sp(B_1)sp(A_1|B_1)$, where the vertical stroke is to be read as "if." If the subjective probability for B_1 to be true is the same whether A_1 is true or not, then $sp(B_1|A_1) = sp(B_1)$ and the subjective probability for both to be true is $sp(A_1B_1) = sp(A_1)sp(B_1)$.

R3: If A_1 is a possibly true answer to Question A and B_1 is a possibly true answer to Question B, then the subjective probability for at least one of them to be true is $sp(A_1 + B_1) = sp(A_1) + sp(B_1) - sp(A_1B_1)$. If A_1 and B_1 cannot be true together, then the subjective probability for at least one of them to be true is $sp(A_1 + B_1) = sp(A_1) + sp(B_1)$. Particularly, if $sp(A_1)$ is the subjective probability for Answer A_1 to be true and sp(not A_1) for A_1 to be false, then $sp(A_1) + sp(\text{not } A_1) = 1$.

R4: If $A_1, A_2, \ldots, A_n$ are possibly true answers to Question A, not any two or more of them can be true together, since there is only one true answer. It follows that $sp(A_1 + A_2 + \ldots + A_n) = sp(A_1) + sp(A_2) + \ldots + sp(A_n)$, for any integer value n. Particularly, if these are the only answers that are possibly true, then $sp(A_1 + A_2 + \ldots + A_n) = 1$.

Specification of Subjective Probability Values

For many questions, there are only two possibly true answers, for instance, "yes" and "no." In the case of complete uncertainty, equal subjective probability, namely, .5 each, for "yes" and for "no," is an adequate expression of the state of knowledge. The smaller the degree of belief for "yes," the closer the specified subjective probability value to 0 and, correspondingly, the larger the

subjective probability for "no." The larger the degree of belief for "yes," the closer the subjective probability value to unity and the smaller the subjective probability for "no."

If the question asks for a parameter or input value, plausibility considerations will often suggest limits such that values beyond these limits cannot possibly be true. Correspondingly, a uniform subjective probability distribution over the parameter value range that lies between these limits is an adequate quantitative expression for the state of knowledge of the true value according to the maximum entropy principle. Information additional to the limits may justify not using a uniform distribution but using one that exhibits some characteristic behavior toward the end points of the range of possibly true parameter values. This may, for instance, be a piecewise uniform or a triangular or truncated normal or lognormal distribution, to name just a few. The choice will need to observe the maximum entropy principle; that is, the distribution chosen must be such that it introduces a minimum of information in addition to what is actually given.

The plausibility considerations will have to take into account every aspect of the question, all possibly true answers, all arguments for and against each answer, and so on. They will use all available relevant information from experience, literature studies, extrapolations from related fields, laboratory experiments, theoretical models, and the input from interviews of experts and patients. If the epistemic uncertainty concerned is judged to be the main contributor to the uncertainty of results of a decision model, and the question to be answered by the model application is of sufficient importance, then the interviews will need to be performed using a structured approach to expert judgment elicitation. This includes methods of how to elicit the opinion of several experts, particularly how to avoid bias from the use of heuristics, and how to aggregate the opinions.

If there is uncertainty about the parameter value of a probability distribution that summarizes variability, and there are random observations of the variable quantity, then the likelihood of the observations can be evaluated for each of the possibly true parameter values. The Bayesian method can then be used to employ this likelihood in an update of the a priori state of knowledge of the true parameter value. The a priori state of knowledge may be based on plausibility considerations only, and the resulting subjective probability distribution is an a posteriori (after the observations) state of knowledge quantification for the true parameter value. A subjective probability interval, read from such a distribution, closely resembles a corresponding classical statistical confidence interval, provided that the Bayesian method starts with a minimum of a priori information (noninformative a priori distribution). Plausibility considerations and results from laboratory studies may often justify starting with an informative a priori distribution. In this case, the close resemblance mentioned above may no longer be given due to the additional information content, unless the number of random observations used in the update is large and the a priori information is not too restrictive.

Examples

Example 1

A patient has to undergo an operation. There is uncertainty about the outcome. The patient asks, "Will the operation be a success?" The physician can only give the patient his subjective probability for "success." A medical aid, on the other hand, may specify the probability (in its frequentistic interpretation) of a successful operation, without specifying the particular patient. If one assigns the value 1 to success and 0 to failure, the average value over many operations of this kind is an estimate of this probability. For example, the probability estimate derived from the recorded operations of a large number of patients may be .65 for "success." The physician may use .65 as his starting point for the specification of $\mathrm{sp}(O_{+Z})$, where O_{+Z} stands for "operation will be successful for patient Z." He finds that his patient shows the characteristic C, which reduces the subjective probability for a successful operation. The physician therefore needs to specify a factor that adjusts the value .65 accordingly. If he could obtain, from the documented operations, the relative frequencies of individuals with characteristic C among all recorded operations and among all recorded successful operations, then he could use them as estimates of $p(C)$ and $p(C|O_+)$, respectively, and use $p(C|O_+)/p(C)$ as the adjustment factor (since $p(O_+|C) = p(O_+)\,p(C|O_+)/p(C)$ according to Rule R2). An estimate

of any of these probabilities that is not based on observations of recorded operations is called a *subjective estimate*. This does, however, not make it a subjective probability. If estimates are not available from the recorded operations, they will need to be specified by the physician on the basis of his or her experience, literature studies, interviews of experts and patients. There will also be other influencing factors, such as relevant differences among patients showing characteristic C, so that additional judgment is needed to finally arrive at the subjective probability $\mathrm{sp}(O_{+Z})$ for the truth of the answer "The operation of patient Z will be successful." Both the patient and the physician may prefer to use the colloquial term *chance* in place of *subjective probability*. The term *chance* does not, however, properly convey the message that the quoted value is the physician's degree of belief held for the truth of the answer "Yes, the operation will be a success."

Example 2

As mentioned above, a priori subjective probabilities derived from plausibility considerations can be updated by random observations using Bayes's theorem. This theorem is a consequence of Rule R2. For instance, patient W shows symptom X, and the physician intends to perform test Y to update the a priori subjective probability $\mathrm{sp}(D_W) = .15$, where D_W stands for "Individual W has disease D." This a priori value may be taken from statistics, saying that among a large number of individuals with symptom X, 15% had disease D. Test Y is known to give a positive result (Y_+) in 90% of the cases with disease, and it is known to give a negative result (Y_-) in 90% of the cases without disease. The test, performed on individual W, gave a positive result (this is the observation). The probability $p(D|Y_+) = p(Y_+|D)p(D)/p(Y_+)$, with $p(Y_+) = p(Y_+|D)p(D) + p(Y_+|\text{not } D)p(\text{not } D)$, is then used as the updated or a posteriori subjective probability $\mathrm{sp}(D_W|Y_+)$. The result of this calculation is $\mathrm{sp}(D_W|Y_+) = .9 \times .15/(.9 \times .15 + .1 \times .85) = .61$. This result may, however, still need some adjustment. The percentages for a positive and a negative test result, derived from many tests performed on a population of patients, may need to account for the relevant characteristics of patient W. For example, $p(Y_+|DX)$ may not be equal to $p(Y_+|D)$; that is, the test may not give a positive result in 90% of the cases with disease D and symptom X, and so on.

Use in Uncertainty and Sensitivity Analysis

A decision model concerned with a population of patients uses imprecisely known probabilities and probability distributions to summarize variability. All probabilities and distributions estimated from statistics are subject to errors of various sources. The limited sample size, leading to wide confidence intervals around estimates of probabilities and estimates of parameter values of probability distributions, is only one (albeit often the most important) source. Consequently, there is epistemic uncertainty about the true value of a probability and the true distribution that summarizes the variability concerned. This uncertainty is quantitatively expressed by subjective probability. The decision model also uses other imprecisely known model parameters and input values such as utilities. Their state of knowledge (or epistemic uncertainty) can also be expressed by subjective probability distributions. Performing a Monte Carlo simulation with the decision model, where each simulation run is a complete evaluation of the decision model for a possibly true set of all ingredients that are subject to epistemic uncertainty, chosen at random according to their subjective probability distributions, provides a quantitative expression of the combined influence of the epistemic uncertainties on the model results. This is called an uncertainty analysis of the model application. If the resulting uncertainty is large (wide subjective probability distribution for model results), the physician will have to know where he or she needs to improve his or her state of knowledge in order to reduce the epistemic uncertainty of model results most effectively. This information is available from a sensitivity analysis that can be performed with the output from the Monte Carlo simulation. The wide uncertainty is less critical if most of it (as measured by subjective probability) lies below (or above) a given threshold value for a decision or if the physician finds from the sensitivity analysis that there is a good chance to improve the state of knowledge for main contributors to uncertainty.

Eduard Hofer

See also Bayes's Theorem; Expert Opinion; Probability; Uncertainty in Medical Decisions

Further Readings

Box, G. E. P., & Tiao, G. (1973). *Bayesian inference in statistical analysis*. Reading, MA: Addison Wesley.

de Finetti, B. (1974). *Theory of probability*. New York: Wiley.

Jaynes, E. T. (2003). *Probability theory: The logic of science*. Cambridge, UK: Cambridge University Press.

Vick, S. G. (2002). *Degrees of belief: Subjective probability and engineering judgment*. Reston, VA: American Society of Civil Engineers.

Winkler, R. L., & Hays, W. L. (1975). *Statistics: Probability, inference, and decision* (2nd ed.). New York: Holt, Rinehart & Winston.

Subset Analysis: Insights and Pitfalls

Subset analysis, also called subgroup analysis, is the statistical analysis of the effect of treatment intervention within subsets of the subjects in clinical trials. The subsets are usually defined by baseline characteristics of the subject population such as age groups, gender, or pretreatment comorbidity. As a supplement to the primary analysis of the clinical trial, which compares the randomized groups as a whole and ignores the potential heterogeneity within each subset, the subset analysis is widely used to explore whether the treatment effect is consistent across various subsets and, if different, which patient subsets might benefit more from the treatment under study. While the subset analysis provides valuable insights to the heterogeneity of subject population, it can be easily mismanaged or misinterpreted. The following sections discuss methodological issues with the subset analysis and point out the pitfalls and ways to avoid them.

An Example of Subset Analysis

In a clinical trial of the effect of reduced blood transfusion on postoperative morbidities after cardiac surgery, the target population is the patients undergoing coronary artery bypass grafting or heart valve surgery. A representative sample of 1,500 eligible patients is recruited from a clinical center. Since the number of transfused red blood cell units during an operation can not be predetermined, these patients are randomized to two transfusion triggers: (1) a liberal transfusion trigger that requests a unit of blood being transfused whenever the patient's hematocrit level (%) drops below 28 during the operation and (2) a conservative transfusion trigger that requests a unit of blood whenever the hematocrit level (%) drops below 24. Under this design, the conservative transfusion group would receive less transfused blood than the liberal group. The primary analysis is to compare, between the two treatment groups, the mean rate of the primary end point, a postoperative composite morbid outcome. In this way investigators are able to draw a conclusion on whether the conservative transfusion strategy is more beneficial than the liberal transfusion strategy for the target population.

In the primary analysis, the two randomized groups of patients are treated as a whole, and the within-group heterogeneity is ignored. However, the investigators speculate that the treatments may have different effects on patients with different body sizes (body mass index <18.5, 18.5–24.9, 25–30, >30), as patients of smaller body size are at higher risk of anemia if insufficient blood is transfused. The primary analysis cannot tell investigators whether the treatment effect is the same across different subsets and, if different, which patient subsets might benefit more from the conservative transfusion strategy. Subset analysis is the statistical analysis exploring such heterogeneity. People sometimes use the term *quantitative heterogeneity* for the case where one treatment is always better than the other across all levels of the subset variable (e.g., body mass index levels) but the magnitude of benefit varies; the term *qualitative heterogeneity* is often used for the case where one treatment is better than the other in some levels of the subset variable but worse in other subsets. Qualitative heterogeneity is rare in clinical studies.

In a subset analysis, one extracts data from the subset of patients and estimates and tests the treatment effect with it. Sometimes, it can yield intriguing insights that help in generating new hypotheses for further investigation. Hence, the subset analysis is an important consideration in the design and

analysis of both clinical trials and observational studies.

The subsets are usually defined by baseline characteristics that are known or speculated to influence the effect of the treatment. Put in statistical terms, the subset variables must be the ones that can potentially form interactions with the treatment groups.

Pitfalls in the Conduct of Subset Analysis

One problem with the subset analysis is the inadequate sample size. Clinical trials are usually designed with a sample size that is just big enough for the primary analysis to have enough power. In subset analysis, only a portion of the original data set is used. The confidence intervals for the treatment effect are expected to be wider, and the statistical tests have less power. If the sample size of the subset is half of the original, the corresponding confidence interval could be 41% wider; if the sample size is a quarter of the original, the confidence interval could double its width. The lack of power may lead to excessive false-negative findings (i.e., inflated Type II error) on the tests of treatment effect within subsets. For example, if the overall treatment effect is significant but one of the subset tests is not significant, it does not suggest lack of treatment effect in that subset.

Another problem is the multiplicity and inflation of Type I error when many subset analyses are conducted simultaneously. A common mistake is to do a statistical test for the treatment effect within each subset for many subsets, and pick out the ones with significant p values, and claim treatment effects on those subsets and no treatment effects on others. For the transfusion trial, four subsets are defined based on body size. Other subsets of interest are those defined by gender and by blood types (A, B, AB, O). Since the Type I error of the statistical tests is usually .05, even if there is no treatment effect in any of the subsets, as long as enough subsets are conducted, one can always find some (on average 5% chance) significant ones. If the family of subset tests is considered as a whole, the family-wise Type I error rate, that is, the probability of finding one false-positive result where none exists, is much higher than .05 (if there are 10 independent tests, this probability is estimated to be 40%). Therefore, any positive treatment effects from repeated subset tests should be viewed with caution. Multiplicity adjustments to the p values are available but do not always work well: The Bonferroni method is too conservative, especially in the presence of many tests; the Sidak method does not guarantee controlling the error rate under the prespecified level when the tests are not independent (which is almost always the case because many subsets are overlapped); resampling-based methods, such as bootstrap and permutation tests, often require the null hypothesis being "None of the subset variables have any effect," which does not always hold.

Subset analysis can be used in observational studies. For those that analyze large databases or registries, the sample size may be less of an issue because such studies often have more than enough sample size for the primary analysis and certain subset analyses. But the multiplicity problem still remains.

How to Avoid Them

In light of the two problems above, the subset analysis is usually not used as the primary or confirmatory analysis in clinical research. For a statistical analysis to be confirmative, the Type I and Type II errors must be strictly controlled. In most cases, the subset analysis is part of the secondary or exploratory analysis, for which strict error control is not required. The results can be used to corroborate the primary findings or to suggest new hypotheses for further research. Although it is often difficult to dispel the uncertainty associated with subset analysis, there are ways to reduce the false-positive or false-negative findings. The following are a few practical suggestions.

Use Interaction Tests Instead of Repeated Subset Tests

As mentioned before, a common mistake is to claim treatment effect on tests within individual subsets. Take the transfusion trial as an example. Suppose, after analyzing a number of subsets, the investigator finds that the conservative strategy significantly reduces postoperative morbidity among patients with bigger body size (body mass index >30) but not in the other three groups; it does not lead to the conclusion that the treatment

effect differs by body size. When many subsets are analyzed, a significant result may be a false-positive finding due to multiplicity. On the other hand, if none of the three subset tests are significant, it does not mean that there is no treatment effect. It may be the case that none of the tests have enough power to claim significance. A better way to assess the treatment effect heterogeneity is to include an interaction term between the treatment group and the subset variable in the model and test for the significance of that interaction. The interaction test produces a single *p* value based on all the data, no matter how many levels the subset variable has. In this way, the multiplicity is effectively reduced. Usually, the sample size of the trial is determined by the overall main effect of the treatment, and the interaction test may have much less statistical power. A clinical trial should be designed such that the conclusion is applicable to all the eligible patient population. If the treatment effect is very different in some subpopulations, they should have been left out and studied in a separate study. The lack of power of the interaction test ensures that only strong evidence can overthrow this assumption of homogeneity. In most cases, the interaction tests are used to check whether everything is in line with the homogeneity assumption.

Limit the Number of Prespecified Subset Analyses

Even when the interaction test is adopted, one may still do a lot of "data snooping" using various subset variables and only report the significant ones. Clearly, this kind of practice leads to excessive false positives. One way to reduce (but not eliminate) the false-positive rate (family-wise Type I error rate) is to prespecify the list of subset variables at the design phase of the clinical study. This is the same rationale as choosing the primary end point for a clinical trial: It has to be chosen prior to data collection. If the investigator has the freedom to choose the primary end point after seeing the data, he or she may be inclined to choose the one in favor of the research hypothesis and hence bias the conclusion. Subset analyses that are not prespecified are sometimes called post hoc subset analyses. Such analyses are also valuable to researchers: Unexpected phenomena may be observed during the conduct of the trial and prompt new subsets to be analyzed; the investigator may want to check for the consistency of treatment effect in various subsets. However, one should always keep in mind the limitations of unplanned subset analyses and interpret and report the results with caution, especially when unexpected results are found. Many scientific and reporting guidelines recommend that all the prespecified and post hoc analyses (including all that have been looked at by the data analyst) should be listed in the report of the study. It is also a good practice to limit the number of prespecified subsets to those that are justified by strong biological rationale. If Types I and II error rates are not strictly controlled, one should not emphasize the finding from subset analyses in the conclusion and should rather report the magnitude of the multiplicity problem informally by calculating how many statistically significant tests are expected by chance alone (this is usually the number of subset tests times .05, the level of significance).

Use the Forest Plot to Summarize the Result From Subset Analysis

The forest plot is a graphical display that shows the treatment effect heterogeneity across various subsets. Figure 1 is an exemplary forest plot using the published results from the MATCH trial (Diener et al., 2004). To make a forest plot, one first calculates the confidence intervals of the treatment effect (in this example, the odds ratio of the primary end point between the two treatment groups) for each subset and the entire data set. The confidence intervals are represented in the forest plot by parallel horizontal lines. The point estimator of the treatment effect is represented by the solid square on the confidence interval, with the size of the symbol proportional to the sample size of that subset. It is clear that the confidence intervals are shorter with bigger subset sample sizes. There are two vertical reference lines. One corresponds to the overall treatment effect. By comparing the subset confidence intervals with this line, one can assess the heterogeneity of subsets visually. The other reference line corresponds to null treatment effect. Note that the forest plot is a tool for descriptive analysis only. The sample size and multiplicity problems with subset tests apply to the subset confidence intervals as well. Therefore, one cannot claim treatment heterogeneity even if some subset confidence intervals do not cross the overall

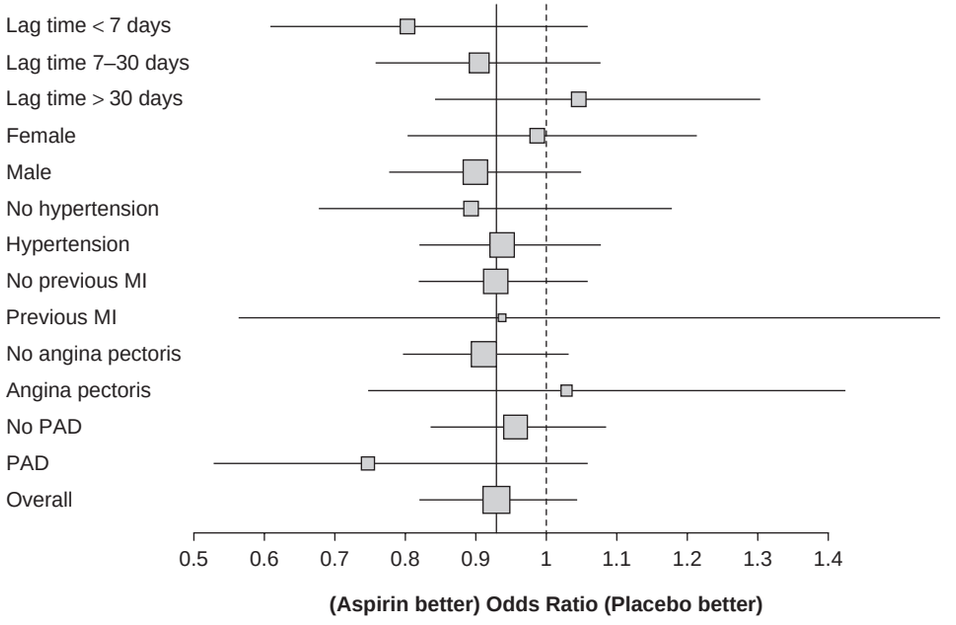


Figure 1 An exemplary forest plot for subset analysis

Notes: Lag time = time from qualifying event for enrollment to randomization. MI = myocardial infarction. PAD = peripheral arterial disease.

effect reference line. It has to come from the test of interaction. Jack Cuzick recommended de-emphasizing the null reference line because it encourages the misinterpretation that confidence intervals that cross this line indicate no treatment effect in those subsets.

If a post hoc subset analysis suggests some interesting treatment effect in a subset and the researcher wants to draw a confirmatory conclusion about it, he or she may conduct a separate meta-analysis that increases the sample size and credibility of the result by combining information from various (independent) sources. One can also design a new trial to specifically answer that question.

Liang Li

See also Analysis of Covariance (ANCOVA); Randomized Clinical Trials; Sample Size and Power

Further Readings

Assmann, S. F., Pocock, S. J., Enos, L. E., & Kasten, L. E. (2000). Subgroup analysis and other (mis)uses of baseline data in clinical trials. *Lancet, 355,* 1064–1069.

Begg, C., Cho, M., Eastwood, S., Horton, R., Moher, D., Olkin, I., et al. (1996). Improving the quality of reporting of randomized controlled trials: The CONSORT statement. *Journal of the American Medical Association, 276,* 637–639.

Cuzick, J. (2005). Forest plots and the interpretation of subgroups. *Lancet, 365,* 1308.

Diener, H. C., Bogousslavsky, J., Brass, L. M., Cimminiello, C., Csiba, L., Kaste, M., et al. (2004). Aspirin and clopidogrel compared with clopidogrel alone after recent ischaemic stroke or transient ischaemic attack in high-risk patients (MATCH): Randomized, double-blind, placebo-controlled trial. *Lancet, 364,* 331–337.

Follmann, D. (2004). Subgroups and interactions. In N. L. Geller (Ed.), *Advances in clinical trial biostatistics* (pp. 121–139). New York: Marcel Dekker.

Hernández, A. V., Boersma, E., Murray, G. D., Habbema, J. D. F., & Steyerberg, E. W. (2006). Subgroup analyses in therapeutic cardiovascular clinical trials: Are most of them misleading? *American Heart Journal, 151,* 257–264.

Wang, R., Lagakos, S. W., Ware, J. H., Hunter, D. J., & Drazen, J. M. (2007). Statistics in medicine: Reporting of subgroup analysis in clinical trials. *New England Journal of Medicine, 357,* 2189–2194.

Subtrees, Use in Constructing Decision Trees

A subtree is a portion of a decision model that is repeated in various places throughout the tree. Since decision models must reflect the real-world complexity of clinical medicine, compact notational representations that highlight repetitive structure throughout the decision model significantly improve comprehensibility. Much like a subroutine in common computer programming languages, the decision tree fragments represented by common subtrees are homologous structures that can be shared by many decision strategies and events.

Subtrees are a powerful notational representation and cognitive tool that can simplify and compact the decision tree representation graphically, emphasize analogies among events shared by multiple paths in the tree, highlight relations among factors in a decision model (e.g., through linkages and bindings), and ensure structural symmetry, helping the decision analyst to avoid inadvertent omissions.

Good Tree-Building Etiquette

A "Primer on Medical Decision Analysis" published in *Medical Decision Making* made a number of recommendations regarding tree structure. These included the following: (a) The tree must have symmetry, and (b) the branches must be "linked." The use of subtrees contributes to these goals.

Symmetry

A common error made in building decision tree models is to neglect to include the same chance events in all strategies of the decision tree. This may occur in several contexts. For instance, in a tree examining diagnostic testing, the modeler may neglect to consider a chance node representing the presence or absence of disease (i.e., disease prevalence) in the strategy that does not involve testing. In a tree examining different treatment options, the modeler may neglect to include a chance node representing treatment-related adverse events in the Do Not Treat strategy. While this may be reasonable in some clinical circumstances, there are many diseases in which the same adverse events may occur in both treated and untreated patients. For instance, patients with atrial fibrillation who receive anticoagulation or blood thinning treatment may have bleeding complications. However, even patients who do not receive this therapy may suffer from similar bleeding events, albeit at a lower risk. Similarly, patients receiving radiation therapy for prostate cancer may suffer from difficulty in urinating. However, even patients who do not receive radiation may have similar problems, albeit at a lower rate. The use of common subtrees to model these events ensures symmetry and reduces the risk of conceptual or programming errors. Of note, an alternative representational notation for decision models is the *influence diagram*, which by its very nature enforces complete syntactic symmetry. An automated decision-tree-critiquing program, BUNYAN, takes advantage of the principle of symmetry to diagnose common errors in decision tree models.

Consider the simple decision tree shown in Figure 1, where the underlying presence or absence of disease is explicitly modeled only on

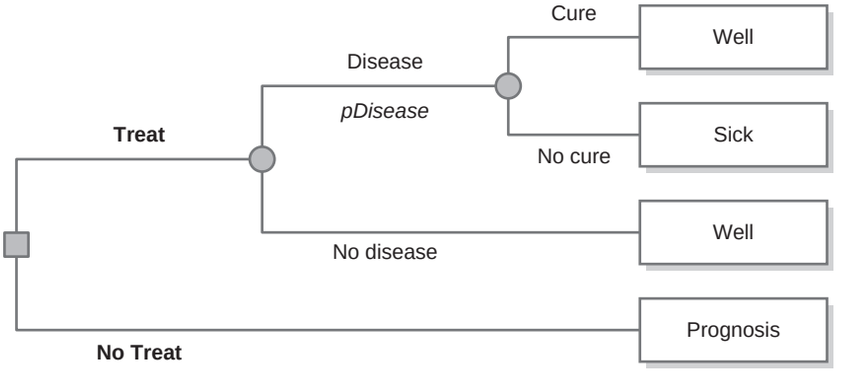


Figure 1 Treat/No Treat decision

the upper Treat branch. If one were to perform a sensitivity analysis on the parameter representing the probability of disease (*pDisease*), increasing *pDisease* would decrease the expected utility or value of the Treat strategy by shifting patients from well to sick. However, variations in *pDisease* would have no impact on the No Treat strategy because this factor is left implicit in the outcome prognosis. This would result in clinically nonsensical sensitivity analysis results—the more likely the patient was to have disease, the worse treatment would appear to be compared with no treatment; whereas in reality patients more likely to have disease are the ones who benefit most from treatment.

Figure 2 shows the same decision tree, now modified to use common subtrees. The curly bracket following the Treat and No Treat strategies signifies that the subtree connects to both strategies. However, while *pDisease* will correctly and explicitly be modeled for both strategies, the probability of cure (*pCure*) will likely be lower in the No Treat strategy. It is reasonable for many diseases that there be a nonzero probability of cure without treatment. Using the analogy of a computer program subroutine once more, the subtree needs to be "called" with the appropriate value for the parameter *pCure*. This is accomplished in some decision software packages, such as Decision Maker and SMLTREE, through a *binding* mechanism, in which the values of parameters, such as probabilities and utilities, can be set at some point proximal in the tree (e.g., to the left of the subtree) to alter the value of those parameters at all points in the tree distal to where the binding is set. Therefore, parameters in subtrees can take on different values depending on their context (i.e., which branches they are attached to). The use of subtrees and bindings also supports the tree-building principle of linkage, discussed below.

Linkage of Branches and Variables

Consider the classic Test, Treat, No Treat decision tree fragment shown in Figure 3. Note that the probability of being sick is described by four different variables. In the Treat strategy, the probability of being sick is *pSickRx*; in the No Treat strategy, it is *pSickNoRx*; and it the Test strategy, it is *pSickRxT+* or *pSickNoRxT–*, depending on a positive or negative test result. What happens if we want to perform a sensitivity analysis on the probability of getting sick? If we do this by examining *pSickNoRx*, we will find that, while the outcome (expected utility) for the No Treat strategy gets better as *pSickNoRx* gets smaller, outcomes for the other two strategies will remain the same and at some point patients not receiving treatment may actually do better than those receiving therapy. Unless the risk of side effects is very high, this is not likely. Furthermore, if the probability of getting sick without treatment decreases, it is also likely that the probability of getting sick with treatment will also decrease. Therefore, what we really want to accomplish by doing a sensitivity analysis on *pSickNoRx* is to examine some underlying probability of getting sick that influences or is linked to all the other probabilities of getting sick (*pSickRx, pSickRxT+*, *pSickNoRxT–*).

Figure 4 shows the same decision tree, now represented through extensive use of subtrees. Since the subtrees may be referenced by different contexts proximal in the tree, the binding mechanism

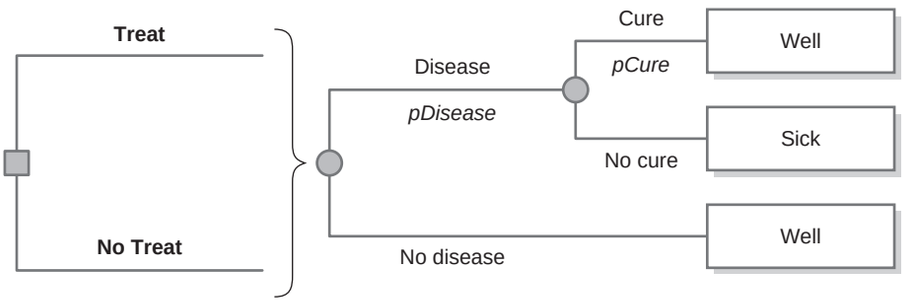


Figure 2 Treat/No Treat decision using a common subtree

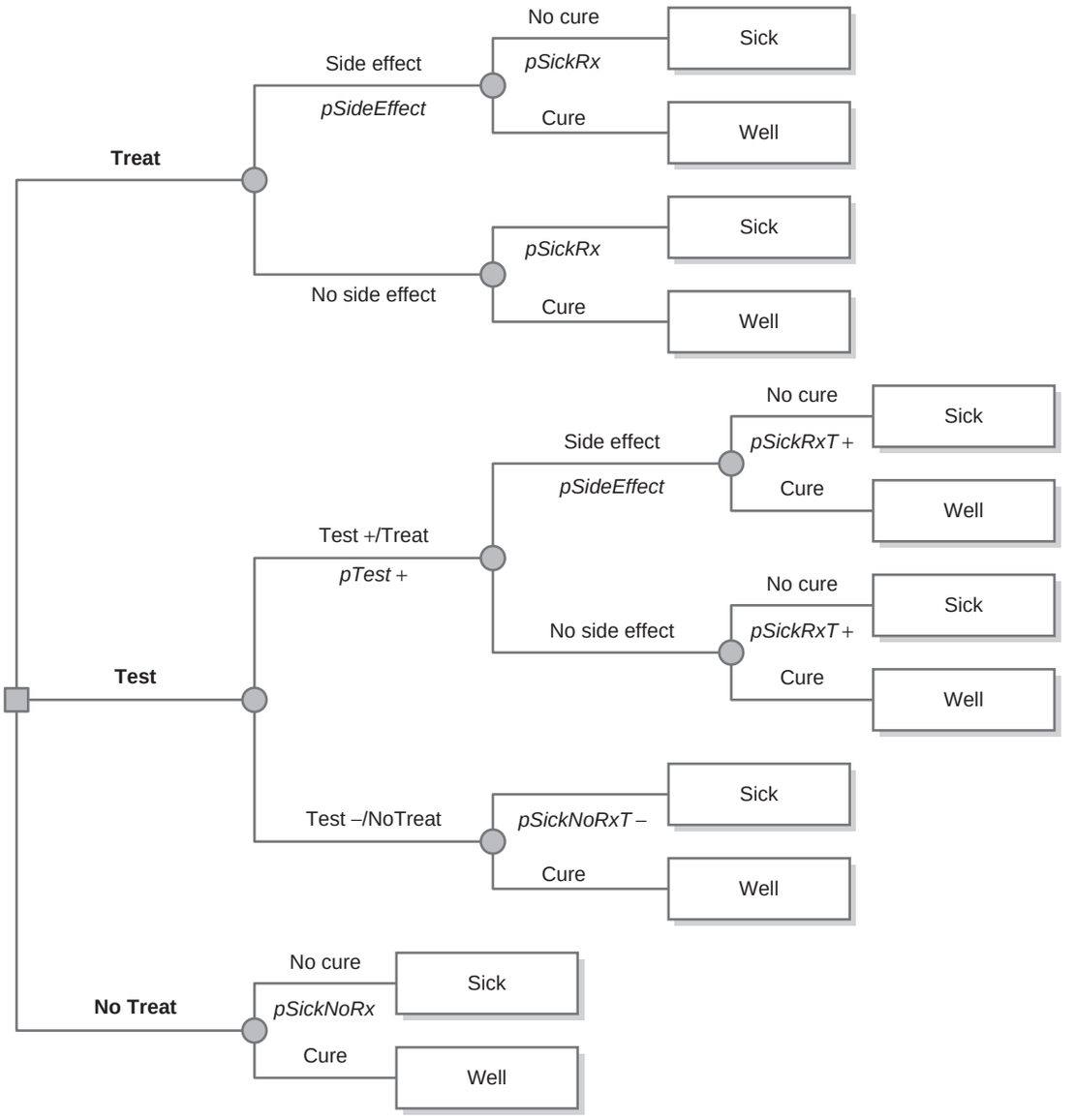


Figure 3 Treat/Test/No Treat decision

is used to pass parameter values to the variables appearing distally in the subtree. For instance, in the bottommost No Treat strategy, *pSick* is set equal to *pSickNoRx* (the underlying risk of getting sick without treatment). In the topmost Treat strategy, *pSick* is set equal to *pSickRx,* which itself is equal to *pSickNoRx*(1 – *effRx*). Using the binding mechanism to link all these probabilities to the underlying risk of getting sick (*pSickNoRx*) allows us to explicitly represent the efficacy of treatment, and if desired, we could perform a sensitivity analysis on this parameter to examine the impact of treatment efficacy without altering the underlying probability of getting sick. In a similar manner (not shown in this simple example), bindings and additional subtree notation can be used to model explicitly the presence or absence of disease causing the symptoms of being sick and appropriately adjust *pSickNoRx* to account for the absence or presence of disease, along with the occurrence of false-positive and false-negative test results.

Mark H. Eckman

See also Decision Trees, Construction; Decision Trees: Sensitivity Analysis, Deterministic; Influence Diagrams

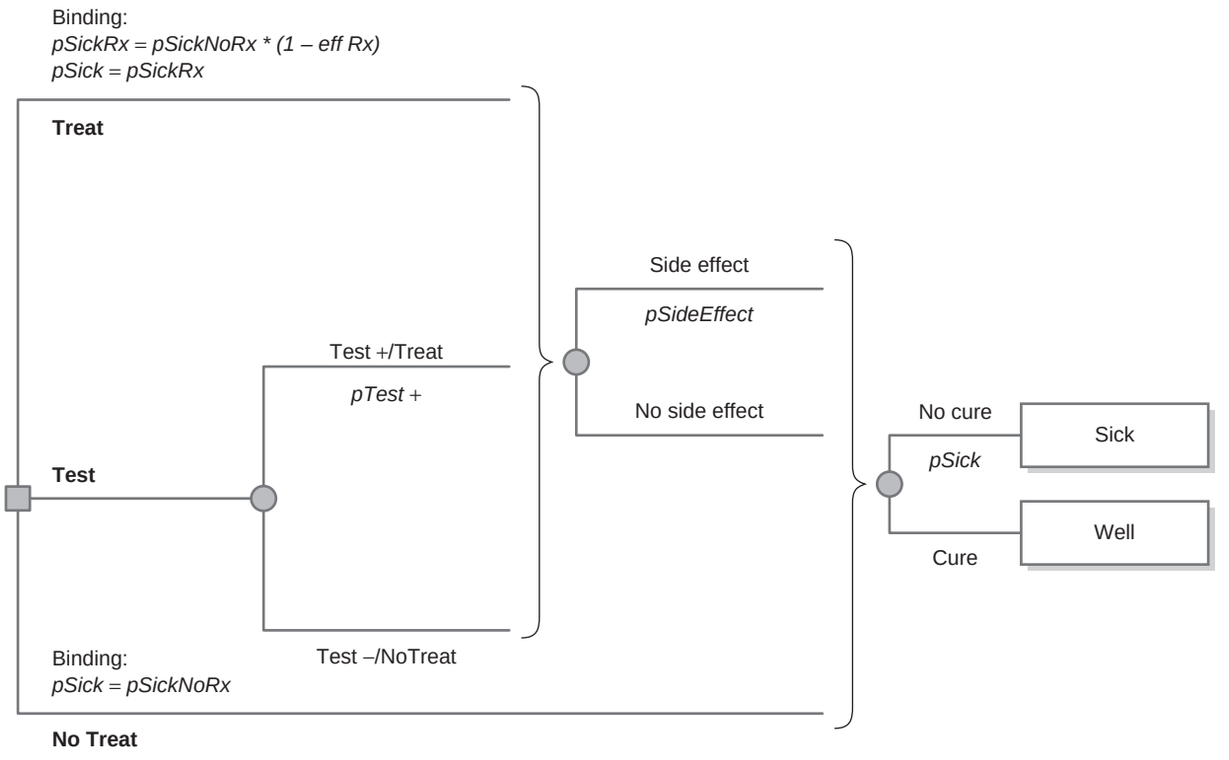


Figure 4 Treat/Test/No Treat decision using common subtrees

Further Readings

Detsky, A. S., Naglie, G., Krahn, M. D., Redelmeier, D. A., & Naimark, D. (1997). Primer on medical decision analysis: Part 2—Building a tree. *Medical Decision Making, 17*(2), 126–135.

Owens, D. K., Shachter, R. D., & Nease, R. F., Jr. (1997). Representation and analysis of medical decision problems with influence diagrams. *Medical Decision Making, 17*(3), 241–262.

Pauker, S. P., & Kassirer, J. P. (1992). Decision analysis. In J. C. Bailar III & F. Mosteller (Eds.), *Medical uses of statistics*. Boca Raton, FL: CRC Press.

Wellman, M. P., Eckman, M. H., Fleming, C., Marshall, S. L., Sonnenberg, F. A., & Pauker, S. G. (1989). Automated critiquing of medical decision trees. *Medical Decision Making, 9*(4), 272–284.

Sunk Costs

The sunk-cost or escalation effect leads a decision maker to continue a course of action into which an initial investment of time, thought, or expense has been "sunk," even after that particular course of action has proven to be a suboptimal choice. Like most cognitive heuristics, maintaining an unsuccessful course of action is often adaptive, as positive outcomes can take time to accrue, and there are always costs involved in switching. However, it can be counterproductive when the decision is based solely (or primarily) on the mere fact of having made a large prior investment instead of on an objective appraisal of current and future prospects.

Sunk Costs in Medical Decisions

Every day, medical practitioners are faced with making decisions where no clearly right or wrong answers exist. These decisions frequently involve weighing evidence in evaluating competing hypotheses, estimating probabilities, and predicting uncertain outcomes—for example, which test should be ordered to confirm or eliminate a particular diagnosis; should a patient continue with a given medication or switch to an alternative one; do the signs

and symptoms presented indicate Disease A, Disease B, or no disease at all? Research has shown that in making these sorts of decisions, physicians behave in a manner very similar to that of expert decision makers in other domains (e.g., physics, logic, chess); that is to say, rather than using time- and labor-intensive computational procedures, they rely on rules of thumb or heuristics. In doing so, they are able to apply a wealth of accumulated knowledge relatively quickly and efficiently. These heuristics are generally an effective way of allocating limited cognitive resources in dealing with uncertain situations, but they can also introduce reasoning biases that adversely affect a decision. The sunk-cost effect is one example of how heuristics can lead to biased decisions in medical decision making.

This effect has been studied extensively in the fields of economics and behavioral decision making, where numerous studies show that decision makers' commitment to a decision increases as a function of the amount of their initial investment, in money, effort, or time. A number of variables can affect individuals' susceptibility to the sunk-cost effect, including personality. Of particular relevance to medical decision making, individuals with a Type A personality—a disposition shared by many physicians—are more likely to demonstrate the effect. In medical decision making, the sunk-cost effect can have important diagnostic or therapeutic implications on an individual-patient level and perhaps even on a national level when looking at the system of delivering medical care.

Individual-Patient Level

An example of the sunk-cost effect at the patient level is one where the physician recommends that the patient continue with further courses of the same medication even after the patient fails to improve on that medication because of the time and money that the patient has invested or the time and energy that the physician has invested in coming up with the present, apparently inaccurate diagnosis or ineffective treatment plan. Research addressing the sunk-cost effect in physicians' reasoning has asked physicians (medical residents) to evaluate different responses to scenarios where an initial diagnosis or treatment is suspect, as a function of the amount of time and/or money that had already been invested (by varying, e.g., the cost of medication). The results showed that residents were surprisingly good at not allowing the sunk costs to affect their decisions in evaluating medical treatment scenarios, but they were no better than laypeople at preventing the sunk cost bias from affecting their decisions in everyday situations outside the medical realm. Although physicians were not susceptible to the sunk-cost bias in judging medical scenarios, they nonetheless considered it more important to continue the original treatment for purposes of consistency when they, as opposed to another physician, had made the original decisions.

These findings would seem to indicate two things: first, that commitment effects are less in the absence of prior involvement, which supports the practice of obtaining additional medical opinions, and second, that medical training and expertise may ameliorate some of the nonnormative effects of biases. Indeed, recent studies involving emergency room doctors have shown that those involved in medical education are less likely to exhibit various forms of cognitive bias than their colleagues who do not mentor students. Although medical decision makers are by no means immune to cognitive biases, these biases are not unavoidable, and they can be ameliorated somewhat through medical education.

National Level

On a more global level, the sunk-cost effect could play a role in a country's continued reliance on a particular kind of medical delivery system, even when that system provides fractured medical care, leaves a large number of patients under- or uninsured, and is financially inefficient. To be sure, questions of how best to optimize healthcare are enormously complex, but public policy debates on the issue—in the United States and elsewhere—contain elements of a sunk-cost rationale (i.e., "We have too much invested in our current healthcare system to start making changes now"). A similar resistance to changing failing policies has characterized other political and policy debates (e.g., the American government's reluctance to pull out of Vietnam). Thus, there is potential for the sunk-cost effect to operate at both the microlevel and the macrolevel of medical decision making.

Brian H. Bornstein and A. Christine Emler

See also Bias; Personality, Choices; Treatment Choices

Further Readings

Bornstein, B. H., & Emler, A. C. (2001). Rationality in medical decision making: A review of the literature on doctors' decision-making biases. *Journal of Evaluation in Clinical Practice, 7,* 97–107.

Bornstein, B. H., Emler, A. C., & Chapman, G. B. (1999). Rationality in medical treatment decisions: Is there a sunk-cost effect? *Social Science and Medicine, 49,* 215–222.

Hall, K. H. (2002). Reviewing intuitive decision-making and uncertainty: The implications for medical education. *Medical Education, 36,* 216–224.

Roswarski, T. E., & Murray, M. D. (2006). Supervision of students may protect academic physicians from cognitive bias: A study of decision making and multiple treatment alternatives in medicine. *Medical Decision Making, 26,* 154–161.

Support Theory

Support theory is a descriptive model of probability judgment that posits that judgments of probability are made based on descriptions of events rather than on events themselves.

Probability theory provides a normative framework for determining the probability of a combination of disjoint events. If two events A and B are exclusive, and have probabilities $p(A)$ and $p(B)$ of occurrence, the probability of one or the other occurring is exactly $p(A) + p(B) = p(A$ or $B)$. "A or B" is referred to as the disjunction of A and B. For example, if A is "The patient's heart rate is between 60 and 70 beats per minute" and B is "The patient's heart rate is between 70 and 80 beats per minute," the disjunction might be expressed as "The patient's heart rate is between 60 and 70 beats per minute or between 70 and 80 beats per minute" (an explicit disjunction) or as "The patient's heart rate is between 60 and 80 beats per minute" (an implicit disjunction).

Support theory attempts to explain the observation that people often judge the probability of implicit disjunctions to be lower than the sum of the probabilities of the constituent events. This property is referred to as *subadditivity* and is contrasted with the normative probability theory model's *additivity* property (and with *superadditivity*, in which the probability of a disjunction is judged to be higher than the sum of the probabilities of the constituent events).

In support theory, descriptions of competing events are evaluated by assessing the relative support (s) for each description, characterized by a nonnegative real number associated with the strength of evidence for that description. Formally,

$$p(A \text{ rather than } B) = \frac{s(A)}{s(A) + s(B)}.$$

Support for a description may be a function of the strength of memories for events matching the description. As the name suggests, it may also be related to the ability to provide reasoned justifications for the described event. A stochastic extension of support theory, random support theory, extends the basic support theory model by representing support for events as a random variable. That is, people are assumed not to assign a fixed level of support to a given description but to sample support at random from a distribution of support. As a result, it is possible to speak of the variance and expectation of support associated with a given description. This enables random support theory to model the calibration of judgments (whether the objective frequencies of events are correctly predicted by their subjective probabilities).

Support theory assumes that the support for an implicit disjunction is less than or equal to the support for an equivalent explicit disjunction. It also holds that the support for an explicit disjunction is less than or equal to the sum of the support for the two descriptions. Formally, if A describes an implicit disjunction of B and C,

$$s(A) \leq s(B \text{ or } C) = s(B) + s(C).$$

Two cognitive processes, *unpacking* and *repacking*, operate on descriptions of events and determine whether they are considered to be implicit or explicit disjunctions. Unpacking an implicit disjunction into its constituents increases the overall support, and thus the judged probability, of the event. A classic illustration is that judgments of "the likelihood of death by any vehicle accident" are often lower than judgments of "the likelihood of death by car accident, death by plane accident, death by bicycle

accident, or death by any other vehicle accident." Moreover, the latter is often lower than the sum of the judgments of "the likelihood of death by car accident," "the likelihood of death by plane accident," "the likelihood of death by bicycle accident," and "the likelihood of death by any other vehicle accident" when elicited individually.

Unpacking may call attention to alternatives that were not considered in the implicit disjunction or may highlight the salience of such alternatives. For example, house officers assessing a hypothetical case for the probabilities of gastroenteritis, ectopic pregnancy, and none of the above (G, E, N) assigned a lower probability to none of the above (N) than house officers judging probabilities of gastroenteritis, ectopic pregnancy, appendicitis, pyelonephritis, pelvic inflammatory disease (PID), and none of the above (G, E, A, PY, PI, N) assigned to "appendicitis, pyelonephritis, PID, and none of the above" (A, PY, PI, N). In this case, it is possible that unpacking none of the above to include, for example, PID may have reminded the residents about the possibility of PID.

Conversely, repacking an explicit disjunction into a corresponding implicit disjunction can occur when the events are very similarly described (as with the heart rate example). Once repacked in the mind, support theory predicts subadditivity for the repacked (implicit) disjunction and, thus, that what appears to be an explicit disjunction can have less support than the sum of the support of its constituent events. In addition, judges may anchor their probability estimate on one of the constituents of the explicit disjunction (e.g., "60–70" bpm) and then determine the probability of the disjunction by adjusting upward from the anchor for each of the other constituents. Insufficient adjustment (which is typical) also leads to subadditivity for explicit disjunctions.

Most research studies have provided evidence for support theory's key insight—that it is support for descriptions, rather than events, that are used in probability judgment. Some researchers, however, have reported cases in which unpacking leads to additive or even superadditive probability judgments. Subadditivity may be limited to conditions in which the events are unpacked into descriptions that do not easily come to mind and yet have higher support than those that do easily come to mind, as noted by Sloman and colleagues. It has also been suggested by Bearden, Wallsten, and Fox that there are other important contributors to subadditivity besides support theory, such as response variability.

Support theory suggests that clinical decision makers must take care when making decisions based on judged probabilities for sets of events that can be variously described, as, for example, a list of differential diagnoses for a patient. The medical research literature facilitates the determination of the likelihood of single diagnoses given clinical findings, but when the problem necessitates considering a disjunctive diagnosis (e.g., "bacterial or viral meningitis"), patients, clinicians, and research subjects may not combine constituent probabilities normatively. Studies of smoking suggest that judged risk of lung cancer is greater when it is presented alone than when it is included in a list of causes of death. By increasing the availability of other causes of death, it may be possible to reduce overestimates of risk of a particularly salient disease by patients.

Alan Schwartz

See also Bias; Differential Diagnsosis; Heuristics; Judgment; Probability

Further Readings

Bearden, J. N., Wallsten, T. S., & Fox, C. R. (2007). Contrasting stochastic and support theory accounts of subadditivity. *Journal of Mathematical Psychology, 51,* 229–241.

Brenner, L., Griffin, D., & Koehler, D. J. (2005). Modeling patterns of probability calibration with random support theory: Diagnosing case-based judgment. *Organizational Behavior and Human Decision Processes, 97,* 64–81.

Fox, C. R. (1999). Strength of evidence, judged probability, and choice under uncertainty. *Cognitive Psychology, 38*(1), 167–189.

Redelmeier, D. A., Koehler, D. J., Liberman, V., & Tversky, A. (1995). Probability judgment in medicine: Discounting unspecified probabilities. *Medical Decision Making, 15,* 227–230.

Rottenstreich, Y., & Tversky, A. (1997). Unpacking, repacking, and anchoring: Advances in support theory. *Psychological Review, 104*(2), 406–415.

Sloman, S., Rottenstreich, Y., Wisniewski, E., Hadjichristidis, C., & Fox, C. R. (2004). Typical

versus atypical unpacking and superadditive probability judgment. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 30*(3), 573–582.

Tversky, T., & Koehler, D. J. (1994). Support theory: A nonextensional representation of subjective probability. *Psychological Review, 101*(4), 547–567.

Support Vector Machines

Support vector machines (SVMs) are machine learning models that share some similarities with neural networks and logistic regression models for classification tasks. Examples of such tasks arise naturally in clinical settings, whenever one is given a set of data descriptors (e.g., lab results, clinical findings, imaging data, genetic information) and wants to predict health status or medical outcome given such data. In the simplest case, which is discussed in this entry, there are only two possible outcomes to predict. In a clinical context, these two can, for example, correspond to healthy and diseased patient states, respectively.

Traditionally, logistic regression and artificial neural network models have been the tools of choice for solving classification tasks as outlined above. In the past 10 years, SVMs have increasingly been used in data-intensive machine learning scenarios in clinical contexts, for example, as decision aids for the classification of mass spectronomy data or imaging data.

The following assumptions and notational conventions are used here: An n-element data set $D = \{x_i, t_i\}$ contains m-dimensional data points (cases) x_i that serve as inputs to the SVM that classifies these cases into the associated class labels t_i $\{1, +1\}$ (outcomes or outputs of the SVM). The data points x_i are mathematical representations of the clinically relevant information that is to be used in the classification task. In a similar vein, the class labels t_i are abstractions of the two classes that should be predicted by the model. Because they are m-dimensional, the data points will sometimes be referred to as vectors. The pattern classification task is to find a model (in this case, the SVM) and associated parameter settings that are able to predict class labels, given the data points, while making the fewest mistakes. This means that, when given a new data point x^* from the same distribution as D, the model output should be the correct class label of x^* as often as possible. For most data sets, it will not be possible to reduce the average error to 0. A model that makes few mistakes on unseen data is said to *generalize well*.

The following presentations are mathematical in nature because SVMs are based on geometrical concepts. Nevertheless, it is hoped that the essence of SVMs (the "what") can be grasped without having to understand all the mathematical details (the "how").

Optimal Separating Hyperplanes

A *hyperplane* is the extension of the concept of a straight line (in 2D) or a plane (in 3D) to $n > 3$ dimensions. A hyperplane, H, defined as the set of all points x that satisfy the equation $\boldsymbol{w} \cdot \boldsymbol{x} + b = 0$, partitions its enclosing space into three parts: (1) the points directly on the plane, (2) the points for which $\boldsymbol{w} \cdot \boldsymbol{x} + b > 0$, and (3) the points for which $\boldsymbol{w} \cdot \boldsymbol{x} + b < 0$. Here, $\boldsymbol{w} \cdot \boldsymbol{x}$ denotes the dot product of the two m-dimensional vectors $\boldsymbol{w}$ and $\boldsymbol{x}$, that is, the result of multiplying all components of $\boldsymbol{x}$ with the corresponding components of $\boldsymbol{w}$ and adding up the results. The two parameters $\boldsymbol{w}$ and b encode the position of the hyperplane in its enclosing space: $\boldsymbol{w}$ encodes the orientation and b the distance to the origin. Together, these two parameters uniquely determine where H lies. A hyperplane H can thus be used as a classification model: All the points on one side of H belong to one class; all the points on the other side belong to the other class. A data set that can be classified in this sense by a hyperplane is called *linearly separable*.

It was one of the fundamental results of early machine learning research that a simple iterative procedure (the *perceptron learning rule*) will always find a separating hyperplane for a linearly separable data set. This result guarantees the existence of a separating hyperplane. It is not clear, however, whether this hyperplane generalizes well. To investigate this question, researchers from the field of statistical learning theory have defined the notion of an *optimal separating hyperplane*: This hyperplane is as far as possible from the two classes it separates. This means that the margin between the two classes is as large as possible; SVMs are therefore also known as *large-margin classifiers*. It can be demonstrated

theoretically (and backed up by empirical results) that optimal separating hyperplanes generalize better than arbitrary hyperplanes.

Below, the computations required for determining this optimal separating hyperplane are outlined. The computations are nothing more than the mathematical apparatus for finding the best possible separating line (in 2D) or hyperplane (in higher dimensions). Remember that the two parameters $\boldsymbol{w}$ and b uniquely determine the position and orientation of the hyperplane. Determining the correct values of these parameters therefore constitutes solving the classification task. Unfortunately, the two parameters do not have an additional interpretation in any useful clinical sense; this is in contrast to logistic regression, where the β coefficients *do* have such an interpretation.

As an example, consider the problem of diagnosing pigmented skin lesions as malignant melanoma or common nevi based on imaging data acquired, for example, by dermoscopy. Several dozen data descriptors such as size, asymmetry, color distributions, and so on, constitute the m-dimensional data points x_i; the class labels t_i encode whether the lesion in question is benign or malignant. This gold standard diagnostic information would have to be obtained by histopathology or follow-up examinations (to rule out malignancy). The data set with established gold standard diagnoses is known as *training data*. Assuming that it were possible to always correctly diagnose lesions as benign or malignant, an SVM could accomplish this diagnosis by calculating the optimal separating hyperplane between the two sets of lesions in the training data. Diagnosing a novel lesion described by a data point x^*, for which the gold standard is not known, then amounts to determining on which side of the separating hyperplane the point x^* lies. The parameters $\boldsymbol{w}$ and b of the hyperplane do not convey any clinical information and cannot be used other than for encoding the position and orientation of the hyperplane.

The mathematical details for calculating the optimal separating hyperplane are as follows. For a point x_0 and a hyperplane H given by the equation $\boldsymbol{w} \cdot \boldsymbol{x} + b = 0$, the value of $\boldsymbol{w} \cdot \boldsymbol{x}_0 + b$ is proportional to the distance of x_0 from H (positive if x_0 is on one side, negative if it is on the other). Note that a given hyperplane can be expressed by an infinite number of equations: If H is given by $\boldsymbol{w} \cdot \boldsymbol{x} + b = 0$, then it is also given by $\alpha(\boldsymbol{w} \cdot \boldsymbol{x} + b) = 0$, for any $\alpha \neq 0$. We can thus make the expression $\boldsymbol{w} \cdot \boldsymbol{x}_0 + b$ arbitrarily large or small by changing the parametrization of H. This also means that there exists a parametrization such that $\boldsymbol{w} \cdot \boldsymbol{x}_i + b = -1$ for the points x_i closest to H on one side and $\boldsymbol{w} \cdot \boldsymbol{x}_i + b = +1$ for the points x_i closest to H on the other side (these points are said to *lie on the margin*). We now want to find a hyperplane such that the distance between these closest points is as large as possible, that is, the margin is largest. By observing that the distance of x_0 to H is given by $|\boldsymbol{w} \cdot \boldsymbol{x}_0 + b|/||\boldsymbol{w}||$, it follows that the distance between the closest points on both sides is $2/||\boldsymbol{w}||$. Here, single vertical bars denote the absolute value of a number, and double vertical bars the length of a vector. Maximizing $2/||\boldsymbol{w}||$ thus yields the value of $\boldsymbol{w}$ that corresponds to the optimal separating hyperplane.

Maximizing $2/||\boldsymbol{w}||$ is equivalent to minimizing $||\boldsymbol{w}||$, or $1/2||\boldsymbol{w}||^2$ (the latter form is chosen for its mathematical convenience). (It is immediately obvious that $\boldsymbol{w} = 0$ minimizes the value $1/2||\boldsymbol{w}||^2$, but this does not represent a sensible solution to the original problem—it would violate the assumption that the closest points satisfy $\boldsymbol{w} \cdot \boldsymbol{x}_i + b = -1$ and $\boldsymbol{w} \cdot \boldsymbol{x}_i + b = +1$, respectively.) All other points have to be further than this from the hyperplane. The problem of determining the optimal separating hyperplane is thus a *constrained optimization problem* of the form (known as *primal format*):

$$\text{Minimize } 1/2||w||^2$$
$$\text{subject to } t_i(w \cdot x_i + b) \geq 1, \text{for } 1 \leq i \leq n,$$

where the two forms of constraints $\boldsymbol{w} \cdot \boldsymbol{x}_i + b \leq -1$ and $\boldsymbol{w} \cdot \boldsymbol{x}_i + b \geq +1$ are written in one format so that they can both be treated the same way. This format can be achieved by requiring all points for which $\boldsymbol{w} \cdot \boldsymbol{x}_i + b \leq -1$ holds to have class label $t_i = -1$ and all points with $\boldsymbol{w} \cdot \boldsymbol{x}_i + b \geq +1$ to have class label $t_i = +1$.

The above problem is a convex quadratic optimization problem that can be solved by standard mathematical methods. In contrast to the error optimization algorithms used in neural network training, this constrained optimization problem has a unique global solution. The standard way of calculating the solution is to use the method of *Lagrangian multipliers* α_i to convert the problem

to its so-called dual format, in which it can be stated as

$$\text{Maximize } \sum \alpha_i - \frac{1}{2}\sum\sum \alpha_i \alpha_j t_i t_j x_i \cdot x_j$$
$$\text{subject to } \alpha_i \geq 0, \text{for all } i \text{ and } \sum \alpha_i t_i = 0.$$

The solution to the problem in this format is exactly the same as the solution to the original problem. Although this dual format seems to be more complicated than the primal format, it has the major advantage that the data points enter the calculation only via their dot products $x_i \cdot x_j$. As we will see later on, it is this particular form of dependence that allows the generalization of linear hyperplanes to nonlinear decision boundaries.

The quadratic optimization problem, whether in primal or dual format, has a unique solution $w = \sum \alpha_j t_i x_i$—that is, as a weighted sum of the inputs. This solution is sparse: In general, few of the α_j are nonzero. The corresponding x_i are exactly those data points that lie on the margin; these are known as *support vectors*. A graphical representation is shown in Figure 1. A hypothetical clinical situation that may be represented by Figure 1 is the distinction between high-risk and low-risk patients in the prediction of cardiovascular disease, given two risk indicators (e.g., systolic blood pressure and total cholesterol). Patients on one side of the hyperplane are deemed to be at high risk, while those on the other side are deemed to be at low risk. It is, however, rare that two health states can be distinguished as easily (by a straight line) as shown in Figure 1. The extensions to basic SVM methodology discussed below explain how to deal with more complicated situations.

Note that the solution is determined completely by the position of the support vectors; all other data points could be removed without changing the solution. Because the support vectors satisfy $t_i(w \cdot x_i + b) = 1$, these vectors can be used to calculate the value of the parameter b.

Based on the parameters $w = \sum \alpha_j t_i x_i$ and b, the decision function for an SVM is

$$\text{sign}\left(\sum \alpha_j t_i x_i \cdot x + b\right).$$

This means that all the points x for which $\sum \alpha_j t_i x_i \cdot x + b > 0$ belong to one class, and all the points with $\sum \alpha_j t_i x_i \cdot x + b < 0$ belong to the other class.

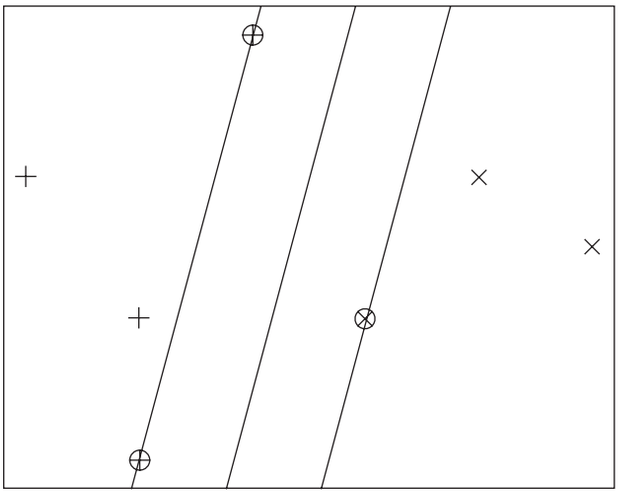


Figure 1 The optimal separating hyperplane for distinguishing between two classes of data points

Note: The solution depends only on the support vectors (shown as circles).

Note the implication of considering only the support vectors and ignoring all other points in determining the SVM parameters: If incorrectly labeled cases are support vectors, the hyperplane will be incorrectly determined. Other classifiers such as linear discriminant analysis, artificial neural networks, classification and regression trees, and logistic regression do not discard information from the majority of cases as SVMs do and are therefore not as easily influenced by the labels and positions of the data points on the margins.

Soft Margin Support Vector Machines

Few data sets that contain biomedical information are linearly separable. In many situations, the above derivations are therefore not directly applicable. To overcome this limitation, Cortes and Vapnik introduced the notion of *soft margin SVMs*. In this extension to the standard SVM methodology, data points are also allowed to lie on the wrong side of the margin. Of course, the total distance of these points from the separating hyperplane, measured by nonnegative *slack variables* ξ_i, should be as small as

possible. In the primal format, the constraints that allow slack variables are $t_i(\boldsymbol{w} \cdot \boldsymbol{x}_i + b) \geq 1 - \xi_i$. The optimization goal is then twofold: Minimize the original optimization goal, and minimize the total sum of slack variables. The trade-off between these two goals is given by a parameter C that weights the contribution of the slack variables. The primal format of the soft margin SVM problem is

$$\text{Minimize } 1/2||w||^2 + C\sum \xi_i$$
$$\text{subject to } t_i(w \cdot x_i + b) \geq 1 \text{ and } \xi_i \geq 0, \text{ for } 1 < i < n.$$

In the dual format, the slack variables vanish, and the only difference to the dual format of the original problem is that C is an upper limit to the weighting terms α_i of the support vectors; that is, there are now n constraints $0 \leq \alpha_i \leq C$. As in standard SVMs, the decision boundary is again a hyperplane for soft margin SVMs. Cases that are not support vectors therefore still contribute no additional information to that gathered from the support vectors and incorrectly classified cases.

Nonlinear Support Vector Machines and the Kernel Trick

Nonlinear decision boundaries can be represented by SVMs by using the following observation: If a data set is not linearly separable, and one can project it onto a higher-dimensional *feature space* F with a nonlinear function Φ (followed by the optimal separating hyperplane construction presented above), then the linear hyperplane in F is a nonlinear decision boundary in the original data space. It was the achievement of Boser and colleagues to apply a shortcut to these calculations: They noted that the kernel function method of Aizerman and colleagues can be applied to optimal separating hyperplanes. This method consists of calculating a so-called kernel function $k(x_i, x_j)$ of two data points instead of the dot product $\Phi(x_i) \cdot \Phi(x_j)$ of the data points after projecting onto F via Φ. One can show that several large classes of functions satisfy the requirements of being such kernel functions. Using these functions, one need not specify a projection Φ or a feature space F—all calculations are performed in the original data space.

The most widely used classes of kernel functions are polynomials (with the degree d of the polynomial as parameter of the kernel) and Gaussian radial basis functions (with the variance σ^2 as parameter). These two classes of kernels are defined as

$$k(x_i, x_j) = (\boldsymbol{x}_i \cdot \boldsymbol{x}_j)^d$$

and

$$k(x_i, x_j) = exp(-1/2||x_i - x_j||^2/\sigma^2).$$

An example of the nonlinear nature of decision boundaries in the original space is given in Figure 2. This situation could, for example, correspond to the diagnosis of breast cancer from two imaging parameters. The distinction between healthy and diseased cases can only be achieved by a curved line.

Other Topics of Interest

The standard SVM algorithm can be modified to allow regression problems (predicting a real value, instead of a class label) to be solved by SVMs. For this, one defines an ε-insensitive loss function that penalizes predictions only when they are more than a value of ε from the correct value. The corresponding optimization problem can be handled in a manner similar to the hyperplane calculations

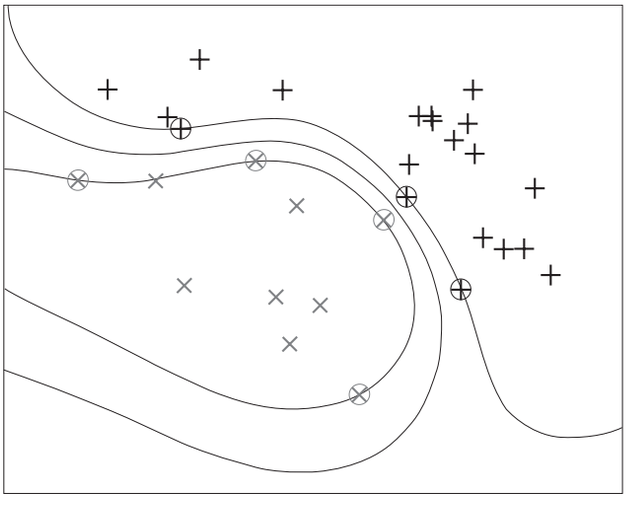


Figure 2 Decision boundary for a nonlinear SVM using a Gaussian radial basis function kernel and parameter $C = 100$

Note: The points on the margin are marked as support vectors.

outlined above. In particular, nonlinearities are again introduced by use of kernel functions.

Another topic of interest, especially in a biomedical context, is the possibility to associate probability values with the outputs of machine learning model. While the outputs of suitably trained neural networks and logistic regression models can be interpreted as class membership probabilities, the same is not true for SVMs, which only provide outputs of −1 or +1. It is, however, possible to fit a logistic model to the "raw" SVM outputs $\sum \alpha_j t_i x_i \cdot x + b$ that determine the sign and distance from the hyperplane. The larger this distance, the higher the probability of belonging to the class associated with this side of the hyperplane. The exact form of the logistic model can be obtained by minimizing a cross-entropy error function.

Stephan Dreiseitl and Lucila Ohno-Machado

See also Artificial Neural Networks; Logistic Regression

Further Readings

Boser, B., Guyon, I., & Vapnik, V. (1992). *A training algorithm for optimal margin classifiers.* Fifth annual workshop on computational learning theory. New York: ACM Press.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning, 20,* 273–297.

Cristianini, N., & Shawe-Taylor, J. (2000). *An introduction to support vector machines and other kernel-based learning methods*. Cambridge, UK: Cambridge University Press.

Herbrich, R. (2001). *Learning kernel classifiers: Theory and algorithms.* Cambridge: MIT Press.

Platt, J. (2000). Probabilities for SV machines. In A. J. Smola, P. L. Bartlett, B. Schölkopf, & D. Schuurmans (Eds.), *Advances in large margin classifiers*. Cambridge: MIT Press.

Schōlkopf, B., & Smola, A. J. (2002). *Learning with kernels: Support vector machines, regularization, optimization, and beyond.* Cambridge: MIT Press.

Shawe-Taylor, J., & Cristianini, N. (2004). *Kernel methods for pattern analysis.* Cambridge, UK: Cambridge University Press.

Vapnik, V. N. (1998). *Statistical learning theory.* New York: Wiley.

Vapnik, V. N. (1999). *The nature of statistical learning theory* (2nd ed.). New York: Springer.

Surrogate Decision Making

Surrogate decision making is the process by which one or more medical treatment decisions, or other decisions relating to healthcare and personal welfare, are made on behalf of adults who are judged to lack the decision-making competence (or capacity) to make such decisions for themselves.

When Is Surrogate Decision Making Required?

Surrogate decision making is required when a medical decision needs to be made for a patient who is judged to be *unable* to give, or withhold, informed consent. Surrogate decision making is distinguished, therefore, from medical decision making *without* consent, which relates to compulsory psychiatric assessment or treatment for adults with mental disorders, and the involuntary detention and treatment of adults with communicable diseases that pose a risk to public health.

In numerous legal jurisdictions, decision-making competence acts as a threshold concept that preserves the right of self-determination. If adults are judged able to make an autonomous decision relating to their medical treatment, that decision must be respected, even if considered unwise. There is little academic consensus about the abilities that need to be demonstrated for an adult to be judged competent; however, the legal criteria for competence revolve primarily around a number of cognitive and communicative abilities. If a patient is, for example, unable to understand information relevant to the decision, to retain that information and weigh it to make a decision, or to communicate a choice, surrogate decision-making procedures will need to be invoked. Adopting a threshold for decision-making competence that is primarily cognitive in nature means that adults with a "mental disability" may require one or more medical decisions to be made on their behalf. This includes men and women with acute mental illness, dementia, intellectual disabilities, brain damage, or head injury or those who are unconscious, confused, or affected by fatigue, pain, or drugs.

Once a judgment of incompetence has been made, attention turns to how a surrogate decision should be made and who should make it. These

questions have shaped research into surrogate decision making, the majority of which lies at the intersection of law, ethics, and medical practice. Empirical studies exploring how surrogate decisions are made in practice have been a relatively recent development. This entry explores the ethical principles and regulatory procedures that have been developed for surrogate decision making for adults and highlights some problems that arise when invoking these principles and procedures in medical practice.

Ethical Principles for Surrogate Decision Making

Ethical and legal engagement with surrogate decision making has exposed three different principles.

Proxy Decision Making

Proxy decision making refers to a set of procedures, built on the primacy of autonomy, that aim to ensure that incompetent patients' *known wishes* guide the surrogate decision-making process. In the ideal scenario, incompetent patients would have made an advance directive and appointed a proxy decision maker, when competent, using a power of attorney. The proxy decision maker is then required to uphold the patient's advance directive if relevant. Despite academic debate about the nature of personal identity, advance directives are believed to best reflect autonomous choice. If no advance directive exists, the surrogate decision-making process should be guided by other principles.

Substituted Judgment

The substituted judgment principle requires a surrogate to attempt to make the same decision that the incompetent patient would make, if he or she were able to do so. The need for the surrogate to be able to "speak" for the incompetent patient means that the next of kin has become established as the most appropriate surrogate. This subjective approach to surrogate decision making requires a detailed inquiry into incompetent adults' lives to make a *best guess* about their preferences, with this guess being substituted for that which is impossible to obtain directly.

Best Interests

The best-interests principle requires a surrogate decision to be made by focusing on achieving the *best outcome*. Traditionally, this principle has been seen as paternalistic, with patients' interests being served by receiving treatment that is clinically indicated, such that surrogate decision making operated on the basis that "the doctor knows best." However, in legal jurisdictions where the best-interests principle dominates, it is now well established that the concept of best interests has both an objective and a subjective component. There is broad acceptance that best-interests judgments should incorporate objective medical evidence relating to the decision at hand, interpreted in the context of subjective values, wishes, and beliefs. In this way, the principle is used both to express ideals and goals and to find reasonable solutions to practical problems, such that the "good" of the decision for the individual is promoted to the greatest extent.

The Pervasiveness of Ethical Dilemmas in Medical Practice

While these ethical principles have developed in the context of the practical challenges posed by individual cases, operationalizing these principles to make a broad range of medical decisions, on behalf of different groups of individuals, remains problematic. For example, examining the subjective values of adults with profound intellectual and multiple disabilities is difficult, or, arguably, impossible. For surrogate decision making at the end of life, drawing on ethical principles might lead to decisions being made that are controversial and potentially unacceptable for people with certain religious beliefs.

For medical practice to avoid potential problems, legal regulation has formalized the procedures for surrogate decision-making around these principles in a number of jurisdictions to enable the surrogate decision-making process to be consistent, transparent, and defensible.

Legal Regulation of Surrogate Decision Making

The legal regulation of surrogate decision making in different jurisdictions revolves around a number

of procedural safeguards. To demonstrate the range of approaches that have been taken, the legal frameworks in England and the United States are compared and contrasted.

The English Approach

In England and Wales, the historical context of surrogate decision making lies in the *parens patriae* jurisdiction. This jurisdiction, which still exists for children, places an obligation on the state to protect its most vulnerable citizens. With the passing of mental health statutes, this jurisdiction fell into desuetude and was replaced nearly two decades ago, in the landmark case of *Re F* [1990], by the best-interests principle. Since 1990, the conceptualization of best interests in the common law has shifted from medical interests, defined objectively, to those relating to patients' broad welfare, defined subjectively.

The introduction of the Mental Capacity Act of 2005 (MCA) formalized the surrogate decision-making process and codified the best-interests principle for "acts in connection with the care or treatment" of incompetent adults. Reflecting developments in the common law, the MCA provides a "best-interests checklist" that requires healthcare practitioners to consider

- whether competence will be regained and, if so, when;
- whether the person can be permitted and encouraged to participate in the decision-making process regardless of incompetence;
- the person's past and present wishes, feelings, beliefs, and values;
- the views of other people who are deemed practicable and appropriate to consult; and
- all other circumstances deemed to be relevant.

Healthcare practitioners must, therefore, incorporate a range of objective and subjective evidence, grounded in the specific decision that is being considered, and based on consultation with individuals in a position to provide advice on patients' preferences. While family members should expect to be consulted, no surrogate or guardian is appointed, and the process of weighing a range of evidence to make the final decision lies with healthcare practitioners themselves.

The best-interests checklist approach represents an amalgamation of both the best interests and substituted judgment principles and is accompanied by additional safeguards for surrogate decision making in relation to medical matters. These include lasting powers of attorney, advance decisions to refuse treatment (the adherence to which by a surrogate decision maker is mandatory), advance statements (which are designed to provide a rich source of evidence about a person's treatment preferences but the adherence to which is not mandatory), and restricted decisions (such as the withdrawal of artificial hydration and nutrition from patients in a persistent vegetative state, nontherapeutic sterilization, and tissue donation), which can only be made by a court.

The U.S. Approach

In the United States, aspects of surrogate decision making are regulated in state law, and this process has become recognized as a way of upholding the protected liberty interests of U.S. citizens to receive beneficial medical treatment under the 14th Amendment of the U.S. Constitution.

In contrast to English law, the incompetent patients' next of kin have been accorded a key role. Following the landmark judgment *In re Quinlan* by the New Jersey Supreme Court in 1976, numerous state courts have drawn on the substituted judgment principle when no advance directive exists. While procedures differ markedly across different states, hierarchies for family members who should be appointed as surrogates have developed, and the decisions made by the individual(s) appointed are binding.

For incompetent patients whose preferences are judged unable to be known because, for example, they have lifelong and profound disabilities or when no appropriate surrogates are available, the best-interests principle has formed the basis of regulation. Acting in the best interests of incapacitated patients requires medical practitioners to exercise their judgment and expertise, with subjective factors being excluded as unreliable or unknowable.

The Empirical Validity of Ethical Principles

Recently, there has been increased interest in exploring surrogate decision making empirically,

with quantitative and qualitative studies examining the empirical validity of substituted judgment and best interests. These empirical studies have produced findings that problematize the conceptualization of surrogate decision making in law and ethics.

Empirical studies of the operation of the substituted judgment principle in practice have shown that there is little concordance between patients' wishes and the substituted judgments of family surrogates and doctors in a number of medical treatment scenarios. The reasons for surrogates and doctors predicting patients' treatment preferences incorrectly are explained on the basis of the liability for these preferences to change substantially over time and the family dynamics and stress associated with the responsibility of making decisions on behalf of loved ones.

Empirical studies of the operation of the best-interests principle in practice have confirmed that surrogate decision makers believe that their personal values and interests are important and relevant to the decisions that they make, challenging the possibility of their being able to ground their decisions in objective and subjective evidence relating to incompetent patients' lives only. It is also clear that familial influences in assessments of the "best" course of action are often underappreciated, particularly in cross-cultural contexts. In Pakistani culture, for example, it has been shown that an individual's wishes and values are seen as synonymous with those of his or her family and that it is common for healthcare practitioners to be placed in the role of a family member.

In practice, the formal legal duty to follow the best-interests checklist under English law will likely facilitate deliberation about the pros and cons of the treatment options available, but not help to determine the "best" outcome. While medical practitioners are obliged to obtain a range of objective and subjective evidence relating to the patient and the decision at hand, there is no guidance about the weight that should be given to different evidence.

Moving Forward

Attempts to both do justice to the ways that surrogate decisions are made in practice and to ensure that these decisions can be defended ethically have led to a reconsideration of the principles that *should* frame surrogate decision making *on the basis of empirical data*. Largely, this has involved reformulating surrogate decision making around a situated and relational approach, with calls for narrative engagement that aims to preserve the dignity and identity of the individual and with the incorporation of community norms to contextualize a broad range of relevant factors.

With regard to the regulation of surrogate decision making, there have been parallel calls for an approach built on legal pragmatism rather than legal formalism. A pragmatic interpretation would recognize that (a) surrogate decision making is a practical experiment; (b) surrogate decisions are made on the basis of a range of information much of which is likely to be limited, conflicting, and emotive; and (c) the decisions that are made may lead to a range of outcomes, many of which might be ethically defensible.

Michael Dunn

See also Advance Directives and End-of-Life Decision Making; Decision-Making Competence, Aging and Mental Status; Decisions Faced by Hospital Ethics Committees; Decisions Faced by Surrogates or Proxies for the Patient, Durable Power of Attorney; Informed Consent

Further Readings

Berger, J. T., DeRenzo, E. G., & Schwartz, J. (2008). Surrogate decision making: Reconciling ethical theory and clinical practice. *Annals of Internal Medicine, 149,* 48–53.

Buchanan, A. E., & Brock, D. W. (1990). *Deciding for others: The ethics of surrogate decision making*. New York: Cambridge University Press.

Coggon, J., & Holm, S. (2008). Best interests: A reappraisal. *Health Care Analysis, 16,* 193–196.

Dunn, M. C., Clare, I. C. H., Holland, A. J., & Gunn, M. J. (2007). Constructing and Reconstructing "best interests": An interpretative examination of substitute decision-making under the Mental Capacity Act 2005. *Journal of Social Welfare and Family Law, 29,* 117–133.

Holm, S., & Edgar, A. (2008). Best interest: A philosophical critique. *Health Care Analysis, 16,* 197–207.

In re Quinlan, 355 A.2d 647, 670 (N.J. 1976).

Kopelman, L. M. (2007). The best interests standard for incompetent or incapacitated persons of all ages. *Journal of Law, Medicine & Ethics, 35,* 187–196.

Mental Capacity Act (2005), c. 9. London: Stationery Office. Retrieved from http://www.opsi.gov.uk/ACTS/acts2005/ukpga_20050009_en_1

Re F (Mental Patient: Sterilisation) [1990] 2 AC 1.

Torke, A. M., Caleb Alexander, G., & Lantos, J. (2008). Substituted judgment: The limitations of autonomy in surrogate decision making. *Journal of General Internal Medicine, 23,* 1514–1517.

Survival Analysis

Survival analysis is the analysis of time-to-event data. The event may be death (hence the term *survival*), or it may be some other event such as cancer recurrence or a stroke. In engineering, when analyzing mechanical failures, the topic is known as reliability analysis. Although these methods are traditionally used for time-to-event data, they are also applicable to other types of data, for example, the total amount paid due to an accident. As with many statistical methods, there are parametric and nonparametric methods for survival analysis. Parametric models assume that the data come from some parametric distribution such as the normal distribution, which is defined by two parameters, the mean and the standard deviation. Nonparametric models do not assume a given parametric distribution.

Data have

1. a density function: $f(t)$ for $t \geq 0$, the probability of the event at time t;
2. a cumulative distribution function: $F(t) = \int_0^t f(t)dt$, the probability of the event before time t;
3. a survival function: $S(t) = 1 - F(t)$, the probability of surviving (event-free) until time t;
4. a hazard function: $\lambda(t) = f(t)/S(t)$, the probability of the event at time t given survival until time t (may also be denoted $h(t)$);
5. a cumulative hazard function: $\Lambda(t) = \int_0^t \lambda(t)dt = \ln S(t)$ (may also be denoted $H(t)$);
6. mean survival time: $\int_0^\infty tf(t)dt$;
7. median survival time: t_{med} such that $S(t_{\text{med}}) = .5$.

As usual in statistics, the density function is non-negative and the integral from 0 to ∞ is 1, the cumulative distribution function ranges from 0 to 1, and the survival function ranges from 1 to 0.

Caution should be exercised as to the choice of time 0. If the event of interest is survival after a clinical procedure, then $t = 0$ is the time of that procedure. For individuals randomized into a clinical trial, $t = 0$ is often the date of randomization. The start of the study is not appropriate as a $t = 0$ for individuals who enter the study after the starting date.

As an example (of a parametric distribution), suppose the event data have an exponential distribution, $f(t) = \lambda e^{-\lambda t}$. Remember that this is the distribution of event times. The density function, $f(t)$, for the exponential distribution is monotone, decreasing with its highest value, λ, at $t = 0$. The cumulative distribution function is then $F(t) = 1 - e^{-\lambda t}$, the survival function $S(t) = 1 - F(t) = 1 - 1 + e^{-\lambda t} = e^{-\lambda t}$, the hazard function $\lambda(t) = f(t)/S(t) = \lambda e^{-\lambda t}/e^{-\lambda t} = \lambda$ (a constant), and the cumulative hazard function $\Lambda(t) = \lambda t$; the mean survival time is $\int_0^\infty t\lambda e^{-\lambda t} dt \quad 1/\lambda$, and the median survival time is $\ln(2)/\lambda$. The exponential distribution was one of the first distribution functions used as the calculations are reasonably straightforward. It is thought that some manufactured items such as light bulbs have an exponential distribution for their failure times. However, it is now recognized that a constant hazard function may be too restrictive. Also, with the advent of readily available cheap computing, models are no longer chosen for computational simplicity.

Parametric survival analysis assumes a parametric distribution for the density function such as the exponential distribution described above. Other density functions frequently used for survival analysis include the Weibull, Gamma, normal, and lognormal density functions.

Often, one does not want to assume a particular parametric distribution. Therefore, nonparametric analyses are desired. If one has complete data on n people or objects of interest, that is, you know the event time t_i, $i = 1, \ldots, n$, for all n observations, the survival function is simply a step

function starting at 1 (at time 0), which drops by $1/n$ at each of the event times until it reaches 0 at the last event time. For simplicity, let the distinct event times be $0 \le t_1 < t_2 < \ldots < t_n$ and the number at risk be $n_1 = n$, $n_2 = n - 1$, …, $n_i = n - i + 1$, …, $n_n = n - n + 1 = 1$. The survival function can be written as

$$S(t) = \prod_{t_j \le t} (n_j - 1)/n_j.$$

Note that the symbol $\prod_{t_j \le t}$ means the product of terms indexed by j, where $t_j \le t$. As an example, suppose there are 10 observations. At the first event time t_1, there are 10 people at risk, and therefore, the survival function at the first event time is 9/10 (only t_1 is less than or equal to t_1). At the second event time t_2, there are 9 people at risk, and the survival function is (9/10)(8/9) = 8/10 (here t_1 and t_2 are the only times less than or equal to t_2). If there are multiple events at a given time, then let the k distinct event times be $0 \le t_1 < t_2 < \ldots < t_k$, the number of events at t_i equal d_i, and the number at risk (just prior to t_i) be equal to n_i. The survival function then drops by d_i/n at each event time and can be written as

$$S(t) = \prod_{t_j \le t} (n_j - d_j)/n_j. \quad (1)$$

Often, one does not have complete data, as some observations are censored. Observations may be censored for many reasons (such as the study ended but a participant had not yet experienced the event of interest or a participant is lost to follow-up). What is known is that the event time is at least some value (such as the time to the end of the study or the time last observed free of the event of interest). In addition to the event or censored times, t_i, one observes an indicator variable I_i, where $I_i = 1$ for an observed event and $I_i = 0$ for a censored observation. The number at risk, n_i, remains the same; however, the survival function only changes at event times, not at censored observations. The survival function is then

$$S(t) = \prod_{j:t_j \le t \text{ and } I_j = 1} (n_j - d_j)/n_j. \quad (2)$$

Here, it should be noted that $\prod_{j:t_j \le t \text{ and } I_j = 1}$ means the product of terms indexed by j such that $t_j \le t$ and $I_j = 1$; that is, the product includes terms whose event time is less than or equal to t (and only includes j such that t_j is time to an event, not a censoring time). Equation 2 is known as the Kaplan-Meier estimate or product-limit estimate of the survival function. Note that the mean survival time cannot be calculated unless $S(t_n) = 0$. Therefore the median survival time is often reported if it exists.

As an example, suppose the data in the table below are observed for 10 participants. It is possible to calculate a standard deviation for the Kaplan-Meier estimate using Greenwood's formula. Also, the log-rank test may be used to compare two or more survival functions calculated using Kaplan-Meier estimates.

Although Kaplan-Meier estimation has been used for 50 years, an older nonparametric method of survival analysis has been in use for a century or more. It is known as life table analysis and has been used extensively in demography and actuarial science. The exact lifetime or censoring time (withdrawal time) of a given individual is unknown, only that the event or withdrawal occurred within a given interval of time (e.g., month or year). The survival function is then

Observation	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*
t_i	3	5	8	11	12	17	20	30	30	30
I_i	1	0	0	1	0	0	1	0	0	0
n_i (at risk)	10	9	8	7	6	5	4	3	2	1
S	*a*	*a*	*a*	*b*	*b*	*b*	*c*	*c*	*c*	*c*

Note: a = 9/10, b = (9/10)(6/7), c = (9/10)(6/7)(3/4).

calculated as a series of probabilities of survival of each interval, given that one has survived to the beginning of the interval.

Parametric models may be estimated with complete or censored data by using the appropriate likelihood function. However, it should be noted that parameter estimates and inferences may not be robust to departures from the parametric assumptions. This is the reason why most survival analyses of medical data are based on nonparametric methods.

The next step in survival analysis is often the incorporation of additional information in the form of covariates such as gender, age, comorbidities, cancer stage, and so on. For parametric models, the parameters are often modeled as a function of one or more covariates. There are no strictly nonparametric methods to incorporate covariates. However, a semiparametric method is available if one assumes that the hazard function includes covariates. For example, if the hazard function is equal to

$$\lambda(t) = \lambda_0(t)e^{b'X},$$

where $\lambda_0(t)$ is an unspecified hazard function and b' a vector of parameters for the vector X of covariates (which may or may not depend on t), you have Cox's proportional hazard model. It should be noted that some parametric models (e.g., Weibull and exponential) are themselves proportional hazard models.

Carol L. Link

See also Cox Proportional Hazards Regression; Likelihood Ratio; Log-Rank Test

Further Readings

Kalbfleisch, J. D., & Prentice, R. L. (2002). *The statistical analysis of failure time data* (2nd ed.). New York: Wiley.

Kaplan, E. L., & Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association, 53*(282), 457–481.

Lawless, J. E. (2003). *Statistical models and methods for lifetime data* (2nd ed.). New York: Wiley.

Miller, R. G. (1981). *Survival analysis*. New York: Wiley.

TABLES, TWO-BY-TWO AND CONTINGENCY

Contingency tables are used in medical research to organize summary data from research studies. The data are arranged in columns and rows according to two or more categorical variables. The term *two-by-two* (2 × 2) *table* is used to describe the specific contingency table that compares two dichotomous categorical variables and thus has two columns and two rows of data. Contingency tables are also sometimes called *row-by-column tables*, or *R-by-C tables* for short. Contingency tables efficiently display summary data about two or more categorical variables. A variety of statistical tests can be quickly calculated with the data contained in a contingency table.

Structure

The "classic" 2 × 2 table will have a dichotomous outcome variable labeled at the top of the two columns and a categorical exposure variable labeled to the left of the two rows. A typical structure for a 2 × 2 table is shown in Table 1.

As in Table 1, rows and columns are typically summed individually with the totals entered into extra cells to the right and to the bottom of the table. Care must be taken when analyzing 2 × 2 tables because the rows and columns may be organized differently by different researchers (i.e., the first column may contain patients with disease in one study but might contain patients without disease in another study).

Application

Contingency tables are an efficient means for quickly summarizing the association between an independent and a dependent variable. Contingency tables are used extensively in epidemiological studies and are very popular for case-control studies. In addition, contingency tables can be used to evaluate the predictive ability of medical tests. The gold standard for the presence or absence of disease can be compared with the dichotomous result of a new medical test using the same format as in the first example. In this way, sensitivity, specificity, positive predictive value, and negative predictive value can be quickly calculated with the data in this table. These characteristics of medical tests form the backbone of medical decision making surrounding medical tests (i.e., Which test is most important for screening? Which test is most appropriate as a confirmation test?).

Contingency tables can also be used to calculate odds ratios, relative risks, kappa statistics, chi-square statistics, Fisher's exact test, McNemar's test, and so on.

Example

Table 2 shows a silly, hypothetical example that uses a 2 × 2 table to display the results of a study. A researcher develops a theory that the mechanical

Table 1 Structure of the traditional 2 × 2 table

	Outcome		
	Developed Disease	*Did Not Develop Disease*	*Totals*
Exposed	*A*	*B*	*A + B*
	Number of exposed patients who developed disease	Number of exposed patients who did not develop disease	Total number of exposed patients
Not exposed	*C*	*D*	*C + D*
	Number of unexposed patients who developed disease	Number of unexposed patients who did not develop disease	Total number of unexposed patients
	A + C	*B + D*	*A + B + C + D*
Totals	Total number of patients with disease	Total number of patients without disease	Total number of patients in the study

Table 2 Example 1: Case control of Alzheimer's disease

	Alzheimer's Patients	*Controls*	*Totals*
History of skipping rope	38 (*a*)	22 (*b*)	60 (*a + b*)
No history of skipping rope	62 (*c*)	78 (*d*)	140 (*c + d*)
Totals	100 (*a + c*)	100 (*b + d*)	200 (*a + b + c + d*)

vibrations experienced by school-age children while skipping rope cause microtrauma to the brain, which could increase the risk of Alzheimer's disease in old age. The researcher performs a case-control study in which she asks Alzheimer's patients and control patients if they had ever skipped rope as a child. The results reveal that 38 out of 100 Alzheimer's patients and 22 out of 100 control patients report a history of skipping rope. The results of this study are reported in Table 2.

The odds ratio can be easily calculated using the cross products from the 2 × 2 table:

$$\text{Odds ratio} = (a \times d)/(b \times c) = (38 \times 78)/(22 \times 62) = 2.17$$

The odds ratio indicates that the Alzheimer's disease patients were 2.17 times more likely to have a history of skipping rope than the control patients.

Calculating a relative risk from this study would not be appropriate because of the case-control

Table 3 Example 2: Skipping rope

	Alzheimer's Patients	*Controls*	*Totals*
Excessive skipping rope	16	11	27
Low to moderate skipping rope	22	11	33
No history of skipping rope	62	202	264
Totals	100	224	324

Table 4 Example contingency table from a matched case-control study

		Matched Controls		
		History of Skipping Rope	*No History of Skipping Rope*	*Totals*
Matched cases	History of skipping rope	8 A (concordant pairs)	30 B (discordant pairs)	38
	No history of skipping rope	14 C (discordant pairs)	48 D (concordant pairs)	62
	Totals	22	78	100

Note: Each cell in this table refers to "pairs" of patients. For instance, cell "A" has 8 pairs of patients, or 16 patients.

design and high frequency of exposures in both groups. Even though it's not plausible, if we assume for the moment that the data in Table 2 were somehow collected in a prospective fashion, then we can calculate the relative risk. The relative risk is defined as the proportion of people who develop the disease among the exposed divided by the proportion of people who develop the disease among the unexposed:

$$\text{Relative risk} = (a/(a+b))/(c/(c+d)) = (38/60)/(62/140) = 1.43.$$

A relative risk of 1.43 would indicate that children who skip rope are 43% more likely to develop Alzheimer's disease later in life than children who do not skip rope.

Continuing with our example, the researcher decides that the categorical variable for "skipping rope" may not be adequately discriminating patients according to exposure. She decides to repeat the survey and asks more detailed questions about the exposure to skipping rope. Patients are then stratified into the following rope-skipping categories: none, low to moderate, and excessive. The new data are reflected in Table 3, which would now be called a 3×2 contingency table.

A colleague points out to the researcher that the risk of dementia increases with age and suggests that age might confound any associations that are observed between Alzheimer's disease and a history of skipping rope. Data for matched case-control studies are generally entered into 2×2 tables as pairs of data, as described in the following example: The researcher decides to match each of the cases to a single control patient of the same age. The results are the same as the first example and reveal that 38 out of 100 Alzheimer's patients have a history of skipping rope compared with 22 out

of 100 control patients. These numbers have been entered into Table 4. With the original information given, we can only fill in the totals of each row and column. Detailed information about the exposures of each pair of patients would be required to fill in the rest of the table. For instance, a matched pair in which the case had a history of skipping rope and the control had no history of skipping rope would be counted in Cell B of the table (note that Cell B = 8, which represents 8 pairs of patients, or a total of 16 patients). Let's assume that the study obtained the results entered into Table 4.

We could try to determine if the cases and controls had statistically different historical exposures to skipping rope using the McNemar chi-square test:

$$\text{The McNemar statistic } (\chi^2) = (b - c)^2/(b + c) = 5.81,\ p = .016.$$

These results indicate that there is only a 1.6% chance that the difference observed between skipping rope and Alzheimer's disease occurred by chance.

Brian J. Wells

See also Basic Common Statistical Tests: Chi-Square Test, *t* Test, Nonparametric Test; Bayes's Theorem; Case Control; Diagnostic Tests

Further Readings

Berger, R. L. (2003). Exact unconditional tests for a 2 × 2 matched-pairs design. *Statistical Methods in Medical Research, 12,* 91–108.

Gordis, L. (2004). *Epidemiology*. Philadelphia: W. B. Saunders.

Haviland, M. G. (1990). Yates's correction for continuity and the analysis of 2 × 2 contingency tables. *Statistics and Medicine, 9,* 363–367.

Pezzullo, J. C. (2005, April). *Contingency tables, cross-tabs, chi-square tests*. Retrieved May 2, 2008, from http://statpages.org

Richardson, J. T. (1994). The analysis of 2 × 1 and 2 × 2 contingency tables: An historical review. *Statistical Methods in Medical Research, 3,* 107–133.

Rigby, A. S. (2001). Statistical methods in epidemiology. VII. An overview of the chi^2 test for 2 × 2 contingency table analysis. *Disability and Rehabilitation, 23,* 693–697.

Teaching Diagnostic Clinical Reasoning

Most medical students in the United States are not systematically taught clinical reasoning with the same rigor as they are taught the medical interview and physical examination. Despite the exponential growth of interest in evidence-based medicine, the integration of such inquiries into clinical reasoning with individual patients remains a relatively weak link.

Diagnostic strategies of inexperienced medical students frequently begin with an exhaustive collection of data, whereas expert clinicians use multiple complex "scripts" gained through reflection on clinical experience. Experts test a limited number of hypotheses starting early in the interview and move quickly to closure, sometimes with inappropriately heavy reliance on the clinical laboratory. Yet if the laboratory is routinely used to rule out improbable diagnoses without estimating pretest probability, the odds of missing serious diseases with false-negative tests, as well as giving people inaccurate diagnoses based on false-positive results, are significantly increased.

The case of Mrs. B is used throughout this entry for illustrative purposes: Mrs. B was a 40-year-old woman who presented to the emergency department with substernal chest pressure radiating to her neck associated with shortness of breath, palpitations, numbness in her hands and lips, and a feeling of impending doom. She smoked one pack per day, and her father had his first myocardial infarction (MI) at age 55. Other than a heart rate of 104, her physical examination was normal, as was her initial electrocardiogram. Her admitting diagnoses were rule-out MI and rule-out pulmonary embolism.

Initial Diagnostic Approaches

Table 1 summarizes the four most common diagnostic approaches used by clinicians at various levels of training. In traditional medical school curricula, beginning students are taught to conduct an *exhaustive review* of a patient's medical history and physical examination before initiating clinical reasoning.

Table 1 Initial diagnostic approaches

Exhaustive review of history and physical exam
Pattern recognition (Gestalt)
Multiple branching (arborization)
Hypothetico-deductive

In our illustrative case, Mrs. B's past medical history included several emergency department visits during her early 20s for similar chest pain that defied diagnosis. Her younger sister was bothered by "anxiety attacks." An exhaustive review of systems uncovered a feeling of unreality and terror during the episode. These data were in the medical student note, unread by the rest of the team.

Experienced clinicians quickly *recognize patterns* of symptoms and signs that mirror previously seen pictures of disease. For example, Mrs. B's attending physician knew that coronary artery disease is frequently underdiagnosed in women and thought that her pattern and risks fit reasonably well. No one on the team matched the pattern of her presentation with panic disorder, so it was not initially considered.

Multiple branching algorithms have been proposed for exploring and evaluating common clinical problems such as chest pain in the emergency department. Here, the clinician asks a series of yes/no questions where the answer determines the next step, usually based on the best available clinical evidence for populations of similarly situated patients. In the example above, Mrs. B was placed on "rule out MI" and "rule out pulmonary embolism" algorithms based on her presentation with substernal chest pain. Troponins were drawn, a ventilation-perfusion scan was ordered, and she was tentatively scheduled for an exercise tolerance test. No estimate of pretest probability for either diagnosis was in the chart. Her ventilation-perfusion scan was "low probability." Her stress test showed minor nonspecific flattening of her T waves at an excellent rate-pressure product.

The validity, accuracy, and efficiency of the *hypothetico-deductive approach* improve with knowledge and experience. Clinicians using this method are actively listening and looking for patterns that "fit" with clinical presentations that they recognize. Experienced clinicians store large numbers of scripts against which a patient's presentation is rapidly tested, moving to closure quickly using a mix of symptoms, signs, epidemiology, pattern recognition, and algorithms. For example, Mrs. B's physicians entertained two main hypotheses based on what they thought was most life threatening, but they did not explicitly estimate the probability of coronary artery disease or pulmonary embolism. If they had considered the underlying epidemiology of chest pain in 40-year-old women and Mrs. B's atypical associated symptoms (feeling of depersonalization, past history of similar episodes in her 20s, family history of anxiety attacks), other diagnoses such as panic disorder might have been considered.

A Six-Step Approach to Teaching Probabilistic Clinical Reasoning

Table 2 outlines a six-step approach to teaching diagnostic clinical reasoning that formally integrates the hypothetico-deductive approach and evidence-based medicine. Although diagnostic accuracy frequently depends on substantial baseline knowledge and experience, being explicit about the process of probabilistic reasoning over time will improve decision making for all levels of clinicians. This six-step process is often used to teach and evaluate undergraduate medical students throughout all 4 years, and the same strategy is used to teach internal medicine and family medicine residents, especially during attending rounds and the morning report. In addition, experienced clinicians can use the same method when confronted with a challenging clinical problem.

Step 1: Generate a Problem List

Care should be taken in the selection of accurate names for problems. For example, "chest pain" describes a symptom with an uncertain cause, whereas "angina" clearly implicates a diagnostic etiology. This problem list should be exhaustive and multidimensional, including all "loose ends." The problem list may include symptoms, physical findings, laboratory abnormalities, past diagnoses, salient psychosocial problems, and family history. The challenge is to provide the reader of the medical record a chance to rethink the diagnosis based on unique ways of putting the constellation together.

Mrs. B's problem list might have included an acute episode of chest pain, shortness of breath,

Table 2 A six-step approach to probabilistic clinical reasoning

Step 1: Generate a problem list
Step 2: Brainstorm diagnostic possibilities around major problem(s)
Step 3: Place general probabilities on diagnoses
Step 4: Decide on probabilistic thresholds for action (or inaction)
Step 5: Select areas for further inquiry (history, physical exam, diagnostic tests)
Step 6: Act once thresholds are achieved

palpitations, numbness in the hands and lips, and feelings of terror and impending doom. The list would also include a one-pack-per-day smoking and a family history of coronary artery disease (in her father), as well as a history of similar episodes in her early 20s and a family history of anxiety attacks (in her sister).

Step 2: Brainstorm Diagnostic Possibilities Around Major Problem(s)

Critical thinking is put aside, and any ideas about what could explain the patient's main problem(s) are encouraged. Loose ends from the problem list should be reviewed to see if any new ideas emerge. Both common and rare conditions should be considered, as well as biomedical and psychosocial explanations.

The clinicians caring for Mrs. B were most worried about coronary artery disease and pulmonary embolism, but they considered no other possibilities. A brainstorm might have added a variety of relatively common (musculoskeletal pain, panic attack, pericarditis) as well as relatively rare (dissecting aortic aneurysm, pheochromocytoma) conditions.

Step 3: Place General Probabilities on Diagnoses

Goodness of fit of the patient's presentation to typical disease patterns and knowledge of basic epidemiology should play major roles in determining probabilities. From the original brainstorm list, diagnoses should be divided into those that are "likely" (best fit), "possible" (could fit), and "very unlikely" (do not fit well at all). The "very unlikely" group should be temporarily set aside, and actual probabilities should be placed on the "likely" and "possible" diagnoses so that the total of all probabilities equals 100%. An attempt should be made to achieve a consensus among participants about the probabilistic range for each possible diagnosis using the best available clinical and epidemiologic evidence.

Both coronary artery disease and pulmonary embolism would be considered both "possible" and "worrisome," yet other possibilities should have been considered. The characteristic pattern of panic attacks along with the genetic and gender predisposition made this diagnosis likely. Musculoskeletal chest pain is also very common but could not explain many of this patient's symptoms. Lack of physical and laboratory findings (and relative rarity) made pericarditis, pheochromocytoma, and aortic dissection very unlikely. Given her clinical presentation, the probabilities might have been panic attack (60%), unstable angina (20%), pulmonary embolism (10%), and musculoskeletal pain (10%).

Step 4: Decide on Probabilistic Thresholds for Action (or Inaction)

All diagnoses in the "likely" group should be ruled in or out if possible. Those that are in the "possible" category should be worked up if life threatening or if making a firm diagnosis would decisively change management. If a diagnosis is only "possible" and the consequences of delay in diagnosis are not significant, the natural history of the disease might determine the diagnosis through watchful waiting. Potentially serious but "very unlikely" diagnoses should not be worked up unless the probabilities change based on subsequent information.

In Mrs. B's case, panic attack should clearly be ruled in since it was the most "likely" diagnosis and not making the diagnosis can have serious consequences. Coronary artery disease and pulmonary embolism were "possible," but the consequences of missing these diagnoses can be lethal,

so they needed to be ruled out. Musculoskeletal etiologies were also "possible," but these can be managed by watchful waiting since the consequences of delay in diagnosis would be minor. Pericarditis, aortic dissection, and pheochromocytoma were all "worrisome" but "very unlikely" and would not be worked up at this time.

Step 5: Select Areas for Further Inquiry

The goal of further inquiry is to reduce diagnostic uncertainty and to cross a threshold toward effective action or inaction. Clinicians should learn to think out loud and justify why further inquiry will help them make a more accurate diagnosis. Inquiry may involve gathering more information from the history, the physical examination, the diagnostic test(s), a therapeutic trial, or watchful waiting.

The key to ruling in the most likely diagnosis of panic disorder would have been to further explore the history of the prior attacks in the early 20s and the family history of anxiety disorders. Concurrent stresses and exposure to stimulant medications, alcohol, or drugs might also have been explored. Even with this additional information, coronary artery disease or pulmonary embolism would not have gone below the threshold to be ruled out. However, the low pretest probability of pulmonary embolism (10%) in conjunction with a ventilation-perfusion scan interpreted as "low probability" took this diagnosis below the threshold for further workup of pulmonary embolism. Similarly, the low pretest probability of coronary artery disease (20%) combined with the relatively negative stress test took Mrs. B below the threshold for further workup of coronary artery disease.

Step 6: Act Once Thresholds Are Achieved

This step illustrates the interface between clinical reasoning and clinical judgment. Most diagnoses are not 100% certain; so even with a "highly likely" diagnosis, clinicians should rethink probabilities if unexpected changes occur. It is wise to track a patient's problem list and loose ends and keep an open mind to the unexpected. Finally, the patient's values and preferences should be taken into account before making any significant clinical intervention.

Mrs. B had a very high likelihood of having panic disorder, and her chances of significant coronary artery disease or pulmonary embolism became "very unlikely." She was given information about the diagnosis and treatment of panic disorder and of the potential consequences if it were left untreated (phobias, depression, hypochondriasis, iatrogenic problems, suicide attempts). The clinician felt that the diagnosis was highly likely (95%) but worried about how to stay alert about the 5% chance that something else was causing the disorder.

Potential Pitfalls to Good Clinical Reasoning

Table 3 outlines several pitfalls that frequently undermine good clinical reasoning, and some questions to counteract them. *Premature closure* is one of the most common. In Mrs. B's case, the mention of chest pain led to exclusive inquiries directed toward coronary artery disease and pulmonary embolism without considering other possibilities. A variant of this pitfall is to *consider only one (or two) hypothesis at a time*. These two diagnoses, while both serious, only had a cumulative probability of 30%, begging the question about what makes up the remaining 70%.

Errors frequently stem in part from clinicians *overly shaping the patient's story*. The mention of chest pain, especially in the emergency department, frequently leads to a highly structured, physician-controlled inquiry about specific symptoms to rule in or out potentially lethal diagnoses. This approach may cause the physician to miss relevant historical data crucial in establishing a more accurate diagnosis. A related phenomenon is the tendency to *overestimate the probability of worrisome diagnoses*, confusing likelihood with clinical concern.

Another potential error is physician-driven *endless inquiry without decisions, leading to a cascade of potentially inappropriate decisions*. In Mrs. B's case, the minor nonspecific T-wave changes on her treadmill test might have led to a stress thallium test and/or a cardiac catheterization, each of which might have had nonspecific findings of uncertain significance. Unless one thinks probabilistically and simultaneously attempts to rule in the most likely diagnosis, there is potential to overinterpret these nonspecific findings, leading to a false-positive

Table 3 Some pitfalls to good clinical reasoning

Pitfall	*Counterbalancing Questions*
Premature closure	*Do we have a complete accounting of the patient's history? What else might cause this picture?*
Considering only one hypothesis at a time	*Let's think of two or three other diagnoses that might explain this patient's problem.*
Overly shaping the patient's story	*Let's go over your experience one more time. Please let me know if I do not have it right.*
Overestimating the probability of worrisome diagnoses	*Let's try to separate what we are most worried about from what is most likely.*
Endless inquiry without decisions, leading to a cascade of diagnostic tests	*What are you going to do differently with the results of the next test?*
Ignoring data that do not fit	*Which data do not fit? What possibilities do they raise?*
Overemphasis on recent experience (recall bias)	*Have you seen a patient with a similar presentation in the past? What did you learn from that case?*
Overreliance on prior diagnoses or tests	*Let's review the primary data that led to this prior diagnosis. How certain are we that it is accurate?*

diagnosis. Moving down such a path might involve *ignoring data that do not fit*, such as the full constellation of symptoms consistent with panic attack, as well as past episodes and family history.

If the physician had recently seen a young female patient with coronary artery disease, he or she may *overemphasize recent experience (recall bias)* and overestimate the statistical risk of coronary artery disease in this patient. Finally, if this patient had falsely been given a diagnosis of coronary artery disease, the next time she presented, there might be an *overreliance on prior diagnoses or tests*. The parts of her story that did not fit would be ignored, labeling her clinical situation as a treatment failure rather than a wrong diagnosis. Healthy skepticism and a need to independently reconfirm prior diagnoses are the basic skills of experienced clinicians.

Continued clinical curiosity about how to account for a patient's unique experience (including the elements that do not seem to fit) in the context of medical diagnostic models, pathophysiology, epidemiology, and probability is essential. Using one's clinical experience while at the same time maintaining the openness of a beginner's mind is an ongoing challenge, and it is sometimes helpful if there are a few beginners in the group thinking through the problem. Reasoning through this six-step process with students and colleagues is good practice for eventually sharing one's thinking as transparently as possible with patients and their families.

Timothy E. Quill, Nancy S. Clark, Kathryn Markakis, Ronald Epstein, Donald Bordley, and Robert Panzer

See also Diagnostic Process, Making a Diagnosis; Diagnostic Tests; Differential Diagnosis; Errors in Clinical Reasoning; Hypothesis Testing; Probability Errors; Problem Solving

Further Readings

Barrows, H. S. (1994). The physician's clinical reasoning process. In H. S. Barrows (Eds.), *Practice-based learning: Problem-based learning applied to medical education* (pp. 12–23). Springfield: Southern Illinois University School of Medicine.

Elstein, A. S. (1999). Heuristics and biases: Selected errors in clinical reasoning. *Academic Medicine, 74,* 791–793.

Epstein, R. M., Alper, B. S., & Quill, T. E. (2004). Communicating evidence for participatory decision making. *Journal of the American Medical Association. 291,* 2359–2366.

Groves, M., O'Rourke, P., & Alexander H. (2003). The clinical reasoning of diagnostic experts. *Medical Teacher, 25,* 308–313.

Groves, M., Scott, I., & Alexander, H. (2002). Assessing clinical reasoning: A method to monitor its development in a PBL curriculum. *Medical Teacher, 24,* 507–515.

Kassirer, J. P. (1989). Diagnostic reasoning. *Annals of Internal Medicine, 110,* 893–900.

Mandin, H., Jones, A., Woloschuk, W., & Harasym, P. (1997). Helping students learn to think like experts when solving clinical problems. *Academic Medicine, 72,* 173–179.

Pauker, S. G., & Kassirer, J. P. (1980). The threshold approach to clinical decision making. *New England Journal of Medicine, 302,* 1109–1117.

Schmidt, H. G., Norman, G. R., & Boshuizen, H. P. (1990). A cognitive perspective on medical expertise: Theories and implications. *Academic Medicine, 65,* 611–621.

Thibault, G. E. (1994). The appropriate degree of diagnostic certainty. *New England Journal of Medicine, 331,* 1216–1220.

Team Dynamics and Group Decision Making

Individuals make decisions differently in teams than they do on their own. Some of these changes are improvements. Teams are more likely to bring different perspectives to the table. Sometimes, however, they can make decision making worse. An extensive literature exists on team building.

Within medical decision making, this literature has been applied to understanding how to improve clinical performance within teams. Particular attention has been devoted to performance improvement and eliminating preventable errors. Although this literature began with the study of individual cognition and performance, there has been increasing recognition that care is provided by multiple providers, working in formal or informal teams. Within institutions, care must be handed off; issues of communications and the transfer of information are thus critical. One element is the need to back up this communication with documentation, for which electronic tools can be valuable. Another question is regarding who can be considered a member of the team. The literature about shared decision making between patients and providers can thus be considered an aspect of team decision making.

This entry discusses several aspects of teams and team building, including the key principles; elements for the formation of effective teams, successful team functioning, and effective meetings; and types of team members. The entry also addresses common problems and the importance of trust. Finally, the entry applies teams and team building to medical care.

Key Principles

The literature on teams and team building suggests key principles, which may be easier or harder to achieve in practice:

- The team must have a leader who is responsible for the outcomes expected of the team, although an effective leader will rely on the team itself to work together to deliver these results.
- The team must be focused on quantifiable goals so that the members have a very clear idea of what they're trying to achieve.
- The team must have clearly defined roles, so that every member of the team knows exactly what he or she must do on a daily basis to avoid crossovers and cross-purposes.
- The team must be willing to share the resources under its control—talent and money—so that the team can achieve its goals.
- The team must establish good communication; that is, it must be frequent and effective. Frequency is fairly easy to achieve; effectiveness is far more difficult but required.
- The team must have constancy and consistency and be fully committed to its goals.

Starting Teams: Forming, Storming, Norming, Performing

In most organizations, individuals are assigned to teams rather than voluntarily choosing to join. The literature suggested several elements seen to make these teams more effective, commonly referred to by the following rhyming terms associated with Bruce Tuckman.

Forming. This term refers to the first stage of teaming, which is designed to allow team member designees to begin their team processes by gathering together. To the extent that what is said and done

at the first few meetings will mold the team and affect team functioning, the literature suggests that leaders with more outgoing personality types can be more effective at making team members feel welcome, facilitating discussion, and clarifying goals. Relevant materials can also be helpful in clarifying team goals.

Storming. This descriptor refers to brainstorming potential solutions to the problem that is presented. Successful teamwork during this stage, therefore, requires good information on what is to be done (e.g., project overview, guidelines, and other specifications) and good structure and guidance during the discussion process (e.g., to prevent it being hijacked by a person who talks too much; getting mired in disagreements, slights, and innuendo; or simply being allowed to go on too long and well past the point when adequate information has been gained).

Norming. Group norming processes are used to bring group members together so that the members begin to function in consonance with each other. These guidelines tend to be enforced initially by the team leader, but in well-functioning teams, they will be shared by all team members.

Performing. Good group performance, of all members, is a necessity for the appropriate completion of work projects. It is not enough for one or two members to conduct their jobs well; all group members need to be able to perform at appropriate levels and to complement one another's contributions and personal behavioral tendencies. This will also involve measuring and tracking protocols, as well as "softer" approaches to enhance good working relationships.

Content, Process, and Successful Group Functioning

The literature thus stresses that both content and process are essential for successful group functioning. There is considerable emphasis on being a good teammate, which has been described as being all about thoughtful behavior and mutual respect (e.g., offering to help teammates; arriving on time for team meetings; listening attentively to what members have to say, accepting idiosyncrasies, and sharing the excitement of others). Teams must be able to process information and deal with it without wasting time, but they must also make good decisions. Within healthcare settings, emphasis is heavily on content.

Types of Team Members

The literature also speaks of two different categories of roles of team members: (1) roles related to *tasks* and (2) roles related to *relationships*.

Task-Related Roles

Task roles can be characterized as *Harmonizer, Analyzer, Gatekeeper*, and *Encourager*. A harmonizer is one who recognizes conflicts and instigates discussion to help resolve differences. An analyzer watches for how well people are working together and takes action (e.g., if the team has lost energy). A gatekeeper stresses communication and participation and ensures that all team members have opportunities to participate. An encourager is one who shows support for the efforts, ideas, and achievements of team members.

Relationship-Related Roles

Relationship roles are the following: *Initiator, Orienter, Fact Seeker*, and *Summarizer*. An initiator helps ensure that there is agreement on how to proceed. An orienter helps ensure that the team stays on the topic. A fact seeker is one who tests reality, ensures that there is adequate information to back up decisions, and ensures that the team has the authority to act on the decisions it makes. A summarizer urges the group to reach a decision.

Effective Team Meetings

The literature often focuses on formal team meetings. The elements noted as being critical to an effective meeting are good preplanning and the ability to enact certain required strategies, including the following: (a) a good game plan, (b) appropriate decision-making processes, (c) effective brainstorming processes, (d) good ground rules, and (e) organized and informative minutes. Decision making can use majority rule or consensus. One suggestion is to reserve consensus forms

for vital tasks while using majority rule decision making for smaller matters not considered important enough to require that level of time and attention. Similarly, processes vary in the extent to which they seek creative solutions (where brainstorming is a popular approach) as opposed to requiring a hard look at all the relevant information available. Possible approaches include dialoging (a rapid process in which team members comment on the ideas as they are presented), force field analysis, fish boning, root cause analysis, and priority gridding.

Various approaches can be used to communicate results to other team members. Flip charts, blackboards, and similar approaches are often useful in meetings; to ensure that material is not forgotten, recording methods can range from minutes (for formal meetings) to various electronic health records.

Common Team Problems

Teams come up against many obstacles that impede or limit their progress. The team may not be composed of the right players, may not have the best mix of complementary skills, or may lack adequate resources and/or effective leadership.

A major challenge is conflict among team members. Teams seeking creative products often welcome conflict as an indication of energy and enthusiasm; a key skill for team leaders is to understand the advantages conflict brings to a team and to manage natural conflicts well for the greater good of the work of the team. If poorly managed, team members may shrink from conflict and enter a paralysis mode that severely limits team progress. Conflict is generally of two types—work related and personality related. Work-related conflicts can be resolved by sorting out the details and the confounding components of the conflict. Personality-related conflicts, however, are of a more challenging nature and might require a variety of interventions. The most regularly used is that of reflecting the effects of the poor behavior back to the perpetrator. The "Three Cs of Conflict Resolution" are the following:

1. Confidently address all instances of conflict when they arise.
2. Carefully investigate the nature and causes of the conflict.
3. Consistently deal with aberrant and detracting behaviors.

Importance of Trust in Teamwork

Trust among team members is deemed essential for the successful functioning of a team. If handled well, conflict resolution will engender trust among team members. Trust typically requires a span of time to build, yet it must also be constantly maintained. The team leader must never assume that once trust is built it will remain constant. Trust is the most fragile of all the relationships that make up a team's functioning component, and it requires constant vigilance to create an ongoing environment in which the members of the team find it possible to trust one another.

Applications to Healthcare

Multidisciplinary teams are an integral part of healthcare delivery. Anne Fleissig and colleagues note that such teams are widely accepted and are a major element of managing cancer care in the United Kingdom, among other countries. In their review, they note that such teams are resource intensive. Purported benefits include ensuring quality of care, both through pooling necessary expertise and ensuring that care is evidence based, and from enhancing continuity of care. Such coordination is clearly most important in more complex cases, where multiple specialists may be involved and where care must be delivered over an extended time period. Advantages to staff include mutual support and reinforcement and improved communication.

Many of the requisites for team functioning are the same as noted above: leadership, trust and mutual respect, information, and communication. Others—particularly funding and administrative support—are implicit but not always specified. However, there are other differences between business and healthcare teams. One issue is the stability and anticipated longevity of teams. Teams intended to accomplish a particular goal may differ from teams expected to function together on an ongoing basis. The staff of an emergency department, for example, is likely to fluctuate over time, and new

staff will need to be oriented and incorporated rapidly. Oncology teams may deal with individuals in other organizations. Decisions must often be made rapidly, and the consequences of poor decisions may be both immediate and severe.

One key rationale for teams in healthcare is the advantages of distributing responsibilities; Vimla Patel and colleagues argue that this, in turn, allows teams to process the massive amounts of information needed to treat patients. Teams thus allow collaboration and provide a safety net for catching errors. Successful multidisciplinary teams allow multiple domains of knowledge to be brought together.

The emphasis on collaborative, interdisciplinary teams has affected clinical education. More emphasis is being placed on group processes and communication skills, both among team members and between clinicians and patients. One clear distinction between clinical teams and those in business is the presence of hierarchical structures, which are reinforced by administrative and professional structures. This is particularly pronounced in teaching hospitals (e.g., the role of residents). Another distinction is the existence of clearly defined domains, each with its own professional knowledge and expertise. Another is the issue of when team members are physically available. In consequence, clinical teams tend to have well-defined demarcated task responsibilities, ideally organized to minimize both duplication of care and falling through the cracks. Communication tends to be focused on specific patient-related problems, with some attention to ensuring smooth team functioning. Patel and colleagues stress the importance of clear definition of roles and delineation of tasks and responsibilities, particularly when there is a potential for overlap. They stress the importance of balancing authority and autonomy and breaking down hierarchical structures when they interfere with smooth functioning. Communication and respect both appear to be key. These studies thus indicate that the general literature on team communication and collaboration considers them to be transferable skills, which should probably be incorporated into educational curricula.

To date, there is little good evidence as to the impact of teams, in part because it is difficult to isolate the impact of any particular factor. However, the limited evidence available does suggest that well-functioning teams do improve communication, patient outcomes, and staff well-being. As Louise Lemieux-Charles and Wendy McGuire have suggested, no single model of team effectiveness is likely to hold; effectiveness is likely to vary by factors such as the patient population, the care delivery setting, the team membership, and the team tasks. Considerable research is required to determine which elements contribute to effective team decision making.

Raisa Deber and Whitney Berta

Note: Some material has been used with permission from Berta, W. (2007). *Team building: The indispensable process of the 21st century* (FastPocket Series). Temecula, CA: Leading & Learning.

See also Choice Theories; Cognitive Psychology and Processes; Medical Errors and Errors in Healthcare Delivery; Shared Decision Making

Further Readings

Daft, R. (2004). *Organization theory and design* (8th ed.). Mason, OH: Thomson South-Western.

Fandt, P. M. (1994). *Management skills: Practice and experience.* St. Paul, MN: West Publishing.

Fleissig, A., Jenkins, V., Catt, S., & Fallowfield, L. (2006). Multidisciplinary teams in cancer care: Are they effective in the UK? *The Lancet Oncology, 7*(11), 935–943.

Institute of Medicine. (Ed.). (2001). *Crossing the quality chasm: A new health system for the 21st century.* Washington, DC: National Academies Press.

Lemieux-Charles, L., & McGuire, W. L. (2006). What do we know about health care team effectiveness? A review of the literature. *Medical Care Research and Review, 63*(3), 263–300.

Marcic, D. (1995). *Organizational behavior: Experiences and cases* (4th ed.). St. Paul, MN: West Publishing.

Patel, V. L., Cytryn, K. N., Shortliffe, E. H., & Safran, C. (2000). The collaborative health care team: The role of individual and group expertise. *Teaching and Learning in Medicine, 12*(3), 117–132.

Tuckman, B. (1965). Developmental sequence in small groups. *Psychological Bulletin,* 63, 384–399.

Technology Assessments

Technology assessment in medical decision making, or health technology assessment (HTA), is a

form of policy analysis that seeks to examine the broader impact of decisions surrounding medical technology in healthcare. Many definitions of HTA exist, but it is generally characterized as an instrument for decision making that bridges the world of research with the world of decision making. Organizations conducting HTA began to proliferate in the 1980s and 1990s, and HTA continues to be an active field internationally. HTA is conducted in very small and very large healthcare settings under a variety of models and using a variety of methods. Subsequently, HTA is used by a large number of those involved in medical decision making, including patients, clinicians, researchers, institutional managers, and public policy makers. HTA is largely practiced by private research firms, academic collaborations, private health plans, specialty societies, governments, and manufacturers of health technologies. HTAs can increase the legitimacy of the health policy decision-making process by giving decision makers the information they need to improve population health.

Defining Technology Assessment

Current Definitions of Health Technology Assessment

Several definitions of HTA are currently in common use, but they are generally consistent with definitions from the field of technology assessment, from which it originated. These definitions generally reflect analytic frameworks for policy analysis, including analysis of the probable political, economic, social, technical, ethical, legal, and environmental impacts of technology use. These approaches to examining multiple dimensions of a policy decision are also referred to as PEST (e.g., political, economic, social, technical), STEP, PESTLE, STEEP, or STEEPLE analyses.

The International Network of Agencies of Health Technology Assessment (INAHTA) defines HTA as "a multidisciplinary process of policy analysis that examines the medical, economic, social and ethical implications of the incremental value, diffusion and use of a medical technology in health care." A more recent definition, developed by the European Collaboration for Assessment of Health Interventions and Technology (ECHTA/ECAHI), a network of HTA producers, suggests that "health technology assessment (HTA) seeks to inform health policy makers by using the best scientific evidence on the medical, social, economic and ethical implications of investments in health care." They additionally suggest that health technology should be broadly defined as technology assessment, and it is not only concerned with machines and devices.

Disciplines Related to Health Technology Assessment

The academic origins of HTA can be traced to the expanding field of applied health research in the 1970s, including public health, evidence-based medicine, clinical epidemiology, health services research, health economics, and medical bioethics. Some have suggested that the field of HTA, with its health policy focus, is strictly confined to decisions regarding funding or reimbursement of health-technology-based interventions. Others have commented that HTA is not different from the broader field of health services research, although training in either discipline may involve different analytic approaches as a result of their different academic history and the widely adopted analytic approaches within them. Other related disciplines are clinical/biomedical engineering; health impact analysis; healthcare environmental health, risk, and safety management; and healthcare information technology analysis.

History of Technology Assessment in Healthcare

Origins

The origins of technology assessment trace back to the origins of the U.S. Office of Technology Assessment, which came into being in 1972. The original bill was introduced by Congressman Emilio Q. Daddario (D-Conn.), Chairman of the Subcommittee on Science, Research and Development in the House of Representatives, and it sought to provide legislation for a formal body that would provide "early indications of the probable beneficial and adverse impacts of the applications of technology." The OTA's mission expanded to healthcare after 1975, but the agency was closed in 1995.

Health Technology Assessment in the 1980s and 1990s

HTA institutions intended to support public policy decisions at the regional or national level began to proliferate in the late 1980s. These institutions generally emerged as academics interested in applied health research responded to government-driven needs for evaluation of high-profile emerging technologies and as a solution to the increasing scarcity of healthcare resources. Following the creation of an agency in France (CEDIT, 1982) designed to look at the impact of technology diffusion in hospitals, subsequent organizations emerged in Sweden (SBU, 1987) and Canada (AETMIS, 1988; CCOHTA, 1989). The reemergence of national technology assessment in the United States was seen with the development of the Agency for Health Care Policy and Research (AHCPR) in 1989 (now the Agency for Health Care Research and Quality).

In 1985, the International Society of Technology Assessment in Health Care (ISTAHC) appeared, a nonprofit organization established to encourage research, education, cooperation, and the exchange of information on the clinical, economic, and social implications of health technologies. The first issue of ISTAHC's journal, the *International Journal of Technology Assessment in Health Care*, appeared as well. Some organizational difficulties led to the replacement of ISTAHC with another society, Health Technology Assessment International (HTAi), in June 2003.

The 1990s

Additional agencies appeared worldwide in the 1990s in France (ANDEM, 1990), Canada (British Columbia Office of HTA, 1991), Spain (OSTEBA, 1991), Switzerland (SWISS-TA, 1992), Israel (ICTAHC, 1992), the United Kingdom (NCCHTA, 1993), the United States (VA-TAP, 1994), Finland (FinOHTA, 1995), and Australia (MSAC, 1998). A majority of these agencies continue to operate today.

Health Technology Assessment Today

More recently, HTA has emerged in newly industrialized countries such as Brazil, South Korea, Singapore, Malaysia, and South Africa and new European Union member states in transition (e.g., Hungary). International networks continue to emerge and include EUnetHTA, a European network of HTA agencies.

National and regional networks have also emerged, including the Swiss Network for HTA, the Italian Society for HTA, and the Canadian Health Technology Analysis Exchange. The emergence of HTA special interest groups in similar organizations, such as the International Society for Pharmacoeconomics and Outcomes Research (ISPOR) Health Technology Assessment Council, reflect the growth of this field.

Settings for Health Technology Assessment

The aim of the activity of HTA is to provide an input for healthcare decisions. As healthcare decision making can occur at the macrolevel (i.e., policy), mesolevel (i.e., institutional management), or microlevel (i.e., healthcare professional), and there is no universal manner in which healthcare is delivered and decisions are made, there is no single correct setting for the conduct of HTA. The structure and processes of any HTA institution must fit the purpose of the decision-making processes that they are supporting. As such, HTA can be seen in very small (e.g., hospital or hospital department) or very large (e.g., regional or national) settings. An HTA organization may be located inside a government or may be completely independent from government. HTA organizations may focus on a narrow set of health technologies (such as drugs) or decisions (such as adoption decisions) or may try to address a broader set of technologies and decisions. HTA organizations may also focus on a narrower set of dimensions of analysis, such as clinical and economic, or a much wider set, including the psychosocial, ethical, and legal implications of decisions surrounding health technologies.

HTA is largely practiced by private research firms, hospital departments, academic collaborations, private health plans, specialty societies, governments, and manufacturers of health technologies.

Hospital-Based Health Technology Assessment

Hospital-based HTA is becoming a more popular setting for the conduct of HTA activities. A hospital-based HTA unit in Montreal reported

that hospital policy was developed from 16 out of 17 HTA findings and that hospital resources were more effectively used, resulting in budgetary savings of Can$3.2 million annually.

Drug Reimbursement

With the advent of public and private plan drug formularies, the application of HTA to drug adoption and reimbursement decisions has also become widespread. Larger national bodies that conduct assessment and provide recommendations to government include Australia's Pharmaceutical Benefits Advisory Committee, Canada's Common Drug Review, Scotland's Scottish Medicines Consortium, the U.S. Medicare Evidence Development and Coverage Advisory Committee (MedCAC), and the U.K. National Institute for Health and Clinical Excellence. More recent examples include Taiwan's Center for Drug Evaluation.

How Is Health Technology Assessment Performed?

Some HTA organizations conduct assessment only, providing relevant findings, while others additionally *appraise* the findings and provide advice or recommendations. The methods and approaches used to assess and appraise information may vary but usually reflect the policy-driven need for reliable information and advice.

The conduct of HTA generally places an emphasis on transparency, comprehensiveness, and rigor. The systematic review has been widely adopted as a tool to examine the technologic (health) impact of any decision as the conduct of systematic review is in itself a transparent and rigorous approach to the identification and selection of appropriate information. Some HTA organizations may rely on publicly available information, while others may work more closely with manufacturers or governments and assess confidential information.

As the quantity and quality of information required to answer specific policy questions are often lacking, many HTA organizations rely on mathematical modeling to further inform policy decisions. Mathematical modeling can assist analysts in estimating the comparative effectiveness of available interventions when direct evidence does not exist and in projecting the long-term consequences of a technology-based decision when data are unavailable.

Economic evaluation in technology assessment usually involves the analysis of health as a commodity or system output and seeks to analyze how health can be optimally exchanged with available healthcare resources within a specific healthcare context. The central role of economics in HTA has led to the recognition that consistency in the approaches to economic evaluation is required to promote consistency in decision making. As such, country-specific guidance for those conducting economic evaluation has widely emerged. ISPOR has a comparative table and database of existing guidelines from specific countries.

Because of the forward-looking nature of HTA and the need to consider multiple sources of evidence for informed decision making, Bayesian methods are becoming more widely adopted to examine the likelihood of future health benefits from technology. Economic evaluation methods have also adopted probablistic methods to examine decision uncertainty when statistical uncertainty is present. More recently, some HTA organizations have used the value of information analysis to compare potential or foregone opportunity costs with the costs of reducing uncertainty through further research.

Like methods to examine the clinical and economic dimensions of health policy, appropriate methods to conduct analyses of additional dimensions of a policy decision, such as political or environmental analyses, are generally borrowed from current acceptable approaches in their respective disciplines.

How Is Health Technology Assessment Used?

Findings from an HTA can be disseminated to appropriate recommendatory or decision-making bodies using appropriate knowledge transfer approaches. HTA organizations that also appraise findings and provide guidance may do so through well-defined and highly collaborative relationships or by using much more passive approaches, giving advice to various levels of decision makers in the hope of influencing decisions. HTA organizations must decide to whom findings and advice are to be communicated and in what format.

Recommendations and guidance from HTA findings usually involve a deliberative process or

method. These deliberative methods may involve deliberation among representatives of all those affected by a decision or deliberation among subject or content experts who make technical recommendations to decision makers. Since decisions involve more than scientific evidence, an emphasis on properly capturing the values of those affected by health policy has seen greater involvement of patients and consumers in deliberative processes. Frameworks for codifying evidence and evaluating recommendatory processes have been developed. One such framework is the "accountability for reasonableness," which suggests that evidence-based recommendations should be transparent and clearly defined.

Don Husereau

See also Costs, Opportunity; Decisions Faced by Nongovernment Payers of Healthcare: Managed Care; Economics, Health Economics; Evidence Synthesis; Government Perspective, General Healthcare; Government Perspective, Informed Policy Choice; Government Perspective, Public Health Issues; Informed Decision Making; International Differences in Healthcare Systems; Medicaid; Medical Decisions and Ethics in the Military Context; Medicare; Meta-Analysis and Literature Review; Moral Choice and Public Policy; Pharmacoeconomics; Reference Case; Value-Based Insurance Design; Willingness to Pay

Further Readings

Banta, D. (2003). The development of health technology assessment. *Health Policy, 63*(2), 121–132.

Battista, R. N. (1996). Towards a paradigm for technology assessment. In M. Peckham & R. Smith (Eds.), *The scientific basis of health services* (pp. 11–18). London: BMJ.

Daniels, N. (2000). Accountability for reasonableness. *British Medical Journal, 321*(7272), 1300–1301.

Health Technology Assessment International: http://www.htai.org

International Network of Agencies for Health Technology Assessment. (2008). *HTA resources*. Retrieved January 2, 2009, from http://inahta.episerverhotell.net/HTA

Lomas, J., Culyer, T., McCutcheon, C., McAuley, L., & Law, S. (2005). *Conceptualizing and combining evidence for health system guidance*. Ottawa, ON, Canada: Canadian Health Services Research Foundation.

Terminating Treatment, Physician Perspective

The involuntary dissolution of an established physician–patient relationship can be one of the more difficult medical decisions a clinician must make. Cultivating a relationship and gaining the trust of patients is a fundamental skill in clinical medicine. Without trust, patients will not provide the necessary information for an accurate diagnosis and are less likely to comply with recommended treatment. The process of gaining trust and establishing a relationship is more overt in primary-care specialties, where it is often necessary to obtain sensitive historical details for clinical decision making. However, the establishment of trust is no less necessary in surgical specialties, where the consequences of clinical decision making are more immediate and where the risks of poor decision making are often higher.

In nonprofessional relationships, trust is acknowledged to require equal participation by both parties. Trust is both earned and given in equal measure by both participants. This is not the general view of clinicians, who understand that there is an unequal power balance between physician and patient and who therefore feel disproportionately obligated to earn the trust of their patients in order to provide good care. During training, it is understood that this empathetic skill will be required of trainees even when patients are personally objectionable or even hostile toward the physician. Physician trainees are considered to bear the majority of the responsibility for establishing a mutually trusting and beneficial therapeutic relationship. As a result of this professional socialization, a physician–patient relationship that has deteriorated to the point of termination can be experienced as a sign of failure on the part of the clinician. This feeling can contribute significantly to the emotional difficulty of the decision to terminate treatment.

Having said this, it is widely acknowledged that not all physicians will be able to cultivate mutual trust with all patients and that a small percentage of physician–patient relationships will need to be terminated. The difficult medical decision comes in determining which relationships

have reached this threshold and when they have reached it. There are three accepted reasons for terminating a physician–patient relationship: (1) noncompliance with treatment, (2) disruptive behavior, and (3) nonpayment of bills.

Noncompliance With Treatment

It is generally understood that irresponsible or unhealthy behavior does not release a physician from the duty to provide care. While patients are expected to be good stewards of their health, noncompliance and poor patient decision making are so widespread that they are viewed as a part of the landscape of care. Indeed, it is broadly recognized that physicians and nurses are some of the least compliant patients. Clinicians also recognize that noncompliance is not always entirely the fault of the patient, for many reasons: (a) Patients may not understand instructions, (b) they may understand instructions but fail to understand the consequences of noncompliance, (c) they may lack the financial resources to follow through, (d) they may be too emotionally overwhelmed with the responsibility of following treatment recommendations, and (e) they may be distracted by other priorities that are temporarily competing with their time and energy.

The reasons for noncompliance are often temporary or relatively easily overcome, and in keeping with the empathetic socialization process discussed above, some physicians will go to extreme lengths to help patients become more compliant. Generally, clinicians are tolerant and understand that the patient's social, emotional, and financial situations will change and compliance can be expected to wax and wane accordingly. Isolated noncompliance is almost never the reason for termination of treatment and is usually combined with one of the other two reasons to provide justification for ending a physician–patient relationship.

Disruptive Behavior

Physicians are not obligated to provide futile treatment. If a patient relationship has deteriorated to the point where the physician feels that further interaction is extremely unlikely to provide benefit to the patient, then the treatment has become futile and the physician is no longer obligated to provide care. It is important to keep the focus on the benefit to the patient when considering whether disruptive behavior meets the threshold of termination of treatment. Countertransference is a psychiatric process whereby a physician transfers the abusive behavior of the patient back to the patient in the form of dislike or outright hatred. Conscientious clinicians are aware of this tendency and will do their best to ameliorate it. As is the case with noncompliance, there are reasons for disruptive behavior that may be temporary: (a) The patient may have an undiagnosed substance abuse problem, (b) the patient may have poor coping skills and be acting out as a result of being overwhelmed by the stress of his or her illness, and (c) the patient may be in extreme pain. Substance abuse disorders can be diagnosed and rehabilitated, coping skills can be strengthened with counseling, and pain can be treated with analgesics. However, all three barriers to treatment must first be uncovered by a clinician who is willing to overlook dysfunctional behavior to get at the reason behind it.

Some patients may not be aware that their behavior is disruptive, because it is considered acceptable in their nonprofessional relationships. Therefore a clinician is obligated to inform the patient that his or her behavior is unacceptable and is hindering care. The next step is to offer help concerning more constructive ways of interacting. This gives the patient the opportunity to change his or her behavior in an effort to demonstrate that he or she wishes to preserve the physician–patient relationship. When repeated attempts at help are rejected and disruptive behavior continues or when there is a concern about physical violence, physicians feel justified in ending a relationship based solely on disruptive behavior. However, disruptive behavior almost always occurs hand in hand with noncompliance, in which case the threshold for termination is lowered. A patient who is disruptive but at least making an effort to get better is still deriving benefit from the physician–patient relationship. A patient who is both disruptive and noncompliant is often viewed as a lost cause.

Nonpayment of Bills

In countries with government-funded healthcare systems, nonpayment of bills is not an issue. However, in the United States, where a large percentage of healthcare is funded by commercial insurance and many physician practices still

function as small businesses, nonpayment is still considered a legitimate reason for termination. Nevertheless, because of their own interaction with an increasingly complex and burdensome payment system, American physicians often have a surprising degree of sympathy for patients who cannot pay their bills. Physicians working in publicly owned American hospitals have an obligation to treat regardless of a patient's ability to pay; and even in private hospitals, the responsibility for admitting uninsured patients is rotated among physicians. Performance of "unassigned call" is usually a condition of maintenance of privileges at private U.S. hospitals.

The obligation to care for patients who present with an urgent threat to life or function and require admission to the hospital does not extend to the outpatient setting. Outpatient care in the United States is seen as discretionary and may be refused by an American physician for any reason, including inability to pay. If an individual physician has an ongoing relationship with a patient who loses his or her insurance or falls on hard times, the physician may be willing to continue providing outpatient care for token payments. In the U.S. system, nonpayment of bills usually accompanies one of the other two reasons for termination, and together, the threshold for the decision to terminate is lowered.

In large private U.S. systems of care where the individual physician is not in direct control of scheduling patients, a new phenomenon of "financial clearance" has arisen that functions as a de facto termination of care on grounds of nonpayment. Financial clearance is a process whereby payment must first be assured by a third-party payer or the patients themselves before an appointment can be made with a physician. The patient without means is thereby excluded from seeing a physician before the nonpayment has actually occurred and without the direct consent of the treating physician. Financial clearance is an increasingly frequent "inadvertent" cause of the termination of a physician–patient relationship for nonpayment of bills.

Mechanics of Termination of a Physician–Patient Relationship

Once a decision has been made to terminate treatment, the process of notifying the patient is mandated by the legal system and/or the patient's insurance company. Third-party payers typically have policies for termination of care by physicians in their panel, and this may affect a physician's response to the patient. Letters are always the preferred method of communication because they can be documented accurately. A letter should be sent by certified mail to ensure that it has been received by the patient. The letter should contain (a) the reason for termination, (b) an offer of alternative care to the patient, (c) ample opportunity to secure alternative care, and (d) an offer of assistance in transferring medical records to a new provider.

The reason for termination and the offer of assistance in transferring records are not required by law and can be omitted, but the other two components are required in the United States to avoid a legal charge of patient abandonment. The offer of alternative care should not be a referral to specific providers but should instead provide the address and phone number of the patient's insurance company or the local medical society, which could, in turn, provide the patient with a list of acceptable alternative providers. Ample opportunity to secure alternative care usually constitutes a window of 30 days from the date of notice, during which the terminating physician is required to continue to provide care for the patient. However, in cases where very subspecialized care is required by the patient and few providers are available in a given geographic area, this period may be extended to 60 or 90 days. U.S. government payers may have much stricter policies regarding involuntary termination of a patient. Medicare and Medicaid can require actions such as offering of anger management classes, intervention by a social worker, or a "second chance," meaning transfer of care to another provider within the physician's group or system. These requirements are usually unique to the individual state in which the federal program is administered.

The decision to terminate treatment and involuntarily dissolve the physician–patient relationship can be one of the more difficult decisions a physician must make in his or her clinical practice. The difficulty of this medical decision arises not from its logical complexity but from the difficulties of successfully navigating any physician–patient relationship and from the emotions it generates as a

result of the professional socialization process. Establishment of an empathetic, mutually beneficial relationship with a patient is seen by many clinicians as every bit as important as making the correct diagnosis or selecting the most appropriate treatment. Failure in this realm of medical decision making can be equally distressing to both physician and patient.

Robert Patrick

See also Decision Making and Affect; Models of Physician–Patient Relationship

Further Readings

Capozzi, J. D., Rhodes, R., & Gantsoudes, G. (2008). Terminating the physician-patient relationship. *Journal of Bone and Joint Surgery* (American Volume), *90*(1), 208–210.

Torres, A., Wagner, R., & Proper, S. (1994). Terminating the physician-patient relationship. *Journal of Dermatologic Surgery and Oncology*, *20*(2), 144–147.

Willis, D., & Zerr, A. (2005). Terminating a patient: Is it time to part ways? *Family Practice Management*, *12*(8), 34–38.

Test-Treatment Threshold

How can physicians decide whether a particular treatment could provide benefits to a patient rather than harm by unnecessary treatment? If a physician has 100% confidence that a particular patient has a disease (100% pretest probability), the physician should treat the patient. In contrast, a physician would not treat the patient if he or she had 100% confidence that the patient did not have the disease (0% pretest probability). Unfortunately, the pretest probability is not always 0% or 100% but usually lies somewhere in between these numbers, which is usually called a gray area. Taking potential adverse events into consideration, what estimated probability of disease do physicians need to decide whether or not to treat?

Threshold analysis is a mathematical approach to determine the answer. The treatment threshold (the probability of disease above which the patient should be treated) can be expressed in terms of the benefit-risk ratio: Treatment threshold = $R/(B + R)$. For example, let us consider a patient who may have an acute disease that has 10% risk of short-term mortality. If a treatment could reduce the risk of short-term mortality from 10% to 5%, the benefit (B) of the treatment is 5%. In contrast, if the treatment can cause a severe adverse event, it would result in a 1% increased risk of short-term mortality; that is, the risk (R) is 1%. Thus, the treatment threshold can be calculated as .01/(.05 + .01) = .1666…(16.7%). Therefore, the treatment could provide benefit if the probability of disease is equal to or greater than 16.7%.

Role of the Diagnostic Test

Whereas the above mathematical approach could provide a threshold for decision making, it is rare for many individuals to select either of the options without any hesitation. Indeed, it is always difficult to answer "yes" or "no" to such questions. A desire for more information for making better decisions tends to be the natural response to such a situation.

In clinical decision making, diagnostic tests usually provide the additional information required by clinicians. The diagnostic process can be defined as a process to increase or decrease the probability of the target disease until it is possible to rule in or rule out specific diseases. The role of diagnostic tests is to increase or decrease the probability of selecting the target disease in clinical decision making.

Figure 1 demonstrates the concept of diagnosis. Physicians can select "Treat" if they have a sufficient number of reasons to estimate that the probability of the target disease is high enough to start a particular treatment. In other words, benefits from the treatment can be assumed to be higher than the potential risk of the treatment. This threshold is referred to as the *test-treatment threshold*, or *test-treat threshold*. In contrast, the physician can confidently choose "Don't treat" if there are a sufficient number of reasons that the probability is low enough to rule out the disease. This threshold is referred to as the *no treatment-test threshold*, or *no treat-test threshold*.

Diagnostic tests are useful as long as they have sufficient information to increase the probability (i.e., posttest probability) above the test-treatment threshold (i.e., rule in) or to decrease it below the

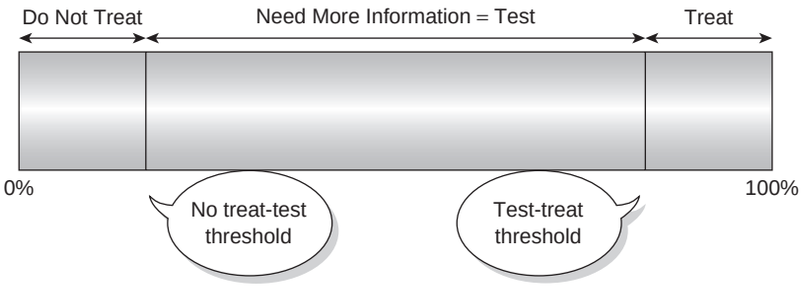


Figure 1 Concept of diagnosis

no treatment-test probability (i.e., rule out). Figure 2 shows the role of diagnostic tests in medical decision making. Assuming that physicians estimate the probability of the target disease as A (%), which is between the no treatment-test and test-treatment thresholds, the diagnostic test is useful if the posttest probability can be increased to C but not to D, which is still below the test-treatment threshold. If the probability is still below the threshold, further diagnostic tests will be needed to increase the posttest probability appropriately. Similarly, the diagnostic test should have sufficient information to decrease the posttest probability to B if the results are negative.

The amount of information required to increase or decrease the probability of the target disease can be quantified by a positive-negative likelihood ratio of each diagnostic test for a particular target disease. A combination of multiple tests can be used if no single test can provide enough information or if the cost of a specific test is too expensive.

No Treatment-Test and Test-Treatment Thresholds

How can a physician determine the two thresholds: (1) no treatment-test threshold and (2) test-treatment threshold? Figure 3 shows a geometric approach to describe these two thresholds.

Let us assume that there is a diagnostic test without adverse effects. If the pretest probability of the target disease is 0, all patients treated would be harmed by the treatment. The expected harm from performing the test instead of not treating can be calculated as Harm × False-positive ratio (FPR). This is the point at which the "test" line intersects the *y*-axis in Figure 3. In contrast, by performing the test, the patients with true-negative test results are spared the harm of treatment. Thus, the expected harm avoided by performing the test instead of treating equals Harm × True-negative ratio (TNR), which is shown in Figure 3 as the vertical distance between the "test line" and the "treat" line along the *y*-axis.

Similarly, when the pretest probability of the target disease is equal to 1, all patients with false-negative test results would miss out on the benefit of treatment (Benefit × False-negative rate [FNR]), which is shown as the height between the "treat" and "test" lines on the right side of Figure 3. Also, by performing the test, patients with true-positive results are given the benefits of treatment (i.e., Benefit × True-positive rate [TPR]).

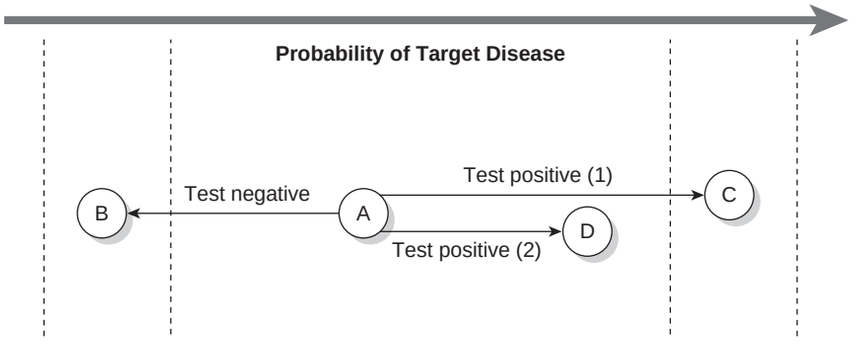


Figure 2 Role of diagnostic test

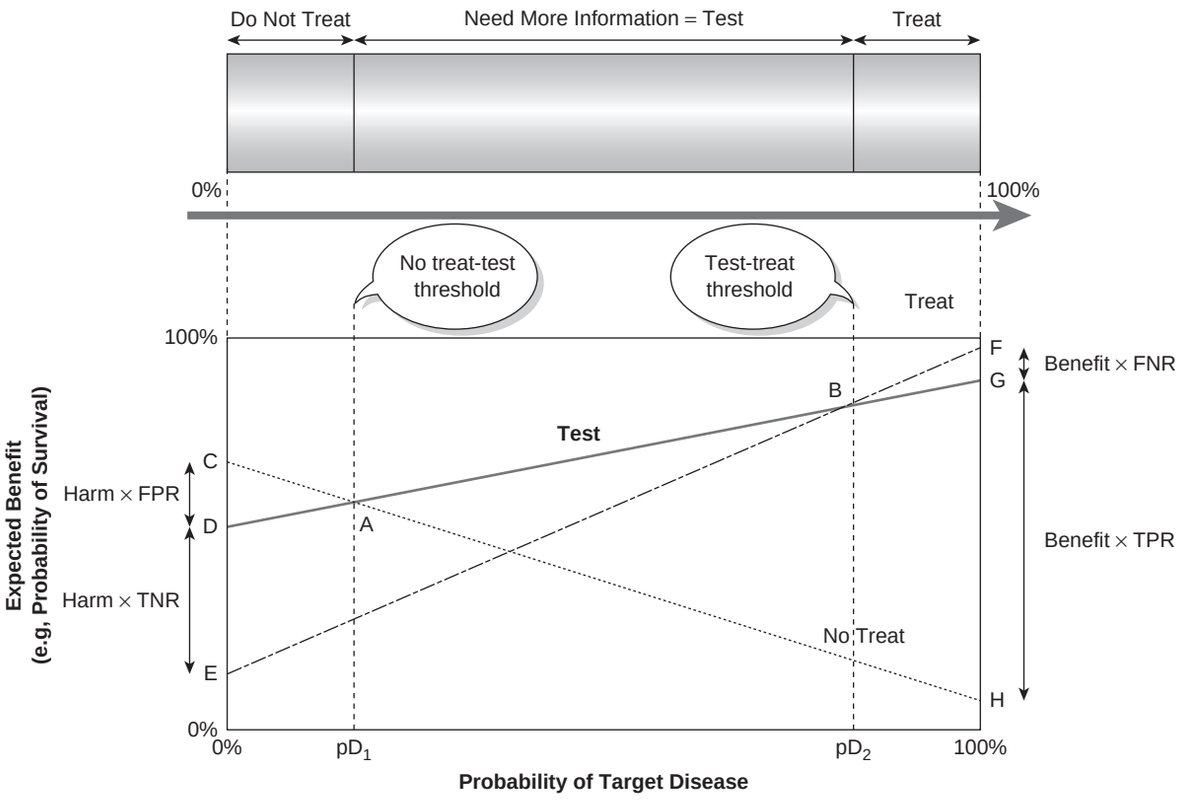


Figure 3 Test-treat threshold

A mathematical formula to calculate these two thresholds can be derived as follows.

To identify the no treatment-test threshold, we should focus on two similar triangles that are made by "No treat" and "Test," ΔACD and ΔAHG. In these triangles, the ratio between pD_1 and $(1 - pD_1)$ equals the ratio between CD and HG. Therefore, we derive

$$pD_1:(1 - pD_1) = (\text{Harm} \times \text{FPR}):(\text{Benefit} \times \text{TPR}),$$

$$pD_1 \times \text{Benefit} \times \text{TPR} = (1 - pD_1) \times (\text{Harm} \times \text{FPR}),$$

$$pD_1 \times ([\text{Benefit} \times \text{TPR}] + [\text{Harm} \times \text{FPR}]) = \text{Harm} \times \text{FPR}.$$

Therefore, no treat-test threshold (pD_1) $= (\text{Harm} \times \text{FPR})/([\text{Benefit} \times \text{TPR}] + [\text{Harm} \times \text{FPR}])$.

Similarly, focusing on two triangles made by "Test" and "Treat" (i.e., ΔBDE and ΔBGF), we derive

$$pD_2:(1 - pD_2) = \text{Harm} \times \text{TNR}:\text{Benefit} \times \text{FNR}.$$

Test-treatment threshold (pD_2) can be calculated as

$$(\text{Harm} \times \text{TNR})/(\text{Harm} \times \text{TNR} + \text{Benefit} \times \text{FNR}).$$

Noriaki Aoki, Michi Sakai, and Sachiko Ohta

See also Expected Value of Perfect Information; Expected Value of Sample Information, Net Benefit of Sampling

Further Readings

Hunink, M., Glasziou, P., & Weinstein, M. (2001). *Decision making in health and medicine: Integrating evidence and values*. Cambridge, UK: Cambridge University Press.

Threshold Technique

The Threshold Technique measures individuals' attitudes toward a single key attribute of a particular "target" therapy. The procedure involves repeatedly varying the levels of the key attribute of primary interest and, with each variation, asking the respondent to choose between the target therapy and an alternate reference therapy.

The key attribute that is varied depends on the study purpose. For example, the attribute could be chance (e.g., the probabilities of side effects or benefits), time (e.g., time waiting for therapy or life expectancy), or distance to travel (e.g., for access to care). Compared with other methods for evaluating therapies, it is close to the Willingness-to-Pay method, in which cost is the key attribute that is varied.

The technique can be appropriately adapted to a wide range of clinical dilemmas in medical decision making. For example, one version can reveal patients' minimal required reduction in the long-term risk of a heart attack that they would want from a cholesterol-lowering agent (the target) before considering it worthwhile compared with lifestyle management alone (the reference). In another example, consider the choice between a current analgesic (the reference) and a new medication (the target) that offers potentially greater pain relief but carries a higher risk of gastric bleed. Different versions of the technique could be designed to reveal either the minimal amount of pain relief that patients would require or the maximal level of risk for gastric bleed that patients would accept before considering the new drug to be preferable to their current medication.

Therefore, the technique is very flexible. However, because it focuses on a single key attribute, it can appropriately address only narrowly defined and context-dependent research questions. Conceptually and procedurally, the technique is very different from other methods for evaluating therapies—including formal Decision Analysis, Conjoint Analysis, the Analytic Hierarchy Process, the Balance Technique, and the Leaning Scale, which work with multiple attributes in a more holistic manner and are designed for different research purposes.

Procedure

The technique is described below by using a particular clinical context and a particular key attribute—the benefit probability—to illustrate its general approach. Although the technique could be carried out using interactive electronic media, the entire procedure is outlined here as it would occur in an in-person interview.

The interviewer uses a preconstructed, study-specific toolkit, consisting of information cards, probability wheels, and sliding scales. Together, the interviewer and the respondent work through three interview phases.

Setting the Stage

First, the interviewer places a Condition Overview Card on the table in front of the respondent. This card outlines the relevant clinical condition's probable causes, signs and symptoms, and natural prognosis. The interviewer reviews this information with the respondent.

When the respondent understands this information, the interviewer explains that there are two relevant therapeutic options, Treatments A and B. The interviewer explains that he or she will systematically present the respondent with cumulative pairs of Treatment Information Cards with "bits" of information about the two options, that they will review each pair together, and that they will not proceed to the next pair until each piece of information is understood by the respondent.

The first pair describes the two treatment protocols. They are arranged side-by-side in front of the respondent, immediately below the Condition Overview Card, and reviewed. When the respondent understands the protocols, the next pair of cards is presented. These describe each treatment's side effects, including the evidence-based estimated probabilities of their occurrence. These cards are arranged side by side, just below the treatment protocol cards, and reviewed. When the side effect information is understood, the next pair of cards is presented. They describe each treatment's anticipated benefits as well as their associated probabilities. These cards are arranged side by side, just below the anticipated benefits cards, and also reviewed. Throughout this stage-setting phase, devices such as moveable probability wheels are used to foster comprehension of the probabilities.

Identifying the Initial Choice

At this point, the Treatment Information Cards are arranged in parallel columns that permit clear across-therapy comparisons. In effect, the respondent sees a map delineating two process-and-outcome paths describing what the treatments would entail, their possible outcomes, and the probabilities of encountering those outcomes. The respondent is invited to indicate which treatment option would be preferable if he or she were actually making this therapeutic choice.

Finding the Threshold

Once the respondent has indicated his or her initial choice, the investigator could proceed in different ways; the particular version of the technique that is used depends on two major considerations.

The First Major Consideration

The study's research objective dictates the answers to two design questions: (1) "Which are the target and the reference therapeutic options?" and (2) "What is the single key attribute that will be systematically varied?"

Suppose our respondents are patients considering lifestyle management alone (Treatment A), which has an estimated 10-year risk of myocardial infarction equal to 15%, versus a new, cholesterol-lowering medication (Treatment B), which has an estimated 10-year risk of myocardial infarction equal to 10% but also has some side effects. The objective is to find the minimal required reduction in the 10-year risk of myocardial infarction that these patients would want from the medication before considering it worthwhile relative to lifestyle management alone. The medication is the target option, lifestyle management is the reference option, and the key attribute is the 10-year risk of myocardial infarction while taking the medication.

The Second Major Consideration

The respondent's identified initial choice, in conjunction with the research objective, dictates the answer to the design question "In which direction should the key attribute be systematically varied?"

Suppose the respondent initially chose Treatment A—lifestyle management alone. Then, the investigator would systematically decrease the risk of myocardial infarction while taking the medication (from 10% to 9% to 8%, etc.) until the respondent switches his or her stated preference from lifestyle management alone to the medication.

On the other hand, suppose the respondent initially chose Treatment B—the medication. Then, the investigator would systematically increase the risk of myocardial infarction while taking the medication (from 10% to 11% to 12%, etc.) until the respondent switches his or her stated preference from medication to lifestyle management alone.

In either case, all respondents are weighing their aspiration for a reduced 10-year risk of myocardial infarction relative to their aversion to the medication's side effects. Regardless of the individual's initial choice, the subsequent assessment steps move up or down a common underlying attitudinal scale; here, that scale is the minimal risk reduction required to render the medication acceptable to the respondent.

Conceptual Basis

Assessing Multi-Attribute Utility Functions

Ralph Keeney extended the axioms and procedures used to elicit an individual's utility function for a particular attribute, such as cost, to the elicitation of utility functions for multi-attribute entities. The initial step in assessing multi-attribute utility functions involves a set of preliminary tasks in which the individual makes different choices between entities and lotteries.

These tasks reveal whether the individual's underlying multi-attribute function satisfies the assumptions of utility independence, mutual utility independence, and additive independence. Verification of the assumptions, in turn, reveals whether the underlying multi-attribute utility function is multiplicative or additive in form. The analyst exploits this insight in subsequent elicitation tasks that quantify the individual's full multi-attribute utility function.

The Threshold Technique basically borrows the assumption-verification strategies used to initially characterize the form of a multi-attribute utility function and applies them in a simplified single-attribute manner to the problem of revealing the respondent's attitude toward the key attribute of primary research interest.

Decisional Conflict

Decisional conflict is induced when there are simultaneous opposing tendencies to accept and reject a course of action. Extremely low or high levels could be dysfunctional, in that they can lead to defective information search/processing strategies and ineffective decision making. On the other hand, some degree of decisional conflict could be functional if it encourages effective information search and processing strategies, greater clarity of one's own preference structure, and a comfortable level of motivation to resolve a decision dilemma.

The Threshold Technique begins with a descriptive, evidence-based overview of the condition and two relevant treatment options. For individuals without a clear prior preference, this presentation of the options could itself induce some decisional conflict as they consider the choice and make an initial selection. In the next steps, however, the decision problem is altered by changing the level of a key attribute so that the initial choice becomes less and less favorable. These changes could induce functional degrees of decisional conflict that motivate the decision maker to focus on the relevant information and arrive at a new choice that explicitly—albeit partially—reveals his or her underlying preference structure. It is important to note that this is a partial revelation; it is only in terms of the key attribute, and it is only relative to the other option.

Applications and Outcomes

Research Applications

The technique has been adapted to assess attitudes toward the different attributes inherent in a wide variety of treatments, including chemotherapy, radiation therapy, surgery, palliative care, and medications for hypercholesterolemia, hypertension, anticoagulant therapy, Crohn's disease, and osteoarthritis. These applications have been motivated by curiosity about patients' attitudes not only toward treatment choice but also toward the design of and entry into clinical trials, the resolution of ethical issues, the construction of practice guidelines, and the development of health policy.

Measurement Outcomes

In each application, the technique was designed to suit the studies' unique research purposes, which determine the target option, the risk/benefit attribute that will be probabilistically altered, and the direction in which that alteration will proceed. Accordingly, in each application, the underlying attitudinal scale is idiosyncratic to the original research problem. Therefore, it is inappropriate to apply the technique to problems that require an absolute preference scale, permitting across-disease comparisons. The technique should not be regarded as an alternative to the Standard Gamble, nor should its results be considered to represent decision-analytic expected utilities for the treatment under consideration. However, when the research problem requires us only to assess individuals' strength of preference for A relative to B, within the confines of the particular clinical context, the disease-dependent nature of the technique is not a concern.

Whether the technique generates internally logical, consistent, and stable results is more important. Test-retest reliability coefficients ranging from the high .70s to the high .90s have been observed, and the method can reveal logically consistent subgroups. This implies that when the underlying preferential attitudes are not expected to be labile, patients report reasonably stable and valid switch points. However, since the relative preference scales are uniquely determined by the particular trade-offs in each decision problem, we cannot talk in terms of *the* psychometric properties of the technique, as if they were characteristics that carry across all applications.

Hilary A. Llewellyn-Thomas

See also Conjoint Analysis; Decisional Conflict; Decision Trees, Construction; Decision Trees, Evaluation; Multi-Attribute Utility Theory; Utility Assessment Techniques; Willingness to Pay

Further Readings

Brundage, M. D., Davidson, J. R., Mackillop, W. J., Feldman-Stewart, D., & Groome, P. (1998). Using a treatment trade-off method to elicit preferences for the treatment of locally advanced non-small cell lung cancer. *Medical Decision Making, 18,* 256–267.

Kennedy, E. D., To, T., Steinhart, A. H., Detsky, A. S., Llewellyn-Thomas, H., & McLeod, R. S. (2000). Is

the probability tradeoff a reliable and valid measure for use in young populations with chronic disease? *Medical Decision Making, 20,* 506.

Kopec, J. A., Richardson, C. G., Llewellyn-Thomas, H., Klinkhoff, A., Carswell, A., & Chalmers, A. (2007). A probabilistic threshold technique showed that patients' preferences for specific tradeoffs between pain relief and each side effect of treatment in osteoarthritis varied. *Journal of Clinical Epidemiology, 60,* 929–938.

Llewellyn-Thomas, H., Arshinoff, R., Bell, M., Williams, J. I., & Naylor, C. D. (1998). In the queue for total joint replacement: Patients' perspectives on waiting times. *Journal of Evaluation in Clinical Practice, 4,* 63–74.

Llewellyn-Thomas, H., McGreal, M. J., Thiel, E. C., Fine, S., & Erlichman, C. (1991). Patients' willingness to enter clinical trials: Measuring the association with perceived benefit and preference for decision participation. *Social Science & Medicine, 32,* 35–42.

Llewellyn-Thomas, H., Paterson, J. M., Carter, J. A., Basinski, A., Myers, M. G., Hardacre, G. D., et al. (2002). Primary prevention drug therapy: Can it meet patients' demands for reduced risk? *Medical Decision Making, 22,* 326–339.

Llewellyn-Thomas, H., Thiel, E. C., Paterson, J. M., & Naylor, C. D. (1999). In the queue for coronary artery bypass grafting: Patients' perceptions of risk and "maximal acceptable waiting time." *Journal of Health Services Research & Policy, 4,* 65–72.

Llewellyn-Thomas, H., Williams, J. I., Levy, L., & Naylor, C. D. (1996). Benign prostatic hyperplasia (BPH): Using a trade-off technique to assess patients' treatment preferences. *Medical Decision Making, 16,* 162–172.

Naylor, C. D., & Llewellyn-Thomas, H. (1994). Can there be a more patient-centered approach to determining clinically important effect sizes for randomized treatment trials? *Journal of Clinical Epidemiology, 47,* 787–795.

Palda, V. A., Llewellyn-Thomas, H., MacKenzie, R. G., Pritchard, K. I., & Naylor, C. D. (1997). Breast cancer patients' attitudes about rationing postlumpectomy radiation therapy: Applicability of the probability trade-off method to policy-making. *Journal of Clinical Oncology, 15,* 3192–3200.

Richardson, C. G., Chalmers, A., Llewellyn-Thomas, H. A., Klinkhoff, A., Carswell, A., & Kopec, J. A. (2007). Pain relief in osteoarthritis: Patients' willingness to risk medication-induced gastrointestinal, cardiovascular, and cerebrovascular complications. *Journal of Rheumatology, 34,* 1569–1575.

Time Horizon

Time horizon is the duration of time being considered when evaluating clinical strategies. When comparing multiple strategies for the same scenario, the time horizon over which the outcomes are evaluated must be the same across all potential strategies to make a fair and appropriate comparison.

Some authors consider two terms: *time frame* and *analytic horizon*. The distinction is made as follows: *Time frame* refers to the specified period in which the strategies are actually applied; the *analytic horizon* is the period over which the outcomes are considered as a result of the strategies considered. In this framework, the analytic horizon can be longer than the time frame because the costs and benefits of an intervention may continue long after the intervention is completed. This is particularly true in the case of preventive interventions (e.g., immunization), in which case the benefits of such activities will occur later in a person's lifetime.

What should be the time horizon? Some have stated that no particular time horizon can be generally recommended since the time horizon for each study should be matched to the intervention and outcomes being studied. The time horizon should be explicitly stated at the beginning of the study.

There are situations in which a very narrow time horizon should be used: When making decisions about acute or short-term events, it may be appropriate to use a time horizon of a few hours or days. For chronic or long-term events, a longer time horizon would be appropriate. The time horizon should encompass the entire decision process being modeled. All events and their resulting effects, including clinical and economic, should be considered within the time horizon when a clinical problem or situation is being modeled.

Decision and cost-effectiveness analyses use time horizons that vary from hours to a lifetime. Changing the time horizon in an analysis can greatly alter the apparent differences in life expectancy among various alternatives. For cost-effectiveness analysis, the Panel on Cost-Effectiveness in Health and Medicine recommends that a lifetime time horizon be used. Analyses of models with lifetime horizons usually employ Markov models, which divide the lifetime horizon into equal periods or cycles.

Of note, if the evaluation is to be over multiple years and the analysis performed is a cost-effectiveness analysis, then the present value of costs and outcomes should be discounted to the base year of the analysis.

Scott B. Cantor and Lesley-Ann N. Miller

See also Cost-Effectiveness Analysis; Costs, Direct Versus Indirect; Costs, Fixed Versus Variable; Costs, Incremental; Costs, Opportunity; Costs, Out-of-Pocket; Costs, Semifixed Versus Semivariable; Costs, Spillover; Discounting; Efficacy Versus Effectiveness; Markov Models

Further Readings

Farnham, P. G., Ackerman, S. P., & Haddix, A. C. (1996). Study design. In A. C. Haddix, S. M. Teutsch, P. A. Shaffer, & D. O. Dunet (Eds.), *Prevention effectiveness: A guide to decision analysis and economic evaluation* (pp. 12–26). New York: Oxford University Press.

Gold, M. R., Siegel, J. E., Russell, L. B., & Weinstein, M. C. (1996). *Cost-effectiveness in health and medicine*. New York: Oxford University Press.

Haddix, A. C., & Shaffer, P. A. (1996). Cost-effectiveness analysis. In A. C. Haddix, S. M. Teutsch, P. A. Shaffer, D. O. Dunet (Eds.), *Prevention effectiveness: A guide to decision analysis and economic evaluation* (pp. 127–155). New York: Oxford University Press.

Haddix, A. C., Teutsch, S. M., Shaffer, P. A., & Dunet, D. O. (1996). *Prevention effectiveness: A guide to decision analysis and economic evaluation.* New York: Oxford University Press.

Tornado Diagram

A tornado diagram (given this name due to its resemblance to a tornado) is a graphical method for displaying a series of univariate (or one-way) sensitivity analyses that has been commonly used in cost-effectiveness analysis. Figure 1 shows an example of a tornado diagram—with incremental cost-effectiveness value ranges arranged from the largest at the top to the smallest at the bottom. In a tornado diagram, the effects of individual parameter variation on results can be compared visually, allowing analysts to intuitively communicate which parameters are more sensitive to variation—that is, parameters whose variation has the greatest effects on the results of the analysis. In the example shown in Figure 1, individual variation of Drug A costs has the greatest effects on model results, and variation of Drug A side effect utility has the least effect. More recently, the use of tornado diagrams and univariate sensitivity analyses has been downplayed due to limitations of these methods and due to the use of other techniques that overcome these limitations.

To construct a tornado diagram, the analyst varies a single parameter over its range and notes the effects of this variation on the analysis results. In Figure 1, the incremental cost-effectiveness varies from $10,000 per quality-adjusted life year (QALY) gained when Drug A costs $54 to $97,000/QALY gained when Drug A costs $139. The vertical line through the bars at $39,000/QALY shows the base case results of the analysis, the expected value of the analysis when all parameters are set at their base case point estimates. The analyst will then repeat this procedure for each of the remaining parameters to be varied. Once the result range for each parameter has been calculated, parameters are arranged in the diagram from greatest to least effect on model results.

Some decision analysis software packages allow construction of tornado diagrams within the program itself. Tornado diagrams may also be constructed using the graphing functions of Microsoft Excel or through the use of a variety of Excel add-in programs.

Tornado diagrams have a number of limitations. They can be useful when two strategies are being compared, since tornado diagrams typically depict changes in the incremental cost-effectiveness ratio between two strategies. However, when three or more strategies are being considered, tornado diagrams become more difficult to use due to the differential effects of individual-parameter variation on multiple competing strategies and, at times, the effects of strategy dominance. Due to these shortcomings, alternative structures for tornado diagrams have been proposed. One such structure uses the net benefit framework to convert incremental cost-effectiveness ratios to either net monetary or net health benefits, thus bypassing some of the shortcomings of the conventional tornado diagram. However, this formulation gives a less intuitive picture of sensitive parameters and requires,

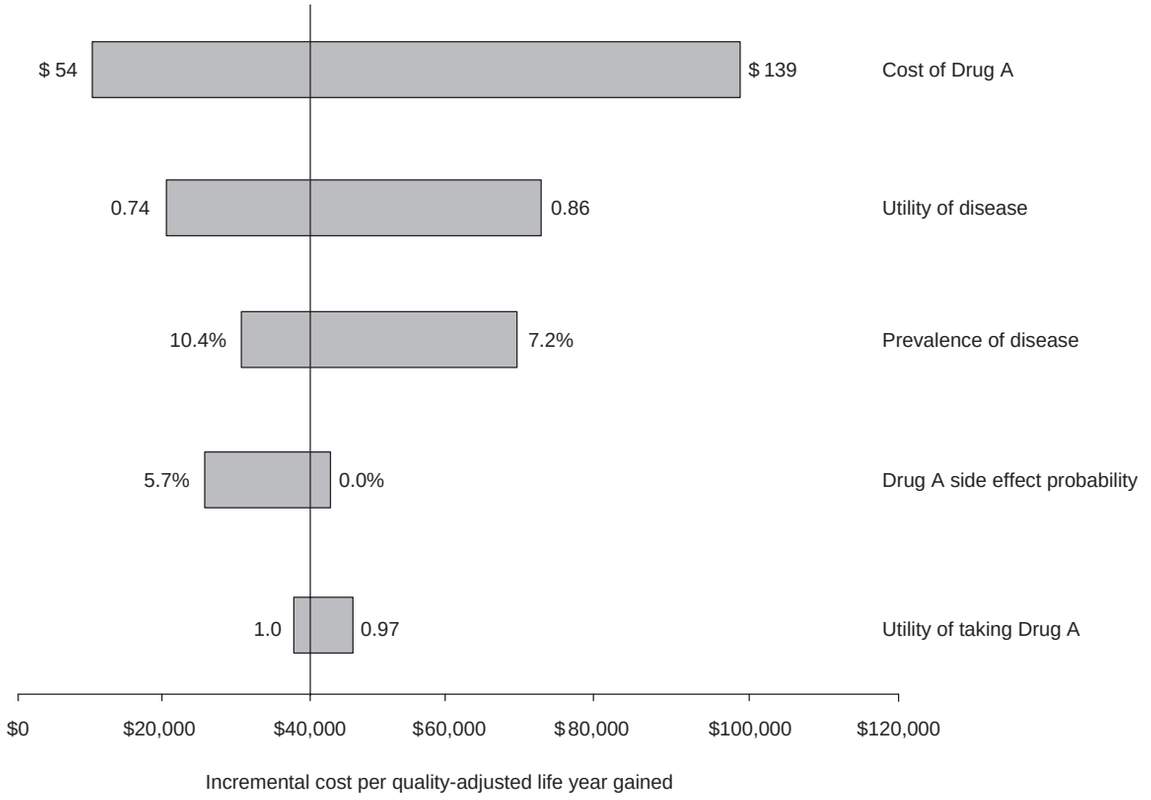


Figure 1 A tornado diagram, depicting changes in incremental cost-effectiveness ratios (the *x*-axis) when individual parameter values are varied

Note: The numbers at either end of the horizontal bar are the high and low parameter values considered in a univariate sensitivity analysis. The vertical line depicts the base case incremental cost-effectiveness ratio of the model.

due to the net benefit calculation, the choice of a single cost-effectiveness acceptability threshold.

Tornado diagrams are also limited by their deterministic nature. The model result ranges depicted in a tornado diagram give no indication of how frequently a given parameter might have a certain value or, similarly, how often a given incremental cost-effectiveness result might occur. From a distributional standpoint, a tornado diagram tacitly assumes that all parameter values have a uniform distribution, where all values in the range are equally likely to occur; however, this almost never is the case.

Another major criticism of tornado diagrams is their reliance on variation of individual parameters, which often may underestimate uncertainty compared with analyses where multiple parameters are varied jointly. Probabilistic sensitivity analysis, where multiple parameters are varied simultaneously over distributions, has been successfully posited as a means of answering the deterministic and univariate limitations of tornado diagrams, leading to the recent decrease in emphasis on tornado diagrams in published cost-effectiveness analyses.

As a result, some have questioned the usefulness of tornado diagrams in medical decision analysis. However, the simple and intuitive structure of the tornado diagram can still be useful in an appropriately structured analysis, showing the parameters whose variation causes the most change in model results. Nonetheless, due to its deterministic and univariate nature, the tornado diagram has become less important in the reporting and interpretation of cost-effectiveness analyses.

Kenneth J. Smith

See also Cost-Effectiveness Analysis; Decision Trees: Sensitivity Analysis, Basic and Probabilistic; Decision Trees: Sensitivity Analysis, Deterministic; Net Benefit Regression; Net Monetary Benefit

Further Readings

Claxton, K., Sculpher, M., McCabe, C., Briggs, A., Akehurst, R., Buxton, M., et al. (2005). Probabilistic sensitivity analysis for NICE technology assessment: Not an optional extra. *Health Economics, 14,* 339–347.

Howard, R. A. (1988). Decision analysis: Practice and promise. *Management Science, 34,* 679–695.

TreeAge Pro Manual. More sensitivity analysis tools: Tornado diagrams (chap. 15). Retrieved July 7, 2008, from http://www.treeage.com/files/pro2008/pdfs/TreeAge%20Pro%202008.pdf

Toss-Ups and Close Calls

The term *toss-up* has different senses in the different contexts in which it is used. In a more technical sense, it is best understood in terms of the flip of a fair coin, where the chance of heads coming up is 50:50 and the chance of tails coming up is also 50:50. The term *close call* has more variation in the medical literature; for example, in the peer-reviewed scientific medical literature, a close call can be interpreted as a near miss in relation to patient safety—for example, during a surgical operation. Yet some individuals may interpret the term *close call* to reflect a decision to be made where the separation of a chance of a benefit accruing to a patient from one of two treatments being compared is, say, 48:52 and not strictly a 50:50 decision between the two treatments.

There are also other senses of the term *toss-up*. In some contexts, it can reflect an unpredictable situation of decision making—for example, in a population, as in an election or any process involving the counting of secret ballots, or in voting, such as on a medical, mental health, surgical, or other ward team. Or there are mixed cases, where, for example, it may not be clear in an individual's care what is the optimal treatment for the patient or which way a vote among medical team members with mixed opinions on what is the best strategy would go if a vote were taken.

In other contexts of speech in medicine, the term *toss-up* simply may be used in an even more general sense to refer to any unpredictable situation and the fact underlying the opinion that there is no systematic way to adequately determine what is optimal care in the patient at a particular time. Jerome Kassirer and Stephen Pauker point to this last sense of the term *toss-up*, where—after careful systematic assessment of the peer-reviewed scientific medical literature and the clinical experience of physicians—the evidence reviewed and assessed shows that there is no difference between treatments and the result is still the same. Thus, the treatments are considered a toss-up from the standpoint of the published peer-reviewed scientific medical literature and a toss-up from the standpoint of clinicians' opinions.

The contemporary published peer-reviewed scientific medical literature contains examples of decisions that are described as "a virtual toss-up" in the areas of screening, diagnostic, and treatment decisions.

Close-Call Versus Clear-Cut Decision Making

Close-call decision making has been separated from more clear-cut decision making and linked to cognitive biases in decision making. Andrea Gurmankin Levy and John C. Hershey studied what they termed *value-induced bias*. They asked volunteers to imagine a serious illness with two possible diagnoses and a treatment with the "same probability" of success for each diagnosis. The authors designed the more serious diagnosis as a clear-cut decision to motivate most subjects to choose treatment. The authors designated the less serious diagnosis a close-call choice. Study participants were randomized to estimate the probability of treatment success before or after learning their diagnosis. The "after" group had the motivation and the ability to distort the probability of treatment success in order to justify their treatment preference. The authors found that in the close-call decision making (but not in clear-cut decision making), individuals may distort relevant probabilities to justify their preferred choices. The authors further argue that those individuals who exhibit value-induced bias in close-call decision making may make suboptimal decisions by distorting relevant probabilities to justify the medical decisions made. This suggests that medical decision making in close-call (or toss-up) decisions is different from decisions that appear to be clear-cut. But there is much more to understand about such apparent toss-ups.

Toss-Up Decisions in Medical Contexts

In medical decision making in the clinic or on hospital wards, for example, the term *toss-up* may be used by a physician or provider loosely in his or her conversation with a patient (or with the patient's family member or significant other when the patient is too ill to make decisions on his or her own behalf) to mean that the physician cannot personally predict how a decision will turn out in the patient's care. Indeed, even an assembled team of experts may not be able to come up with any agreement regarding what is the best course of treatment for the patient with a particular stage of disease at a particular time.

Toss-ups can also be the result of a careful search of the peer-reviewed scientific medical literature on a particular topic in the case of a patient over time. Here, let us consider the case of two treatments, Treatment 1 and Treatment 2, both of which were used in a randomized controlled trial that compared Treatment 1 with Treatment 2 for a particular disease at a certain stage under study. Here, at the end of the research study after the data are analyzed, the results show the same benefits and the same risk for both Treatment 1 and Treatment 2. If this research study was submitted to the peer-reviewed scientific medical literature and survived the peer review process, the resultant published scientific article would state that there were no statistically significant differences between Treatment 1 and Treatment 2 based on this research study, and hence, Treatment 1 and Treatment 2 would be toss-ups in the patient's care. The question that arises here is the degree of surety regarding whether the careful search of the peer-reviewed scientific medical literature was broad enough and had enough depth to capture all relevant scientific research published in the peer-reviewed literature.

When Kassirer and Pauker addressed the toss-up in medical decision making, they took a strict approach to the toss-up in the patient–physician relationship. In a toss-up, the patient can make a decision in any way he or she wants because there is nothing in the scientific evidence or clinical experience of the physician that is available to shed light on the decision. The key word here is *available*—the conditions that exert control on what scientific medical information is available or not available at a particular time.

Kassirer and Pauker's view depends highly on the peer-reviewed medical literature and what it contains and what local experts have to say in case there is no solid peer-reviewed medical literature available to access for guiding the patient's decision making. Where there is no scientific medical evidence or clinical experience on what direction treatment should take, then it is completely up to the patient to base his or her decision on whatever grounds he or she sees fit.

Why Toss-Up Situations Exist in Medicine

This analysis of why toss-up situations exist in medicine may ferret out answers to the question why some toss-up situations are allowed to continue to exist in medicine.

First, there may be no peer-reviewed scientific medical literature on a certain topic because there have been no scientific studies at all carried out on the medical condition, its diagnosis, or its treatment. Rare diseases may fall into this category. The literature may have a letter to the editor of a journal or a case report but no research studies and no clinical case series describing the disease across a set of patients. Hence, there is no scientific research that has been conducted and no published clinical experiences of experts or other physicians managing and treating the disease in question.

Second, there may be no high-quality studies that involve randomized controlled trials among a set of treatments that could be offered to a patient with a particular type of disease at a particular stage. Indeed, the disease itself may be so rare that there is no way to stage the disease because there have not been enough patients with the disease to describe the disease as part of a taxonomy.

Third, there may have been trials, but the trials may not have been completed, and there may be no published scientific articles on the topic in question.

Fourth, there may be cases where the disease is not rare and there are patients with the disease, but these patients may be widely scattered, or there are physicians and researchers would who like to study the disease, but there is no public or private financing available to fund the study of the disease. Here, there is a need to establish the diseases as important enough to study and to receive governmental or public financing.

Fifth, there may have been trials on the disease and its treatment, but the organization that financed the study may not allow the research study to be published. For example, if a prescription medicine does not do well in terms of benefits or risks as contrasted with the other prescription medicines tested in the trial, the manufacturer may not allow the results to be published. Unless there are regulations requiring product manufacturers to publish results of all trials enrolling human study participants, the results may never be available.

Sixth, product manufacturers funding clinical trials on human study volunteers may have principal investigators and representatives of academic medical centers sign contracts that allow only the product manufacturer the right to decide when and if risk data are released to the public.

In addition, product manufacturers may use ghostwriters to help write the scientific articles that are submitted for publication in peer-reviewed medical journals. These ghostwritten articles may attempt to enhance the product's benefits and mute the product's risks. It can be argued that such ghostwritten papers (a) reduce the public trust in product manufacturers, (b) distort the peer-reviewed medical literature as to what the real benefits and real risks of the medical product are, and (c) make the medical decisions based on the enhanced benefit data and muted risk data seem more like toss-ups or close calls when in reality—if all scientific data were released and all facts were known—the medical decisions would be more clear-cut. Full and open discussion of all data straightforwardly expressed in terms of benefit and risk is needed so that medical decisions can be made with the best available information.

Dennis J. Mazur

See also Biases in Human Prediction; Construction of Values; Subjective Probability

Further Readings

Bernstein, M., Hebert, P. C., & Etchells, E. (2003). Patient safety in neurosurgery: Detection of errors, prevention of errors, and disclosure of errors. *Neurosurgery Quarterly, 13,* 125–137.

Eckman, M. H., Levine, H. J., & Pauker, S. G. (1992). Decision analytic and cost-effectiveness issues concerning anticoagulant prophylaxis in heart disease. *Chest, 102,* 538S–549S.

Engelbrechta, M. R., Jagera, G. J., & Severensb, J. L. (2001). Patient selection for magnetic resonance imaging of prostate cancer. *European Urology, 40,* 300–307.

Jenicek, M. (2003). *Foundations of evidence-based medicine.* Boca Raton, FL: Parthenon.

Kassirer, J. P., & Pauker, S. G. (1981). The toss-up. *New England Journal of Medicine, 305,* 1467–1469.

Koenig, H. G., Ford, S. M., & Blazer, D. G. (1993). Should physicians screen for depression in elderly medical inpatients? Results of a decision analysis. *International Journal of Psychiatry in Medicine, 23,* 239–263.

Kuipers, B., Moskowitz, A. J., & Kassirer, J. P. (1988). Critical decisions under uncertainty: Representation and structure. *Cognitive Science, 12,* 177–210.

Levy, A. G., & Hershey, J. C. (2008). Value-induced bias in medical decision making. *Medical Decision Making, 28,* 269–276.

Mello, M. M., Clarridge, B. R., & Studdert, D. M. (2005). Academic medical centers' standards for clinical-trial agreements with industry. *New England Journal of Medicine, 352,* 2202–2210.

Ng, A. K., Weeks, J. C., Mauch, P. M., & Kuntz, K. M. (1999). Decision analysis on alternative treatment strategies for favorable-prognosis, early-stage Hodgkin's disease. *Journal of Clinical Oncology, 17,* 3577–3585.

Tierney, W. M., & Gerrity, M. S. (2005). Scientific discourse, corporate ghostwriting, journal policy, and public trust. *Journal of General Internal Medicine, 20,* 550–551.

Treatment Choices

Treatment choices are the range of options that people may use to deal with a health issue or illness. This might involve different approaches, or combinations of approaches, to treating a health condition. The approaches may range from self-care, with or without medical advice, to deciding to stop treatment. Choices could be made about where, when, how, and by whom to be treated and may include complementary or alternative approaches. The choices that people make are at the heart of patient involvement in medical decision making.

Changes in Consultation Style

In the *paternalistic* ("Doctor knows best") form of healthcare, relatively widespread until the end of the 20th century, the doctor was consulted for treatment advice, and the patient was expected to follow the advice more or less unquestioningly. Sometimes, the treatment was even performed on the patient without discussion; often there was little by way of information or explanation for a treatment regime. Patients who did not follow advice (and there were, of course, many) were deemed noncompliant.

In the past few decades, the conduct of the consultation and the relationship between patient and doctor have changed from the paternalistic model to a more equal style of consultation, sometimes termed *mutualistic*. In this model, the doctor may be seen as an expert in the diagnosis and treatment of the condition; but the patient has a unique experience of the disease, as well as preferences and values that may affect the choice of treatment. Patients' expectations, influenced in part by a more consumerist and less deferential attitude to the medical profession, have combined with changes in clinical training to promote a more balanced encounter between patients and professionals. Instead of talking about compliance with doctors' orders, or even adherence to treatment plans, concordance between patients and doctors has become the aim.

Rationale for Patient Involvement in Treatment Choices

When someone is diagnosed with an illness, he or she wants to hear that there is a remedy, that it has minimal adverse effects, and that there is agreement in the medical profession that this is the best course of action. Of course, where many conditions are concerned, the patient may find out that almost the opposite is true: A definitive cure is still being sought; the adverse effects are off-putting; and specialists in the field have different ideas about the best form of treatment, leading to variation between, or even within, different treatment centers. This can be (at the very least) disappointing and confusing for the patient.

When the outcomes of treatment are more than usually uncertain, the treatment choice may rest heavily on the patient's own priorities and the patient's attitude to the limitations and adverse effects that are associated with the different treatments. In such situations, the outcome that is best for the patient is only likely to be achieved if he or she is involved in making the choice. For example, a surgeon cannot detect through clinical examination whether a woman would prefer a lumpectomy or mastectomy for breast cancer. Surgeons expected that women would prefer conservative treatment, yet when women were given the opportunity to make the choice, many surgeons were surprised how often women chose to have a mastectomy. Among the reasons that affected these women's choice were the belief that if the whole breast is removed, there must be less chance of recurrence. For some women, the fear of recurrence is, quite understandably, greater than the desire to conserve their breast.

There are several possible advantages in involving patients in treatment choices. Many patients leave the consultation apparently willing to follow directions but then either do not get their prescription filled or do not complete the recommended course of tablets. If the patient is involved in making the choice about treatment, it is more likely that he or she will follow and complete the course. In U.K. general practice, it is not unusual for patients to be given a delayed prescription for an antibiotic for a self-limiting condition, such as a sore throat. If the symptoms do not improve within a specified time, the prescription can be filled. This gives the patients the option of deferring their choice about whether they want to embark on the course of antibiotics and helps avoid unnecessary, wasteful, and potentially harmful use of antibiotics. As this example demonstrates, patient involvement in treatment choice does not necessarily mean that more resources are used.

There is also evidence that the public (who, after all, directly or indirectly pays for medical care) *wants* more involvement in treatment choices. A 2006 survey of eight European countries by the Picker Institute found that only a quarter of the respondents from the general public believed that doctors should choose the treatment on their own. Half of the respondents thought that the patient and the doctor should make the choice together, and the other quarter said that the patient alone should decide.

"Real" Choices

To make a choice, people need to be aware that there are options. There is always the option of having no treatment, but people often do not see this as a "real" choice. Women who were offered the option of having adjuvant chemotherapy after treatment for breast cancer did not perceive it as a choice. If the alternative was no treatment, it was simply seen as no choice. Prenatal screening in pregnancy may be optional, but many couples regard it as an inevitable part of prenatal care and are often not even aware that they have exercised a choice. Treatment for prostate cancer includes the option of watchful waiting, sometimes known as active monitoring. Despite uncertainty about the benefits of any treatment for the disease and the unpleasant, and sometimes long lasting, side effects (which include impotence and incontinence) of the standard treatments of surgery, radiotherapy, chemotherapy, and hormone therapy, relatively few men choose watchful waiting. A qualitative interview study of men with prostate cancer sought to understand why nearly all those for whom watchful waiting could have been an option had chosen active treatment. Many men did not feel that they could live with the idea of the disease "hanging over them" and preferred to go ahead with treatment despite the prospect of unpleasant adverse effects. Some initially thought that they might try watchful waiting but were persuaded by their families, and sometimes by support groups, that they should have active treatment. Reasons included preferring to act quickly and decisively and feeling that it must be better to treat than to leave the cancer to possibly grow. The few who had chosen watchful waiting had dedicated time to learning as much as they could about the treatments and the uncertainties; were particularly keen to avoid the common adverse effects of treatment; or had found others to support their choice.

Information for Treatment Choices

When patients are diagnosed, they often have little knowledge about the condition or the treatment options that might be available to them. Choices cannot be made without information. The information that patients may use to help them choose between treatments includes clinical evidence about the outcomes (including survival, quality of life, symptom control) of the various treatments; the likely consequences of having no treatment; the frequency of adverse effects in different treatment regimes; what their own doctor would choose to do in similar circumstances; and what it has been like for other people who have followed the various treatment paths. Some of this information is routinely included in decision-making aids. New information and communication technologies have made all types of information much more easily accessible to patients.

The Internet, in particular, has had a dramatic effect on patients' ability to call on a wide variety of types of information to support treatment choices. Patients often feel that only others who have been through what they are going through can really understand and guide them. It is not surprising therefore that people who use the Internet for health information seek out other patients' stories and experiences online. Clinicians may feel frustrated when patients seem to be more influenced by the experiences of friends, family, or Internet acquaintances than by research evidence. However, the experiences and accounts of other people who have faced similar issues may help patients establish the relevance and salience of different courses of action in their own lives. For example, parents who need to make a decision about whether to enter their child in an immunization program may benefit from hearing how other parents came to make their decision.

Negative Consequences of Patient Involvement in Treatment Choices

Some commentators have suggested that, while greater autonomy and access to information have been widely welcomed, the expectation that patients should make their own treatment choices has gone too far. There is a danger that the expertise of health professionals can become downgraded, or even lost, if the informed advice that they can offer is viewed as equivalent to any other factor or influence on the choice.

In some circumstances, when facing some health problems, patients simply do not want to have to make treatment choices. For example, patients who have just been diagnosed with a life-threatening illness or who have heard that their child is seriously

ill often want to feel *cared for* and not faced with decisions. A person who has just heard bad news is unlikely to be in the best frame of mind to make consequential treatment choices, even in the rare event that he or she knows enough about the condition and treatment options involved.

A qualitative interview study of women with ovarian cancer found that those who were asked to choose between two chemotherapy regimes often felt that their doctors were abnegating their responsibility to advise them about treatment. The option had been offered because a trial comparing the two regimes was at that time incomplete and there was thus no evidence about any difference in overall quality of life or survival. The side effects of the treatments were known to differ and were a matter of individual preferences and priorities. However, women who had either not been told or had not understood the reason that they were being asked to make the treatment choice felt that they were being abandoned by their doctors, sometimes suspecting that this was because of fear of litigation. Women could not understand why their doctors would not tell them what they thought was the best option—especially when they responded with visible relief when the woman made the "right" treatment choice. Patients may not make the choice that their doctor recommends, but they do usually like to know what their doctor would do in similar circumstances. Doctors are sometimes wary of disclosing this information because they are concerned that it will harm their relationship with the patient and be too strong an influence on the choice.

There is often a difference between what people anticipate they might do in a health crisis and what they choose to do when they are actually diagnosed. Barry Schwartz quotes research showing that while 65% of people in a survey said that they would want to choose their own treatment if they were diagnosed with cancer, research with people who had really been diagnosed with cancer showed that only 12% actually wanted to make their own choice. This might be because in a life-threatening situation, the need to feel cared for by a trusted health professional may overtake the desire for autonomy.

Another negative aspect for patients can be living with the responsibility of making their treatment choice. When they are making choices, people often consider which options they are more likely to regret. Choices that are supported by health professionals, as well as friends and family, and that use available information, may be easier to live with because the responsibility is less stark.

Discussion of patients' involvement in treatment choices highlights the importance of good information and clear communication between doctor and patients. It can be hard for the doctor to admit uncertainty to the patient—and no less hard for the patient to hear, but it is a key reason why treatment choices need to involve the patient. This has to be balanced with the humane recognition that patients need to feel cared for and often want some guidance in making their choice.

Sue Ziebland

See also Choice Theories; Decision Quality; Informed Decision Making; Patient Decision Aids; Shared Decision Making

Further Readings

Charles, C., Gafni, A., & Whelan, T. (1999). Decision-making in the physician-patient encounter: Revisiting the shared treatment decision-making model. *Social Science & Medicine, 49,* 651–661.

Schneider, C. (1998). *The practice of autonomy: Patients, doctors and medical decisions*. New York: Oxford University Press.

Schwartz, B. (2004). *The paradox of choice: Why more is less*. New York: HarperCollins.

Ziebland, S., Evans, J., & McPherson, A. (2006). The choice is yours? How women with ovarian cancer make sense of treatment choices. *Patient Education and Counseling, 62,* 361–367.

Tree Structure, Advanced Techniques

The basic decision tree is a graphical representation of a decision model consisting of nodes, branches, variables, and expressions. The standard node types are decision, chance, and terminal nodes. Since decision analysis first became widely used in healthcare, analysts have gradually extended the basic representation to include a number of

advanced features that make models more flexible and versatile. This entry describes these advances and their use in decision modeling.

Branches and Bindings

A branch connects two nodes in the tree (Figure 1) and specifies a specific context. Each branch, or collective path of branches, between nodes in a tree represents a different clinical context, having potentially different probabilities of subsequent events, costs, and utilities. When such values are expressed in terms of symbolic expressions, any parameters may be affected by the corresponding series of events. Bindings are mathematical expressions that reassign the value of parameters in a tree context. For example, in Figure 1, the parameter *pCure* should be different in the contexts of "Empiric therapy" and "Observation." This can be implemented by applying a binding of the form

$$pCure := pCureRx$$

on the branch "Empiric therapy." This expression uses the assignment operator ": =" to indicate that the variable on the left of the operator is assigned to the value of the expression on the right. The assignment of the new value applies to all tree contexts downstream from the binding expression but does not affect other tree contexts at any other points of the tree. While the previously bound value of the variable may be used in the binding expression, the new value overrides the previous value in the subsequent contexts.

Local Versus Global Variable Values

The binding mechanism creates local values for variables. The values apply only to tree contexts downstream from the binding. In some cases, it is useful to create global values, which apply in all parts of a tree. An example is precalculating parameters such as mortality rates and incidence of disease that depend on other variables, such as age. The calculated variables can be used in all parts of the decision model and not only downstream from where they are defined. However, the same effect can often be accomplished by setting a binding at the root of the tree.

Applying Bayes's Rule in a Decision Tree

A common application of bindings in decision models is application of Bayes's rule to calculate the posttest probability of a test from the pretest probability and the test characteristics (sensitivity and specificity). Figure 2 shows a tree with an additional decision node branch labeled "Test."

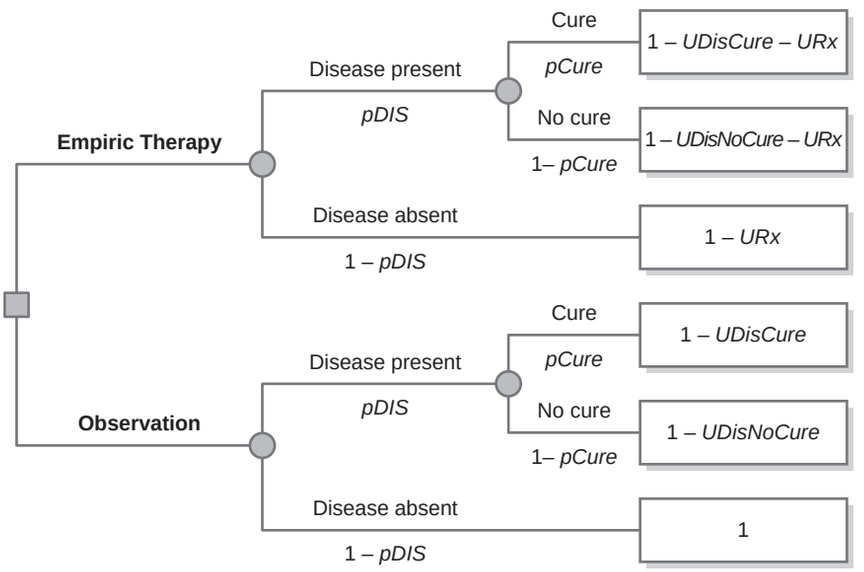


Figure 1 Decision tree

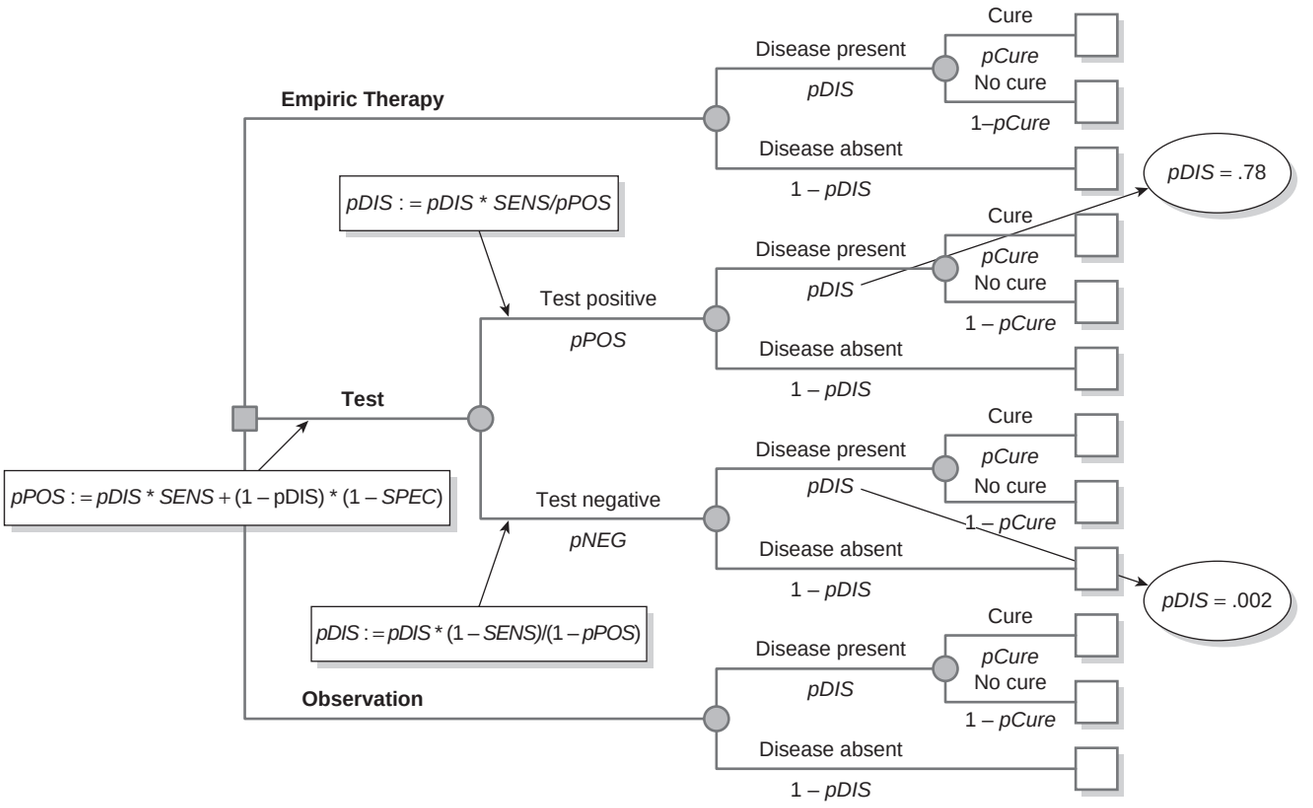


Figure 2 Tree with diagnostic test and bindings

The Test branch leads to a chance node with two branches, "Test positive" and "Test negative."

The probability of the "Test positive" branch (pPOS) is equal to $pDIS \times SENS + (1 - pDIS) \times (1 - SPEC)$, where *SENS* is the test sensitivity and *SPEC* is the test specificity, and *pDIS* is the probability of disease. Note that it is common to use short mnemonics for variable names, using the first character to indicate type, such as "p" for probability variables. This formula can be applied as a binding on the "Test" branch to convey an updated, posterior probability of positive test results in contexts in which disease is present. Given the values for *pDIS* in Figure 4, *SENS* = .95 and *SPEC* = .99, *pPOS* = .19. It is higher than *pDIS* because of false-positive tests.

Bayes's rule can be applied by placing bindings on the "Test positive" and "Test negative" branches, as shown in Figure 7. The bindings result in *pDIS* = .78 following "Test positive" and *pDIS* = .002 following "Test negative," reflecting the likelihood of detecting the disease in each scenario.

Shared Subtrees and Symmetry

Every node or branch of a tree is effectively the root of a unique subtree. However, many decision trees contain repeated structures. For example, in the tree shown in Figures 1 and 2, the two chance nodes with branches "Disease present" and "Disease absent" have structurally identical subtrees. The only difference between these occurrences is the upstream nodes and branches and the values of variables. A convenient modeling construct is to allow nodes to share common subtrees, often denoted by placing a bracket around all the upstream branches, so that the subtree appears only once, as in Figure 3. The bracketed notation is helpful in depicting very large decision models in a compact illustration.

In computer-based implementations of decision trees, subtrees may be represented by attaching them to upstream branches by means of *label* or *link* nodes, as depicted in Figure 4. The label node is indicated here by two vertical bars. The labels on

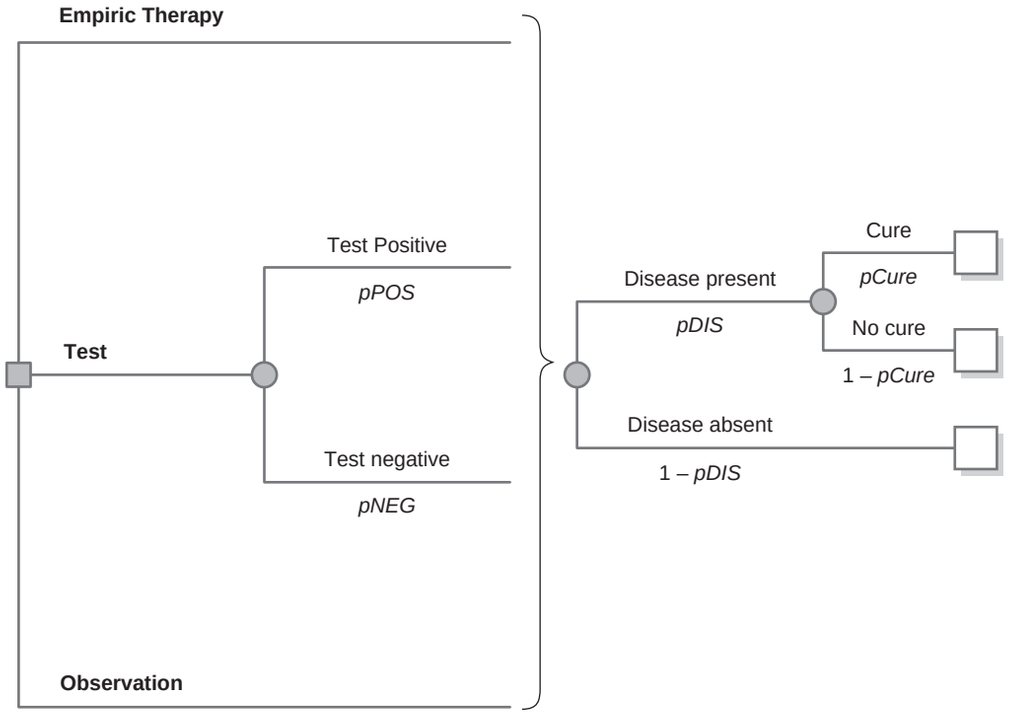


Figure 3 Subtrees: Bracketed notation

the unique branch following each label node refer to the subtree structure downstream from the Disease branch that needs to be represented only once in the model. However, the subtree is evaluated differently in each context because of the action of the bindings. Figure 4 indicates the bindings needed to specify the probability of disease in the "Test positive" and "Test negative" contexts, due to the differences in the bindings to the variable *pDIS*, which appear on the two label node branches that lead to the Disease subtree.

Shared subtrees are important not only as a convenience but also to enforce *symmetry*. Shared subtrees ensure that similar elements of prognosis are modeled the same way for each strategy of the decision tree. Attempting to replicate the subtree structure manually can lead to unintended differences in structure, parameters, or bindings that may lead to errors in evaluation. For example, if the prognosis of disease depends on a particular variable in one part of the tree but appears as a hard coded value or a different variable in another, it can lead to errors when performing sensitivity analysis on that variable.

Chance Nodes With Multiple Outcomes

Chance nodes may contain any number of branches that must represent a set of mutually exclusive, collectively exhaustive events. Therefore, the probabilities of the branches of a chance node must sum to 1. When there are only two branches, this presents no problem because the probability of the second branch can be assumed to be 1 minus the probability of the first branch. The previous decision tree examples in this entry involved a binary disease state: Disease was either present or absent. In many situations, the patient may have one of several diseases. In Figure 5, a chance node models a situation where the patient may have one of two diseases, but they are mutually exclusive, each having an independent marginal probability. In this case, the probability of the "No disease" state must therefore be the remaining complement of the disease states. In reality, the probability of mutually exclusive disease states is not independent. When the probability of one state goes up, the probability of the other goes down. This situation must be represented by a functional relationship between

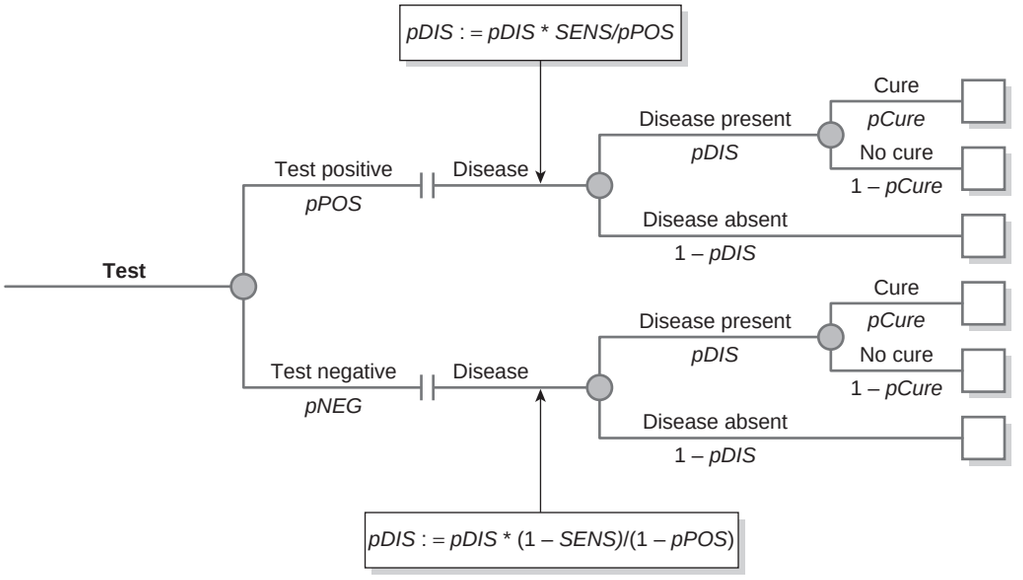


Figure 4 Subtrees with label node and bindings

the variables using algebraic expressions or by retrieving the precalculated probabilities from tables (see Figure 5).

When disease states are not mutually exclusive (e.g., when the model includes coronary artery disease and diabetes), the tree structure in Figure 5 is clinically inaccurate because it doesn't allow for both disease states to occur in the same patient. In this case, the chance node structure shown in Figure 6 is required. Instead of representing all combinations of disease states together as mutually exclusive events at a single node, based on the joint probabilities of events, disease states can also be represented as *sequential* chance nodes in which the probability of the second event may be conditional on the first. There are four possible outcomes, representing either disease alone, both diseases, or neither disease. This structure also ensures that no set of probability estimates will violate the assumption that the combined probability of mutually exclusive events is 1.

Embedded Decision Nodes Versus Normal-Form Strategies

The simplest decision trees have a single decision node at the root of the tree, as shown in Figures 1 and 2. However, decision problems often involve sequential decisions. For example, the first part of the decision may be whether to do a diagnostic test. The second decision may involve which of several treatments to use if the test is positive. This can be illustrated with a tree as shown in Figure 7. The decision node at the root of the tree represents the test decision. The branches of the chance node representing the test result lead to a second decision node, referred to as an *embedded* decision node, which models the treatment decision. A relatively efficient heuristic algorithm to evaluate this model involves first folding back the tree distal to each branch of the second decision node and for each, and selecting the branch with the highest expected utility. Then the root decision node would be evaluated, substituting the values of the

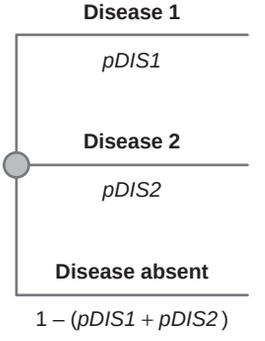


Figure 5 Chance node with three branches

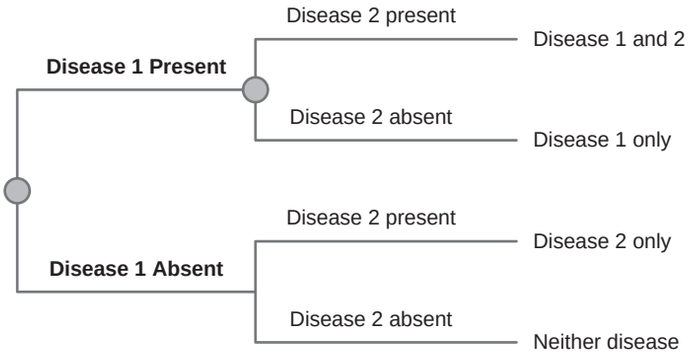


Figure 6 Two nonmutually exclusive diseases

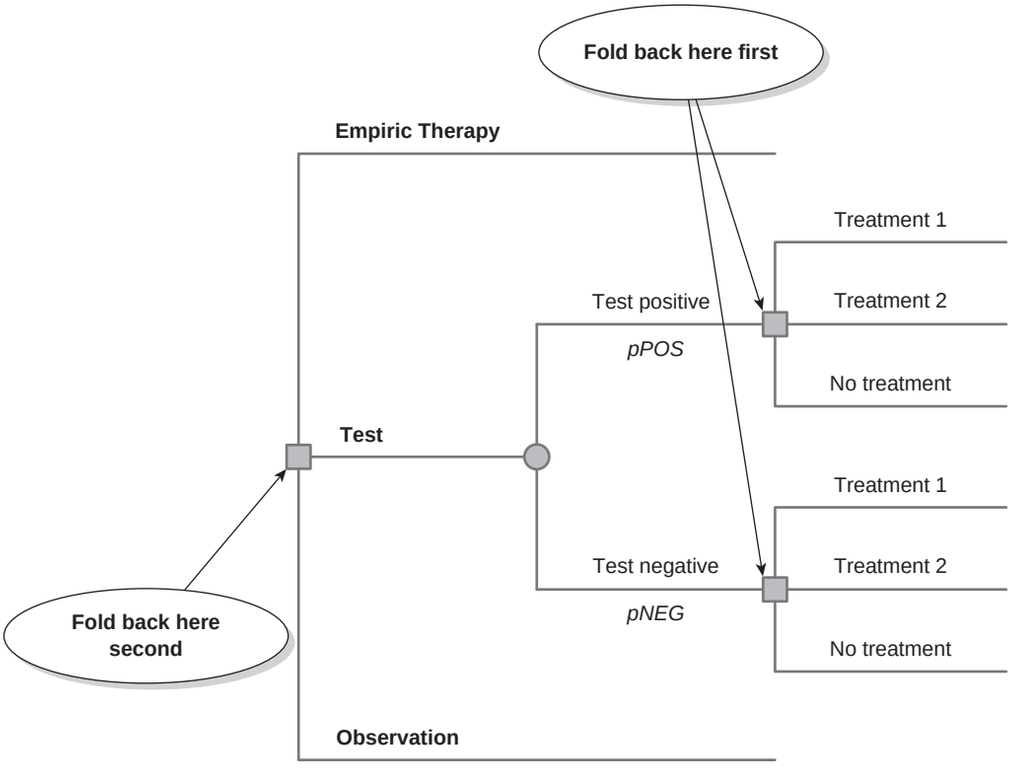


Figure 7 Embedded decisions

embedded decision nodes for the branches of the Test node. However, this evaluation method only results in an overall optimal combination of decisions if one can assume that the optimal embedded decision will always appear in the optimal policy. Furthermore, this algorithm breaks down completely in cost-utility models because there is no defined decision rule at embedded decision nodes; the preferred option depends on the willingness-to-pay threshold and varies according to the threshold chosen.

The proper evaluation algorithm involves considering all possible choices, policies, or strategies and may be represented explicitly by converting such a tree into a structure in which each combination of decisions appears as a top-level alternative, as shown in Figure 8. Instead of representing the downstream decisions as embedded decision nodes,

Figure 8 Decision node in normal form

each combination of decisions is represented as a separate branch of the root decision node. This is sometimes referred to as *normal form*. The combination of initial choice and contingent downstream choice (e.g., "Test, treat if positive") is referred to as a *strategy*. In this case, there are possible strategies that are not represented by branches of the decision node because they do not make sense. For example, the strategies "Test, Treatment 1 if negative" or "Test, No treatment if positive" do not make logical sense because if treatment is indicated after a negative test, then empirical treatment is better than testing. If treatment is not indicated after a positive test, then observation is better than testing. These illogical strategies are sometimes referred to as *degenerate* strategies and may be pruned from the tree so that they will not be subject to unnecessary consideration. Nevertheless, the number of strategies that can be represented in normal form can be very large and is determined by the number of sequential decisions and the number of possible choices for each.

Boolean Nodes and Dynamic Tree Structure

There are times when the structure of the tree needs to change depending on the context or the characteristics of a patient or population. For example, a decision problem may consider an event that occurs only when certain characteristics of the patient are present. While such cases might be represented in terms of chance nodes having mutually exclusive probabilities of 0 and 1, a simpler and more explicit representation can be achieved using a *Boolean* node, which may have any number of branches, only one of which is active. The active branch is determined by the truth of *logic expressions* (whose truth values, true and false, are analogous to the probabilities, 1 and 0, of branches of a chance node). Any branch with a false logic expression is ignored. One such example, shown in Figure 9, models events that might occur with postmenopausal hormone replacement. One of the events of interest is endometrial cancer, but only in women who have a uterus. This can be modeled by using a Boolean node. The risk of endometrial cancer is included in the model only when the logical variable "Uterus" is true. Another example of the use of Boolean nodes is to control the structure of the tree based on a condition. For example, the tree may be a diagnostic test that is to be considered only when a Boolean variable (e.g., "Do test") is true. If the test branch is reached through a Boolean node and its logic expression is "Do test," then the test will be performed only when "Do test" is set to true. Logic expressions may be of arbitrary complexity and may incorporate control variables (such as "Do test") and patient characteristics (e.g., "age < 60").

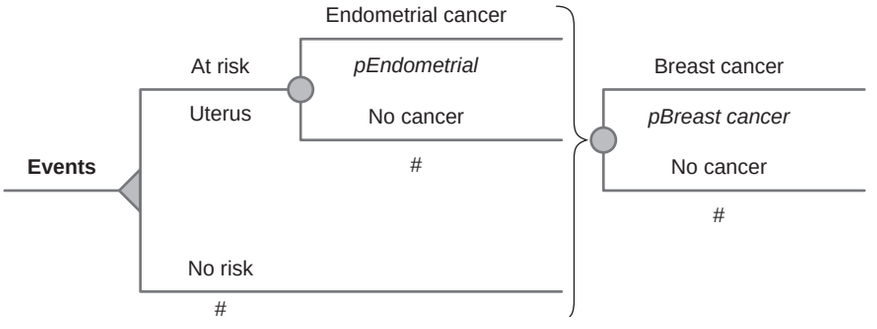


Figure 9 Boolean node

Recursive Trees

The decision trees illustrated so far indicate events occurring simultaneously (e.g., performance of a test, obtaining a result, and having or not having a disease) or in an unspecified time frame (e.g., being cured of a disease) and consider each event to occur only once. This structure may be inadequate to model situations where (a) an event may occur more than once or (b) the timing of an event is uncertain and the time when it occurs is important (e.g., because the risk of the event or its consequences change over time). Although they have been almost entirely replaced by Markov models, recursive trees were an early means of modeling these situations.

Figure 10 shows an example of a tree that uses a recursive tree to model the outcome during a 6-month course of anticoagulant therapy. For each month considered, the "Event" chance node models whether an episode of bleeding occurs. If so, a terminal node is reached (perhaps because it is assumed that anticoagulant therapy would be stopped). If no bleeding occurs, the patient is "At risk" for further events. "At risk" has two branches, one of which leads back to the Event node. Therefore, a cycle, or recursion, is set up. This structure requires either a *stopping criterion* (*exit condition*) or probability functions, which are certain to decrease below some threshold, to ensure that evaluation is not infinite. The Boolean node "At risk" models the stop condition using a counter variable, "Month," which represents the number of months on anticoagulant therapy. The "At risk" node returns to Event if Month < 6. Month is set globally to 0. At each return to the "At risk" node, a binding increments Month by 1. This ensures that the model doesn't try to evaluate more than six passages through the Event node. The recursive tree is one method for modeling recurrent events or risk using a simple tree. Recursive trees have been replaced almost entirely by Markov models in which the iterative parameters are explicit and easier to control. However, technically any Markov model can be represented and evaluated by a corresponding recursive tree.

Functions and Tables

Mathematical expressions can be used in probabilities, utilities, and bindings to represent relationships (linkages) between variables. Functions can be used

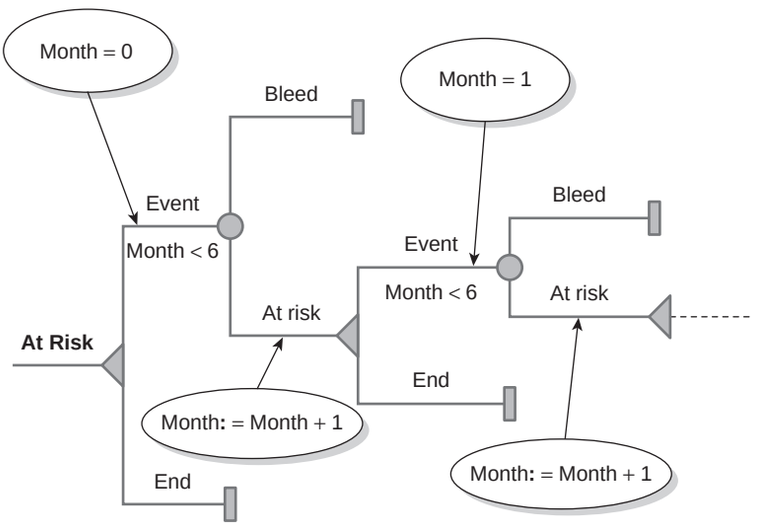


Figure 10 Recursive tree

to express the values when there is a convenient closed-form expression describing the relationship. However, many relationships do not fit this requirement. Examples are the relationship between age and mortality, age-related incidence of disease, and rates of complications after a clinical event. These situations can be modeled conveniently using *tables*. These associate each value of an independent variable (e.g., age) with a unique value of a dependent variable. The independent variable is referred to as the *index* of the table. A common example is the relationship between age and mortality rate. A portion of the corresponding table showing annual mortality rates for ages 50 to 60 is given below:

Age	*Annual Mortality Rate*
50	.0024
51	.0030
52	.0033
53	.0036
54	.0039
55	.0043
56	.0048
57	.0053
58	.0059
59	.0065
60	.0072

Mortality rates are related to age by inserting an expression into a probability or binding of the form TableName[index]—for example, mAge[Age]. This is replaced during evaluation by the value of the table corresponding to the value of Age. If the value of Age is in between two index values of the table, the decision analytic software can perform interpolation to determine the appropriate table value.

Frank A. Sonnenberg and
C. Gregory Hagerty

See also Bayes's Theorem; Decision Trees, Construction; Decision Trees, Evaluation; Decision Trees: Sensitivity Analysis, Deterministic; Diagnostic Tests; Markov Models; Markov Models, Applications to Medical Decision Making; Markov Models, Cycles

Further Readings

Lau, J., Kassirer, J. P., & Pauker, S. G. (1983). Decision maker 3.0. Improved decision analysis by personal computer. *Medical Decision Making, 3*(1), 39.

Pauker, S. G., & Kassirer, J. P. (1981). Clinical decision analysis by personal computer. *Archives of Internal Medicine, 141*(13), 1831.

Roberts, M. S., & Sonnenberg, F. A. (2000). Decision modeling techniques. In G. B. Chapman & F. A. Sonnenberg (Eds.), *Decision making in health care: Theory, psychology, and applications*. New York: Cambridge University Press.

Sendi, P. P., & Clemen, R. T. (1999). Sensitivity analysis on a chance node with more than two branches. *Medical Decision Making, 19*(4), 499–502.

Sonnenberg, F. A., & Pauker, S. G. (1987). Decision maker: An advanced personal computer tool for clinical decision analysis. In W. W. Stead (Ed.), *Proceedings of the Eleventh Annual Symposium on Computer Applications in Medical Care*. Washington, DC: IEEE Computer Society Press.

Trust in Healthcare

This entry examines the conceptual issues and empirical research regarding patients' trust in healthcare. First, the meaning of the concept *trust* is explained. Second, the impact of trust in healthcare systems is explained, and the elements of trust that may be particularly important in this context are discussed. Finally, several relevant results that emerge from a review of the literature on the topic are described.

Concept

Trust has been defined as "the optimistic acceptance of a vulnerable situation in which the trustor believes that the trustee will care for the trustor's interests." According to Lucy Gilson, this definition emphasizes several key features of trust. First, it implies that trust is a relational notion; that is, it emerges from a set of interpersonal

behaviors. These behaviors are guided by sets of institutional rules, laws, and customs. At the microlevel, we can distinguish several types of trust relations: between an individual patient and a physician, between two physicians, and between a physician and his or her manager. At the macrolevel, we can consider patient and public trust in physicians and managers in a particular healthcare organization and in the whole healthcare system. Second, the definition of trust involves a notion of vulnerability and risk, and it is rooted in the expectation that the other will have concern for one's own interests. That is, trust appears to be necessary to cope with situations that have an element of uncertainty regarding the motives, intentions, and future actions of other individuals and organizations that we depend on. Trust appears to be particularly important in the healthcare area because this is a setting characterized by uncertainty regarding the competence and intentions of the physician on whom the patient is reliant. The need for interpersonal trust emerges from the vulnerability associated with being ill as well as the information asymmetries and unequal relationships specific to the medical domain. This explains why most definitions of trust combine expectations about ability, competence, and knowledge of the physician with expectations about his or her ethics, integrity, and motives.

Impact

What are the consequences of trust in the healthcare system? There is substantial evidence that trust mediates healthcare processes. In fact, it has been argued that trusting patient–physician relations have a direct therapeutic effect, although evidence to support such claims is still scarce, mainly because of the lack of intervention studies examining the effect of trust on health outcomes. There is a broad agreement that trust has an indirect influence on health outcomes through its impact on patient satisfaction, adherence to treatment, and continuity with a provider. Trust also encourages patients to access healthcare and to disclose the information necessary to make an accurate and timely diagnosis. Trust, therefore, underpins patient behaviors important for effective treatment. Trust is also a quality indicator, with patients suggesting that high-quality doctor-patient interactions are characterized by high levels of trust. From an organizational perspective, trust is believed to be important in its own right; that is, it is intrinsically important for the provision of effective healthcare and has even been described as a collective good, similar to social capital. Specific organizational benefits that might be derived from trust as a form of social capital include the reduction in transition costs due to lower surveillance and monitoring costs and the general increase in efficiency. It has been argued that trust also has some costs and dangers. Although trust may provide legitimacy for the exercise of power, trusting too much, without caution, may also enable the abuse of power in the form of exploitation, domination, or conspiracy against others. This is a particular danger for healthcare given the vulnerability of all patients, but particularly those from disadvantaged backgrounds, in relation to healthcare providers.

Research

Trust research started to gain momentum with the 1990 publication of Anderson and Dedrick's trust in physician scale. Trust continues to gain increasing attention in the medical and health literatures. The dominant focus of this research is on patients' interpersonal trust in a specific physician. Studies of dimensions of trust in broader medical institutions are lacking. Despite this limitation, several interesting points emerge from the extant literature on the topic. First, in addition to the scale noted above, several research groups have published scales for studying medical trust, each one addressing a patient's trust in his or her individual physician. Second, regarding the core of patient trust, researchers tend to agree that trust depends heavily on a patient's overall assessment of a physician's personality and professionalism and that it is driven fundamentally by the vulnerability of patients seeking care in a compromised state of illness. Accordingly, trust in physicians consists of the following domains in the following order of importance: loyalty or caring, competency, honesty, and confidentiality. All reported trust scales that include these dimensions have high internal

reliability and good construct validity in that they show expected associations with other measures, such as being positively correlated with the length of the relationship. Some of them have been found to have good predictive validity for outcomes expected to be sensitive to trust, such as following treatment recommendations and staying with the same physician.

Research into trust relations also explored the nature and form of trust in terms of its different types; the factors that build, sustain, or detract from trust; and the effects of high or low trust. Overall, this research suggests that while patients retain high levels of trust in individual physicians ("your own doctor"), lower levels of trust are found for healthcare institutions. In fact, patients' trust in their personal physicians has stronger elements of faith than does trust related to other social or economic areas—perhaps more like the form of trust that exists in intimate or fraternal interpersonal relationships. This result suggests that the relationship between the perceived performance of the healthcare system at the microlevel and the perceived quality of healthcare provision at the macrolevel is a complex one. Research needs to examine how institutional trust influences interpersonal trust and vice versa.

Taken together, empirical research clearly complements theory and suggests that developing a trustworthy healthcare system requires more than competent physicians. More important, it needs health workers that have the motivation and capacity for empathetic understanding of patients, as well as institutions that sustain ethical behaviors and so provide a basis for trust.

Rocio Garcia-Retamero and Mirta Galesic

See also Decision Making and Affect; Heuristics; Informed Decision Making; Motivation; Patient Rights; Patient Satisfaction; Shared Decision Making

Further Readings

Anderson, L., & Dedrick, R. F. (1990). Development of the trust in physician scale: A measure to assess interpersonal trust in patient physician relationships. *Psychological Reports, 67,* 1091–1100.

Calnan, M., & Rowe, R. (2004). *Trust in health care: An agenda for future research*. London: Nuffield Trust.

Calnan, M., & Rowe, R. (2006). Researching trust relationships in health care. *Journal of Health Organization and Management, 20,* 349–358.

Gilson, L. (2006). Trust in health care: Theoretical perspectives and research needs. *Journal of Health Organization and Management, 20,* 359–375.

Hall, M. A. (2006). Researching medical trust in the United States. *Journal of Health Organization and Management, 20,* 456–467.

Hall, M. A., Dugan, E., Zheng, B., & Mishra, A. (2001). Trust in physicians and medical institutions: What is it, can it be measured, and does it matter? *Milbank Quarterly, 79,* 613–639.

Hall, M. A., Zheng, B., Dugan, E., Camacho, F., Kidd, K., Mishra, A., et al. (2002). Measuring patients' trust in their primary care providers. *Medical Care Research and Review, 59,* 293–318.

Kao, A., Green, D. C., Davis, N. A., Koplan, J. P., & Cleary, P. D. (1998). Patients' trust in their physicians: Effects of choice, continuity, and payment method. *Journal of General Internal Medicine, 13,* 681–686.

Mechanic, D. (1998). Functions and limits of trust in providing medical care. *Journal of Health Politics, Policy and Law*, *23,* 661–686.

Thom, D. H., Hall, M. A., & Pawlson, L. G. (2004). Measuring patients' trust in physicians when assessing quality of care. *Health Affairs, 23,* 124–132.

U

Uncertainty in Medical Decisions

Uncertainty has many definitions and conceptualizations. In decision making, uncertainty refers to unknown or uncertain (probabilistic) outcomes of decisions. Probability is the mathematical expression of the degree of uncertainty. Medicine is fraught with uncertainty, and medical practice involves dealing with it on a day-to-day basis. Due to the uncertainty inherent in the environment, optimal decisions are not guaranteed to give the desired outcome. Furthermore, uncertainty can lead to variability in medical decisions, with the same (type of) patient being treated differently by different physicians. Uncertainty is therefore a central issue in medical decision making. This entry (a) explores uncertainty in the medical tasks of diagnosis, treatment, and prognosis, providing examples; (b) uses a decision analytic framework to identify types of uncertainty in clinical decisions; and (c) identifies ways of coping with uncertainty and facilitating decision making.

Uncertainty in Diagnosis, Treatment Decisions, and Prognosis

Uncertainty characterizes all core activities in medicine: diagnosis, treatment decisions, and prognosis. Differential diagnosis is essentially a process of dealing with uncertainty in the interpretation of information relating to the symptoms and signs of disease and the results of diagnostic tests. Examples include whether chest pain indicates angina, pulmonary embolism, musculoskeletal cause, or dyspepsia; or whether a normal electrocardiogram (ECG) can exclude acute coronary syndrome. Uncertainty in treatment decisions relates to the probability that an individual patient will be benefited or harmed, for example, a crucial choice between ventilation and palliative care for a patient with chronic obstructive pulmonary disease who is in acute respiratory distress. In addition, decisions about particular treatments within a healthcare system will revolve around issues of uncertainty in cost-effectiveness, for example, the best strategy for a patient with dyspepsia (test for H-pylori, treat anyway, or just prescribe proton pump inhibitors [PPIs]). Prognosis is probably the most uncertain of all medical tasks due to the unpredictability of future events in relation to specific patients, for example, the risk of developing postoperative complications, the likelihood of cancer recurrence, or a patient's life expectancy. Prognosis influences treatment choice and decisions; therefore, uncertainty in prognosis will increase uncertainty in treatment decisions.

Collecting more information can increase certainty; however, this often implies some cost, financial or otherwise (delay, inconvenience to the patient, pain, risk of injury). On the other hand, if there is no cost involved or the cost is small, physicians may collect more information than required or information that is not guided by specific hypotheses. This may increase physician confidence without necessarily altering the objective probability of the disease or may provide data that are difficult to interpret.

Types and Sources of Uncertainty

From a statistical viewpoint, there are two types of uncertainty: first- and second-order uncertainty. Take the example of whether a patient will benefit more from surgery or medical therapy, where the rates of cure at 5 years are 40% for medical therapy and 60% for surgery. First, one cannot predict precisely whether an individual patient will benefit or be harmed. In spite of the observed mean differences obtained by research (40% vs. 60%), individuals in a population are either "cured" or "not cured," not "20% cured." This is first-order uncertainty and is governed by chance and the mathematical laws of probability. There is also the uncertainty about the precision of the estimate of the relative effectiveness of surgery or medical therapy. This is known as second-order uncertainty and can be due to insufficient or conflicting evidence. If there is high second-order uncertainty about a treatment procedure (rates of success or failure, complications), the physician's confidence in its efficacy will be low. Second-order uncertainty is expressed as statistical variation around a point estimate of probability and can be reduced by greater information in terms of research findings. For example, although the absolute risk difference is 20% in favor of surgery, the 95% confidence interval of that estimate may cover the range of a 5% harm to a 45% benefit.

Using a decision analytic framework, different types of uncertainty can be identified in medical decisions:

1. Uncertainty about the options available for diagnostic and treatment strategies, for example, diagnostic tests, medications, and treatment procedures
2. Uncertainty about the outcomes of each option, for example, the potential interactions and side effects of drugs and complications from treatment
3. Uncertainty about the probability of each outcome, for example, the positive predictive value of a test and the success or failure rate of a treatment
4. Uncertainty about the utility of each outcome, that is, the values patients place on outcomes (e.g., miscarriage due to amniocentesis vs. a baby with Down syndrome). Patient preferences and values about different health outcomes are volatile, and utilities are difficult to measure. It is also debatable whether a mean utility, used in health policy decisions, is appropriate for decisions concerning individual patients.

These types of uncertainty may stem from the lack of published, scientific evidence; the difficulty of the physician to keep up-to-date with medical developments; and his or her ability to assess the published evidence critically.

Lack or unreliability of scientific evidence is the most widely cited source of uncertainty in medicine. However, a second, major source of uncertainty relates to the application of evidence to individual patients. Patients respond differently to the same risk factor, drug, or treatment. They experience the same disease differently and exhibit different symptoms. Moreover, they communicate their symptoms differently. These patient-related individual differences make diagnosis, management, and prognosis in medicine different from related tasks in other domains, for example, fault diagnosis in man-made systems.

Uncertainty in Policy Decisions

Over and above the decisions that physicians make about the management of individual patients, healthcare systems, insurers, and government departments make decisions about the provision of care to patient populations with particular conditions and the treatments that they are able to receive. The accepted means of incorporating evidence into this process has become (a) the use of a probabilistic model that encapsulates the second-order uncertainty in the relevant parameters and (b) value of information analysis to determine the trade-offs between taking a policy decision (e.g., recommending adoption of a test, drug, or treatment procedure) and commissioning more research on the topic.

Coping With Uncertainty

Bounded Rationality

The concept of bounded rationality was introduced by Herbert Simon to describe the ability of

humans to make good decisions in complex, uncertain situations and under time constraints, despite their limited memory and information processing capacity. Humans are limited in their ability to encode and retain information in memory, retrieve information when needed, and manipulate information as required for optimal decisions. At the same time, they have access to limited or uncertain information and must decide more or less quickly. Examples of such situations in medicine range from managing rapidly deteriorating patients in the operating theater, emergency room, or intensive care to diagnosing a patient with chest pain in the primary care clinic.

Despite the difficulties outlined above, doctors will make good decisions or help patients make good decisions, most of the time. This is because humans are adaptive decision makers and have learned to cope with uncertainty. They have adapted to the demands of their environment and developed strategies that allow them to make satisfactory, albeit not necessarily optimal, decisions. This is known as *satisficing*. People thus manage to survive and operate successfully in complex environments. Examples of such "approximate" strategies include pattern matching, various heuristics (rules of thumb), and the tendencies to consider only a few options, to evaluate options sequentially, and to seek confirmatory evidence. These are unconscious strategies and people are not necessarily aware that they are using them. These approximate strategies can sometimes lead to error, for example, when the probability of an event (e.g., a disease) is judged by how easily it is remembered. This is known as the *availability heuristic*. This heuristic can lead to wrong diagnoses, as memorability is influenced by a number of factors other than true frequency of an event (in this case, the true prevalence of the disease in the population of interest).

Coping Strategies

Physicians employ a number of strategies for coping with uncertainty, for example, reducing, acknowledging, or suppressing uncertainty.

Reducing Uncertainty

Uncertainty can be reduced by collecting more information (e.g., diagnostic tests or a second opinion); deferring the decision until more information becomes available (i.e., watchful waiting); or initiating management on the basis of a hypothesized diagnosis and monitoring how it influences the patient's symptoms. Uncertainty can also be reduced by extrapolating from the available information, for example, by making assumptions on the basis of circumstantial and contextual information (e.g., how the patient looks, his or her age, sex, area of residence, occupation, hobbies). Experienced physicians are known to make more and better use of contextual information than less experienced ones.

Acknowledging Uncertainty

Physicians may acknowledge uncertainty and make contingency planning when selecting a course of action, for example, by *safety netting* for a serious but less likely disease. This could be considered as an attempt to minimize regret, should the patient be found to suffer from the serious disease. Safety netting may involve close follow-up, further tests and investigations, or referral of the patient to a specialist. Safety netting is also employed when the physician is sufficiently unsure of the diagnosis or thinks that the patient is sufficiently unwell or worried about his or her symptoms.

Suppressing Uncertainty

Finally, uncertainty may be suppressed and physicians may appear more confident and categorical than the available information warrants. This may be an unconscious process and a survival strategy in situations where reducing uncertainty and contingency planning are not possible but an action is still required.

Physicians who have been longer in medical practice report less stress from uncertainty than their younger colleagues. Physicians with less experience tolerate uncertainty less well and seek to reduce it by gathering more information; this can sometimes involve performing unnecessary tests and investigations.

Decision Support

Uncertainty can be reduced by following clinical algorithms and practice guidelines. These can be available to the physician on paper or as part of a decision support system. Guidelines that are vague and general do not reduce uncertainty effectively,

as they cannot easily be applied to specific patients. Simple and easy-to-remember diagnostic or prediction rules are more effective, especially when they are incorporated in a computerized decision support system, activated at the point of care.

Olga Kostopoulou

See also Bias; Bounded Rationality and Emotions; Decision Trees: Sensitivity Analysis, Basic and Probabilistic; Expected Value of Perfect Information; Heuristics; Prediction Rules and Modeling

Further Readings

Klein, G. (1998). *Sources of power: How people make decisions*. Cambridge: MIT Press.

Kostopoulou, O., & Wildman, M. (2004). Sources of variability in uncertain medical decisions in the ICU: A process tracing study. *Quality & Safety in Health Care, 13*(4), 272–280.

Lipshitz, R., & Strauss, O. (1997). Coping with uncertainty: A naturalistic decision-making analysis. *Organizational Behavior and Human Decision Processes, 69*(2), 149–163.

Moskowitz, A., Kuipers, B., & Kassirer, J. (1988). Dealing with uncertainty, risks, and tradeoffs in clinical decisions. *Annals of Internal Medicine, 108*(3), 435–449.

Simon, H. (1990). Invariants of human behavior. *Annual Review of Psychology, 41,* 1–19.

Unreliability of Memory

Patients may exhibit many different forms of cognitive vulnerabilities. The areas where cognitive vulnerabilities take on an important form is when there is a strong need for patients' active participation in medical decision making. Oftentimes, in the discussion of medical decision making there is limited discussion of cognitive processes after the decision is made. And these after-the-fact cognitive processes fall under the rubric of memory. Memory is defined as the processes in both living and nonliving entities that acquire, store, and retrieve information. Two areas where memory plays a key role in medical decision making are in consent and informed consent in clinical care and informed consent in research on humans.

Memory in Medical Decision Making

In the area of medical decision making, medical decision makers would prefer a state of the world where everyone had perfect memory. But there is often confusion regarding what is meant by the term *memory*. In addition, there is a need to explore what is meant by the concept of "memory" when one alleges that a person's memory is unreliable.

Memory is key in medical decision making. Both the individual who has to come to a decision involving his or her medical care and the individual involved in disclosing information related to that decision ideally must have optimal short- and long-term memory about what actually transpired in that decision-making session. There are two sessions that are key in medical decision making.

First, there is the consent or informed consent session in clinical care between the patient who comes to a physician (or other healthcare provider) for care and the physician (or other healthcare provider) who discloses the information about the medical intervention or procedure the physician is recommending (if there is one optimal medical intervention), the risks, and the alternatives involved in the patient's decision.

Second, there is the informed consent session in research on humans between the individual being approached for research study participation as a study volunteer and the principal investigator of the study (or the principal investigator's designee).

Ideally in both of the above cases, both parties of the decision of whether to embark on the physician-recommended medical intervention or whether to enroll in the research study involving humans would have highly accurate recall of what was said and what transpired in each session. However, this is not the case, and the existence of such an ideal memory about events held between two parties is often fraught with errors of memory. The research on two types of memory—*gist memory* and *verbatim memory*—helps understand what is going on with such errors of memory in two-party decision-making sessions.

Gist Memory

Andrew Budson and colleagues define *gist* as the general meaning or general idea conveyed by a collection of items. There is an increasing recognition among memory researchers, for example, Valeria

Reyna and Allan Hamilton, that individuals—during and after consent and informed consent sessions—behave more as having gist memory. In cases of consent and informed consent, for these authors, gist memory is a "selective" memory of things that are important to individuals that are gleaned from disclosures that are made by physicians or research investigators in consent or informed consent sessions.

Verbatim Memory

Verbatim memory is memory that involves the actual accurate remembering of what went on in a decision-making session in terms of facts and details. Verbatim memory is what everyone would like to have to carry out a successful prospective decision-making session. In addition, verbatim memory is what courts would like patients to have when patients allege lack of physician or principal investigator disclosure and when the courts are called in to adjudicate problem cases alleging lack of disclosure related to risks in consent in clinical care or lack of disclosure related to risks in informed consent in clinical care settings or in research studies on humans.

Illustrative Study

Andrew Lloyd and colleagues surveyed 71 patients on the waiting list for carotid endarterectomy (CEA). CEA is a surgical intervention that is designed to remove a blockage at the site of the carotid artery to prevent future stroke, but it is also a procedure that carries with it the risk of stroke during or shortly after the performance of the CEA itself.

Lloyd and colleagues assessed these 71 patients on the CEA waiting list regarding their recall of risk and benefit to health information provided by their surgeon related to the CEA they were waiting to undergo in the surgeon's care. Patients were surveyed 1 month after their initial consultation, and a subgroup was surveyed again on the day before their operation.

The researchers found that patients' estimates of their baseline risk of stroke without surgery were significantly different from what they had been told by the surgeon. Patients' estimates of stroke risk due to surgery ranged from 0% to 65% (actual local risk 2% that was provided to the patient by the surgeon at the initial consent session).

The researchers also found that patients also had unreasonable expectations about the benefit of the operation for their health that were well beyond the information provided to them by their surgeon. Estimates of stroke risk correlated positively with the degree of expected benefit from the operation ($r = .29$, $p = .05$). When resurveyed the day before the operation, patients' perceptions of both risk and benefit had increased significantly.

The authors concluded that most patients in their study failed to understand the risks and benefits associated with CEA and did not recall the risk disclosure information provided to them by their physician. Some patients' estimates of stroke risk were actually greater than the perceived potential benefit of surgery in terms of risk reduction.

In their analysis of Lloyd and colleagues' study, Reyna and Hamilton clarify what is going on in Lloyd and colleagues' study in terms of verbatim memory and gist memory in relation to patient memory related to physician disclosure of risk and benefit information.

First, verbatim memory in Lloyd and colleagues' study involves remembering the facts and details of disclosed information. Second, gist memory in Lloyd and colleagues' study involves the understanding and interpretation that is placed on that disclosed information by the individual hearing the disclosure message.

Reyna and Hamilton argue that what needs to be more widely recognized is that verbatim memories—that is, memories of the facts and details of the disclosed message in consent and informed consent sessions or other decision-making sessions—are the memories that fade rapidly, while the gist memories of the individual—that is, memories of how that disclosed message is understood and interpreted by the individual hearer of the message—are the memories that endure with the patient over time. Thus, unreliability of memory can be understood in terms of memory traces of gist memory, which are the memories that endure over time and the failure of persistence of verbatim memories that are apt to fade rapidly after disclosures are made.

Dennis J. Mazur

See also Cognitive Psychology and Processes

Further Readings

Budson, A. E., Todman, R. W., & Schacter, D. L. (2006). Gist memory in Alzheimer's disease: Evidence from categorized pictures. *Neuropsychology, 20,* 113–122.

Haque, O. S., & Bursztajn, H. (2007). Decision-making capacity, memory and informed consent, and judgment at the boundaries of the self. *Journal of Clinical Ethics, 18,* 256–261.

Lloyd, A., Hayes, P., Bell, P. R., & Naylor, A. R. (2001). The role of risk and benefit perception in informed consent for surgery. *Medical Decision Making, 21,* 141–149.

Reyna, V. F., & Hamilton, A. J. (2001). The importance of memory in informed consent for surgical risk. *Medical Decision Making, 21,* 152–155.

Veliz-Reissmüller, G., Agüero Torres, H., van der Linden, J., Lindblom, D., & Eriksdotter Jönhagen, M. (2007). Pre-operative mild cognitive dysfunction predicts risk for post-operative delirium after elective cardiac surgery. *Aging Clinical and Experimental Research, 19,* 172–177.

Utilities for Joint Health States

Utility is a quantitative expression of an individual's preference for a particular state of health under the condition of uncertainty. Similarly, utilities for joint health states are the quantitative expression of an individual's preference for having two or more health states or comorbid disease conditions, for example, a person with both asthma and diabetes. Same as utilities for single health state, utilities for joint health states are assessed on a scale where 0 represents death and 1 represents perfect health.

One important usefulness of utility assessment arises from the ability to compute quality-adjusted survival, measured in quality-adjusted life years (QALYs), which has gained increasing attention from clinical investigators to evaluate the treatment outcomes for medical decision making. QALY is achieved by multiplying quantity by quality of life, as measured by utility. For example, if a patient lives 1 year of life with utility .8, this amounts to $1 \times .8 = .8$ QALYs. If we were to measure the patient's utility routinely, until death, we could compute his or her quality-adjusted survival. If all patients were followed until death, we would sum the products of these quantity by quality of life values.

Purpose of Research

Standard catalogs of utilities have been developed, primarily for single health state or one disease condition, to facilitate the clinical decision analyses or treatment comparisons from a population perspective. The challenges arise when these off-the-shelf utilities are not readily available for those with joint health states or multiple disease conditions. There is no standard method to calculate the utility of comorbid disease conditions except by directly sampling from the population, which could be costly and time-consuming.

Existing Methods to Estimate Utilities for Joint Health States

There currently exists no gold standard method to deal with joint health states or comorbidities, with most being selected for convenience. That said, several simplistic estimators have been proposed to assess utilities for joint health states when only the separate utilities of the individual health states are available. The commonly suggested ones include multiplicative, minimum, and additive estimators.

Multiplicative Model

Given its appealing simplicity, the multiplicative model [$u(A\&B) = u(A) \times u(B)$] is probably the one that has been most commonly used. For example, if the average utilities for asthma and diabetes are .71 and .63, respectively, the utility for individuals with both asthma and diabetes would be estimated as $.71 \times .63 = .45$ based on the multiplicative model.

The multiplicative model has been empirically examined using the preference-based scores derived from Health Utility Index Mark 3 within a Canadian community population and the findings supported the use of the multiplicative model. It is worth noting that an extra step of rescaling, or "purification" as referred to in the study, was applied. This step was conducted by dividing the observed utilities by the utility of persons reporting no disease conditions to achieve the purified utilities. This step adjusted for health problems other than the disease

conditions being studied such as loss of functional health attributable to unknown factors. However, a difficulty with rescaling (purification) in medical decision making is that the utility of persons reporting no disease conditions may likely differ between a specified sample and the general population. That utility also may not be readily available for clinical investigators or decision analysts.

The multiplicative model was also examined using the preference-based scores derived from EQ-5D (EuroQol) within a U.S. national representative community-dwelling population. The results showed that multiplying the two utilities of single health states has a large difference from the utility of those who actually have both health states. Such a difference is larger than most of the other simplistic estimators with or without purification. The reason the multiplicative estimator does not empirically work well was explained that it is almost impossible for individuals to report all their comorbid disease conditions during a questionnaire survey. Therefore, even though a list of comorbid conditions is reported for certain individuals, it does not mean that such a list is comprehensive and exhaustive. That is to say, when multiplying two utilities together, researchers are likely using utilities of patients with more than one health state, and comorbid conditions are likely correlated. So multiplying two utilities is double counting because some patients naturally have both and not all have just one disease condition. A difficulty with any study in this context is that it is impossible to be sure that a patient has no other comorbidities beyond those queried and that those queried were measured without error.

Minimum Model

Minimum model [$u(A\&B) = \text{Min}(u(A), u(B))$] is another method that has been used to assess utilities for joint health states when only the separate utilities of the individual health states are available. For example, if the utilities for asthma and diabetes are .71 and .63, respectively, the recommended utility score assigned for individuals with both asthma and diabetes would be the minimum of .71 and .63, in this case .63, based on the minimum model.

It has been found that the minimum estimator has the smallest difference among a list of simplistic estimators, including the multiplicative and additive estimators from the same aforementioned study. However, none of the estimators was found unbiased from the utility of those who actually have both health states. Nonetheless, the identified bias for the minimum estimator is smaller than the minimally important difference of the preference-based EQ-5D index score ranging from .033 to .07 in existing literature. Those identified biases for both multiplicative and additive estimators are larger than such a range. Thus, the minimum estimator is still preferred to the multiplicative estimator for measuring utilities of joint health states with the empirical evidence from a U.S. national representative population.

Using the time trade-off utilities in prostate cancer, another study compared the multiplicative, minimum, and additive estimators and also found that the minimum estimator is the least biased and most efficient among the three. Similar to the previous study, none of the estimators provided unbiased results.

Additive Model

Researchers have also suggested the additive model [$u(A\&B) = u(A) + u(B) - 1$] to estimate the utilities of joint health states. To use the same example, if the utilities for asthma and diabetes are .71 and .63, respectively, the recommended utility score assigned for individuals with both asthma and diabetes would be $(.71 + .63 - 1) = .34$ based on the additive model.

The predictive values of the additive and the multiplicative models were compared using the standard gamble technique in a sample of patients with recurrent rectal cancer. It was found that the multiplicative estimator predicted the utility of joint health states better than the additive estimator. Such a finding was echoed by the aforementioned study. Both studies found that the additive estimator has a larger difference from the utility of those with both health states than the multiplicative estimator, although none of these two estimators is preferred.

Regression Models

The mapping method with regression models was used in recent literature to estimate the utilities

for joint health states when the utilities of the individual health states are available. This method is typically applied based on a large sample of population, nationally representative sample preferred, to get the marginal disutility for each disease condition. For example, if the utility for asthma is 0.71 and the marginal disutility for diabetes is −.035, the recommended utility assigned for individuals with both asthma and diabetes would be .71 + (−.035) = .675 based on this method.

The concept of marginal disutility associated with each disease condition is intuitively attractive. Using regression modeling, marginal disutility of a certain disease is the utility difference between patients with that disease and those without after controlling for other covariates in the model. A series of related studies have been conducted to identify the marginal disutilities for most of the common disease categories using the preference-based scores derived from EQ-5D (EuroQol) within a U.S. national representative community-dwelling population. The regressions can provide marginal disutilities for not only each disease condition but also other covariates that were controlled in the model.

However, this method has not been found to be widely used in medical decision making likely due to the following reasons. First, the estimated marginal disutilities can be sensitive to the appropriateness of the regression modeling. This includes both the choice of the regression model and the covariates that are controlled in the model. It is well established that utility score itself is far from being normally distributed. It has a ceiling at 1 with a significant number of people rating themselves in full health. Methods designed for continuous data, such as the ordinary least squares (OLS) regression, are often inappropriate for such data. Several other methods have been proposed including the Tobit model, the censored least absolute deviations estimator (CLAD), and the two-part model. The appropriateness of these models is still in need of further research. Additionally, the estimated marginal disutilities are likely to vary based on the selection of included covariates in the model. Second, while disease conditions and other covariates are adjusted for in regression models, statistical assumptions are made about how these conditions and covariates interact with one another. A simple linear regression can omit or oversimplify these interactions, which may bias the estimated marginal disutilities. Third, the estimated marginal disutilities are highly dependent on the data source that is used for the regression modeling. It is likely that even with the same model, different database may produce distinct marginal utilities for the same health state. Fourth, the method of marginal disutilities is not easy to implement, which discourages the use by clinical investigators.

Other Methods

Other methods that have been proposed include more robust modeling techniques to convert the single health state utilities into those for joint health states. For example, utilities may be statistically estimated as

$$u(A\&B) = \beta_0 + \beta_1 \times \text{Multiplicative}(u(A), u(B)) + \beta_2 \times \text{Min}(u(A), u(B)) + \beta_3 \times \text{Additive}(u(A), u(B)),$$

where the utility of joint health states A and B is a linear combination of the multiplicative, minimum, and additive estimators. The βs are the coefficients that need to be estimated from least square linear regressions. If any of the βs are not statistically significantly different from 0, the corresponding estimator can be omitted from the equation. This method uses the estimated β coefficients to construct correction factors, which may produce more flexible models than those simplistic models aforementioned. However, it also requires a wealth of data and is not straightforward to implement by clinical investigators.

Alex Z. Fu

See also Disutility; EuroQoL (EQ-5D); Expected Utility Theory; Health Utilities Index Mark 2 and 3 (HUI2, HUI3)

Further Readings

Dale, W., Basu, A., Elstain, A., & Meltzer, D. (2008). Predicting utility ratings for joint health states from single health states in prostate cancer: Empirical testing of 3 alternative theories. *Medical Decision Making, 28,* 102–112.

Drummond, M. F., Sculpher, M. J., Torrance, G. W., O'Brien, B. J., & Stoddart, G. L. (2005). Cost-utility

analysis. In *Methods for the economic evaluation of health care programmes* (3rd ed., pp. 137–209). Oxford, UK: Oxford University Press.

Flanagan, W., McIntosh, C. N., Petit, C. L., & Berthelot, J. (2006). Deriving utility scores for co-morbid conditions: A test of the multiplicative model for combining individual condition scores. *Population Health Metrics, 4,* 13.

Fu, A. Z., & Kattan, M. W. (2008). Utilities should not be multiplied: Evidence from the preference-based scores in the United States. *Medical Care, 46,* 984–990.

Li, L., & Fu, A. Z. (2008). Methodological issues with the analysis of preference-based EQ-5D index score. *Value in Health, 11,* A181.

Sullivan, P. W., & Ghushchyan, V. (2006). Preference-based EQ-5D index scores for chronic conditions in the United States. *Medical Decision Making, 26,* 410–420.

Sullivan, P. W., Lawrence, W. F., & Ghushchyan, V. (2005). A national catalog of preference-based scores for chronic conditions in the United States. *Medical Care, 43,* 736–749.

Utility Assessment Techniques

In medical decision making, a *utility* is a measure of the relative desirability that a person attributes to a particular health state, under conditions of risk. Two points are noteworthy here.

First, a utility is a relative measure. A person's utility for the health state of interest—the target health state—is revealed by comparing it with two extreme health states. One extreme is preidentified as highly undesirable and arbitrarily ascribed a utility of 0.00. The other extreme is preidentified as highly desirable and arbitrarily ascribed a utility of 1.00. Unless the person considers the target health state to be preferentially equivalent to either extreme, his utility for that state falls somewhere between 0.00 and 1.00.

Second, a utility is measured under conditions of risk. The assessment process involves asking the individual to choose between the certainty of the target health state or a hypothetical lottery with the probabilities p and $1 - p$ of entering, respectively, either the highly desirable extreme health state or the highly undesirable extreme health state. Therefore, the reported utility for the target health state incorporates the individual's personal attitudes toward the risks of gains and losses in health.

Utilities in Medical Decision Making

An investigator elicits and uses utilities in different ways depending on whether her ultimate research purpose is operating at the patient-population level or at the individual-patient level.

At the patient-population level, for example, utilities are computationally integral to the performance of formal medical decision analyses, to the conduct of cost-utility analyses, and to the derivation of quality-adjusted life years. The aggregated results of these computations then may be used in the subsequent development of evidence-based health policies and clinical guidelines.

Another example involves eliciting utilities for baseline and outcome health states from patients in clinical trials who have been randomized to receive different therapeutic interventions. Then across-time and across-group comparisons of the reported utilities can be used to make inferences about the interventions' relative effectiveness, in terms of the patients' own reported health-related quality of life.

A third example at the patient-population level occurs when a new, preference-based, population health index is under development. After specifying which health states will most likely be captured by her new index in the future, the investigator elicits utilities for these anticipated health states from large groups of people in the community. The observed utilities are then aggregated and incorporated into the new index, as the "weights" to use in future applications of the new index.

Utilities also come into play at the individual-patient level, when a clinical investigator is providing decision support for a patient. A patient may wish to be informed about the distributions of utilities other patients have reported for the health states they experienced after therapy and then use that information in arriving at his own treatment choice. On the other hand, an investigator could elicit the patient's own utilities for the different possible outcome states, incorporate those patient-specific utilities into an individualized decision tree, and then use formal decision analysis to identify the treatment that would have the highest *expected utility* for that patient.

Standard Gamble Technique

The standard gamble technique, devised by John von Neumann and Oskar Morgenstern, is an application of expected utility theory. Because it is based on explicit axioms of rationality, it is often cited as the criterion method for eliciting utilities for health states.

When an investigator wishes to elicit a rater's standard gamble utility for a particular target health state, her elicitation strategy depends on her ultimate research purpose. One strategy is used if the target health state is one of several outcome health states on a decision analytic tree, and the investigator needs the rater's utility for each of these *imagined* health states. A different strategy is used if the target is the rater's baseline health state when he is first entering a clinical trial, and the investigator needs the rater's utility for this current *experienced* health state.

Imagined Health States

Preliminary Steps

The investigator first provides the rater with a number of cards. Each card bears a description of one of the relevant health states. Suppose there are five health states in the set: (1) K (mild symptoms and functional problems), (2) L (moderate symptoms/problems), (3) M (severe symptoms/problems), (4) Perfect Health, and (5) Immediate Death. The states Perfect Health and Immediate Death are customarily included when the investigator needs to compare her observed utility distributions with those reported by other investigators; for such across-study comparisons, the raters in the different studies must have reported their utilities relative to the same two extreme health states.

The rater is asked to rank order the set of health states from the most preferred to the least preferred. Suppose he ranks them as follows: Perfect Health, K, L, M, and Immediate Death. The investigator then arbitrarily assigns the utilities 1.00 and 0.00 to the top- and bottom-ranked health states, respectively. Now the challenge is to elicit the rater's standard gamble utilities for K, L, and M, relative to Perfect Health (utility = 1.00) and Immediate Death (utility = 0.00).

The investigator first uses imaginary money lotteries to illustrate the standard gamble technique. The rater is repeatedly asked whether he would pay a particular amount of money to buy a ticket for a lottery with particular chances of either winning a larger amount of money or losing the money he would pay for the ticket. Each time the question is asked, the cost of the lottery ticket is kept constant, but the chances of winning/losing are systematically varied. Often, visual aids such as probability wheels are used to illustrate these shifts in the probabilities. Once the rater indicates that he understands the logic involved in choosing between a certainty and a lottery with a particular set of probabilities, he is ready to use the standard gamble technique to reveal his utilities for K, L, and M.

Assessing the Utility of a Single Imagined Health State

Suppose the investigator begins with health state L. She asks the rater to imagine the certainty of living in health state L for the rest of his natural life. Formal life expectancy tables can be used to specify a reasonable lifetime duration for a rater of a particular age. For illustrative purposes, suppose a rater has an estimated remaining life expectancy of 30 years.

Next, the rater is asked to imagine a hypothetical situation in which he is asked to choose between living for the next 30 years in health state L or accepting an imaginary lottery. The lottery has two possible outcomes. The "best" outcome is the top-ranked health state—Perfect Health—which has already been arbitrarily assigned a utility of 1.00. The "worst" outcome is the bottom-ranked health state—Immediate Death—which has already been arbitrarily assigned a utility of 0.00.

This lottery is initially presented to the rater with the probability (p) of Perfect Health = 1.00 (or 100%) and the counterpart probability ($1 - p$) of Immediate Death = 0.00 (or 0%). Under these conditions, this is a "dominated" choice. This initial version of the lottery is, in effect, an assurance of a lifetime of excellent health, and the rater's rationally logical choice is to opt for the lottery.

Then the investigator asks the rater to suppose that the lottery's probability (p) of Perfect Health is lowered from 1.00 to .95, while the probability ($1 - p$) of Immediate Death is raised from .00 to .05 (again, visual aids such as probability wheels

are used to illustrate these probabilities, which also can be expressed as percentages). The rater is asked to choose between the certain health state L and this new lottery. Suppose the rater considers health state L—with its moderately severe symptoms and functional problems—to be quite undesirable, and he again chooses the lottery. He is indicating that, to have a .95 chance to be in excellent health instead of health state L, he is willing to accept a .05 risk of death.

Next, the investigator asks the rater to suppose that the lottery's probability (p) of Perfect Health is lowered further, from .95 to .90, and the probability ($1 - p$) of Immediate Death is, by corollary, raised from .05 to .10. Again, the rater is asked to choose. Suppose he again chooses the lottery; he's indicating that, to have a .90 chance to be in excellent health instead of health state L, he's willing to accept a .10 risk of death.

In this manner, the probabilities in the lottery are systematically altered in 5% decrements/increments, until an *indifference point* is reached at which the rater cannot choose between the certainty of continued life in state L and the lottery. Suppose this point is reached when the lottery's chance for Perfect Health has decreased to .75 and its risk of Immediate Death has increased to .25. (The investigator may alter the values of p and $1 - p$ in 1% increments/decrements around this point, in an attempt to precisely estimate the indifference point. This refinement is referred to as *ping-ponging*.)

Any lottery's expected utility = (p of the "best outcome")(utility of the "best outcome") + ($1 - p$ of the "worst outcome")(utility of the "worst outcome"). Therefore, the lottery at this rater's indifference point has an expected utility of (.75)(1.00) + (.25)(0.00) = .75. According to the axioms of rationality that underlie the standard gamble, *the expected utility of the lottery at the rater's indifference point* is, by substitution, *this rater's utility for health state L*, because the rater is indicating that this particular lottery and health state L are preferentially equivalent. Therefore, in our example, the rater's standard gamble utility for health state L over the next 30 years is equivalent to .75.

Then the investigator elicits the rater's utilities for K and M, the remaining target health states in the set, by repeating each of the steps outlined above.

If the investigator is working at the individual-patient level, the utilities for K, L, and M would be incorporated into the underlying decision analytic tree to reveal the therapeutic option with the highest expected utility for that particular patient. If the investigator is working at the patient-population level, she would carry out the entire procedure with each of the raters in the study sample, generate a distribution of utilities for each outcome health state, incorporate the "average" utilities into the underlying decision analytic tree, and thereby reveal the therapeutic option with the highest expected utility for that patient population.

Experienced Health States

Preliminary Steps

The investigator begins by asking the rater to report his current status on a set of relevant symptomatic and functional dimensions (Y_1, Y_2, Y_3, Y_4). Then these self-reports are compiled into an overall description of his current experienced health state, X.

Before launching into the standard gamble technique, the investigator illustrates the technique using questions about imaginary money lotteries, as outlined above. Once the rater has indicated that he understands the logic involved in choosing between a certainty and a lottery with a particular set of probabilities, he is ready to use the standard gamble technique to reveal his utility for his current experienced health state, X.

Assessing the Utility of a Single Experienced Health State

The investigator begins by asking the rater to imagine the certainty of living in health state X for the rest of his natural life (in this example, for 30 years). Then she asks him to imagine a hypothetical situation in which he must choose between that certainty or an imaginary lottery with two possible outcomes.

If the investigator plans to eventually compare her rater's utilities for his current experienced health state with those reported in other studies, she may need to use Perfect Health (with an arbitrarily assigned utility of 1.00) and Immediate Death (with an arbitrarily assigned utility of 0.00) as the lottery outcomes. However, if across-study

comparisons are not scientifically indicated, and it is only necessary to ensure that the utility assessment procedures are internally consistent, she may elect to use different win/lose outcomes in the hypothetical lottery. For example, the best lottery outcome could be a health state in which there are no problems on any of the dimensions Y_1, Y_2, Y_3, and Y_4 (with an arbitrarily assigned utility of 1.00), and the worst lottery outcome could be a health state in which there are extremely severe problems on all the dimensions Y_1, Y_2, Y_3, and Y_4 (with an arbitrarily assigned utility of 0.00). This strategy offers two advantages. First, replacing Perfect Health as the best lottery outcome reduces the likelihood of eliciting confounded utilities when the raters are actually experiencing two or more comorbid conditions. Second, replacing Immediate Death as the worst lottery outcome reduces the likelihood of eliciting invariant, skewed utility distributions due to the rater's aversion to gambles involving death.

In any case, the lottery is initially presented to the rater as a dominated choice, with the probability (p) of the best outcome = 1.00 (or 100%) and the counterpart probability ($1 - p$) of the worst outcome = 0.00 (or 0%). Then the investigator proceeds as outlined above. The rater is repeatedly asked to choose between the certainty of living for the rest of his natural life in his current experienced health state and different lotteries. Appropriate visual aids are used to illustrate lotteries in which the probability (p) of the best outcome is lowered from 1.00 in 5% decrements, while the probability ($1 - p$) of the worst outcome is raised from 0.00 in 5% increments. Eventually, the rater's indifference point is identified, the expected utility of the lottery at this point is computed, and the utility for the rater's current experienced health state is thereby inferred.

The investigator could proceed to elicit utilities for the current health states experienced by the other raters in the study. The particular timing of these utility assessments would depend on her overall research purpose. For example, in a clinical trial, the research purpose may be to compare treatments in terms of their effects on patients' health-related quality of life. This purpose could be met by repeatedly assessing the utilities for raters' current experienced health states as they progress through treatment and then comparing the observed distributions in terms of across-group and across-time differences.

In another example, the research purpose may be to investigate the raters' ability to accurately predict their utilities for future experienced health states. This purpose could be met by asking the raters to describe the health states that they anticipate entering after their therapy, assessing their utilities for these anticipated health states, then following the raters through their therapy to see not only whether they entered their anticipated health state but also whether their utilities for this state are the same once it actually becomes an experienced health state.

Other Assessment Techniques

Two other techniques—the time trade-off and the rating scale—are often used to elicit evaluations for health states; occasionally, their results are also referred to as utilities. Some purists argue that responses collected on rating scales should be referred to as *rating scale scores,* that responses to the time trade-off technique should be qualified as *time trade-off utilities,* and that the term *utilities* should be used only to refer to evaluative scores obtained using the standard gamble technique.

Time Trade-Off Technique

The time trade-off technique was developed by George Torrance, as an alternative, perhaps cognitively easier, strategy for eliciting evaluative scores for health states. However, note that, unlike the standard gamble technique, the time trade-off technique operates under nonrisky, trade-off conditions of measurement.

As in the standard gamble technique, the investigator begins by presenting the rater with a description of the target imagined or experienced health state, *Y*. The rater is asked to consider a hypothetical situation in which he has to choose between the certainty of living either in health state *Y* or in Perfect Health for the rest of his natural life (e.g., for 30 years). Under these conditions, this is a dominated choice; the rater's logical choice is to opt for 30 years in Perfect Health.

Then the investigator asks the rater to imagine another hypothetical situation, in which he has to choose between either 30 years in health state *Y* or 25 years in Perfect Health. Suppose he again chooses Perfect Health. In effect, he is indicating that to be

in excellent health instead of health state *Y* for the rest of his natural life, he's willing to give up—or trade off—5 years of his lifetime expectancy.

Next, the investigator asks the rater to imagine that he has to choose between either 30 years in health state *Y* or 20 years in Perfect Health. Suppose the rater considers health state *Y* to be quite undesirable, and he again opts for a shorter expected lifetime in Perfect Health. The time trade-off technique proceeds by systematically reducing the hypothetical lifetime in Perfect Health in 5-year decrements, while maintaining the lifetime in health state *Y* at 30 years. (Often, visual aids such as sliding scales are used to illustrate how the expected lifetime in health state *Y* remains constant, while the expected lifetime in Perfect Health is steadily reduced.)

As the time trade-off technique continues, the choice becomes more and more difficult, in that the rater becomes less and less willing to trade off any further years in lifetime expectancy in Perfect Health. Suppose that, when the hypothetical lifetime in Perfect Health has been reduced to 20 years, he cannot choose between this option and the alternative of a full 30 years in health state *Y*. (The investigator may alter the time at the 20-year point in 1-year increments/decrements, in an attempt to precisely estimate the indifference point.) At this point, the rater is indicating that, for him, 30 years in health state *Y* is preferentially equivalent to 20 years in Perfect Health, and this rater's time trade-off utility for *Y* is computed as 20/30 = .67.

Rating Scale Techniques

In rating scale techniques, the rater directly assigns an evaluative score to the target health state, in terms of its desirability/undesirability relative to two extreme health states that "anchor" the ends of a scale. Note that, unlike either the standard gamble or the time trade-off technique, rating scale techniques operate under nonrisky, non-trade-off conditions of measurement. Hence, methodological purists argue that the term *utilities* should not be applied to scores obtained using rating scale techniques.

The 10-point rating scale and the 10-cm linear analog scale are two examples of rating scale techniques. Here, the techniques' general principles are illustrated using the 10-cm linear analog scale.

As in the standard gamble and the time trade-off techniques, the investigator begins by presenting the rater with a description of an imagined or experienced health state, *Y*. The rater is asked to imagine living in health state *Y* for the rest of his natural life (e.g., for 30 years). Then the rater is asked to consider a 10-cm horizontal line that's anchored by a highly undesirable state (e.g., Death, with a preassigned value of 0.00) and a highly desirable state (e.g., 30 years in Perfect Health, with a preassigned value of 100). The rater is asked to make a vertical mark across the horizontal line at the point that indicates his opinion about the desirability/undesirability of living in health state *Y* for 30 years, relative to these two extremes. The rating scale score for health state *Y* is then determined by measuring the distance, in millimeters, from the zero-valued end of the linear analog scale to the rater's mark—yielding, for example, a rating scale score of 65.

Hilary A. Llewellyn-Thomas

See also Bounded Rationality and Emotions; Cost-Benefit Analysis; Decision Analyses, Common Errors Made in Conducting; Decision Tree: Introduction; Decision Trees, Evaluation; Expected Utility Theory; Health Utilities Index Mark 2 and 3 (HUI2, HUI3); Quality-Adjusted Life Years (QALYs); Shared Decision Making

Further Readings

Keeney, R. L., & Raiffa, H. (1976). *Decisions with multiple objectives: Preference and value trade-offs.* New York: Wiley.

Llewellyn-Thomas, H. A., Sutherland, H. J., Tibshirani, R., Ciampi, A., Till, J. E., & Boyd, N. F. (1982). The measurement of patients' values in medicine. *Medical Decision Making, 2,* 449–462.

Patrick, D. L., Bush, J. W., & Chen, M. M. (1973). Methods for measuring levels of well-being for a health status index. *Health Services Research, 8,* 228–245.

Read, J. L., Quinn, R. J., Berwick, D. M., Fineberg, H. V., & Weinstein, M. C. (1984). Preferences for health outcomes: Comparison of assessment methods. *Medical Decision Making, 4,* 315–329.

Torrance, G. W. (1976). Social preferences for health states: An empirical evaluation of three measurement techniques. *Socio-Economic Planning Sciences, 10,* 129–136.

Torrance, G. W., Boyle, M. H., & Horwood, S. P. (1982). Application of multi-attribute utility theory to measure social preferences for health states. *Operations Research, 30,* 1043–1069.

von Neumann, J., & Morgenstern, O. (1953). *The theory of games and economic behavior.* New York: Wiley.

Weinstein, M. C., & Fineberg, H. V. (1980). *Clinical decision analysis.* Philadelphia: W. B. Saunders.

V

Value-Based Insurance Design

A value-based insurance design is a health insurance policy where selection of services and payment for services are decided by evaluating clinical benefit as well as cost, where decision making is condition specific if not patient specific, and where the evaluation process takes into consideration multiple time periods. It is not the value of the service to the provider, the clinical developer, or the payers that matters but the clinical benefit and costs to the patient and society. Value-based insurance design offers a potential solution to the healthcare-financing crisis. The concept of value, defined as the clinical benefit achieved for the money spent, is largely absent from most insurance policies and design discussions but is perhaps the most important aspect of insurance for purchasers and consumers. This entry discusses what is meant by insurance design and what it means to have value as the basis for insurance design decision making. It then considers who it is that decides value and the evidence base for their decisions.

Insurance Design

Insurance design is a wide-ranging concept that encompasses selection of services that may be covered (physician services, hospital services, specific drugs), the level of payment for these services for patients (full, a percentage, all except for a copayment), the level of payment for these services for providers (full, a percentage, some predetermined amount), how providers are paid (prospective, retrospective), which providers are eligible for payment (exclusive network, preferred network, open access), and funding (individual savings accounts, employer self-funded, insured). In some cases, insurance design issues are packaged in ways that aren't easily separable. The simplest form of traditional insurance offers full payment to providers for all services from whomever the patient sees, with the patient being responsible for an initial amount (deductible) and a percentage of the remaining expenses. At the other end of the spectrum, the prototype managed-care organization is a health maintenance organization that offers a wide range of services under predetermined payments to providers (capitation) and modest copayments on the part of patients within a closed panel of providers on a fully insured basis. The offerings by insurance companies often include choices along each dimension of design.

With the subset of insurance design issues that are before the consumer—selection of services and payment for these services—there are notable philosophical differences about the most appropriate decision-making process. The wave of managed care that grew during the 1980s was associated with greater payer oversight on covered services and lower cost sharing on the part of the consumer. By exerting oversight through both market share and payment options, managed-care organizations achieved substantial cost savings. However, the push-back from the decision making by managed-care organizations was pronounced during the early 2000s. Providers and many consumers

and consumer rights organizations didn't appreciate oversight, at times for good reasons, which is partly the reason why more than 30 states have enacted patient's bill-of-rights legislation.

Of course, managed care has not completely gone away. Managed care is still the modal form of insurance coverage in the United States, with there being more than a one-third market share of preferred provider organization coverage and a one-quarter market share of health maintenance organization coverage. Within Medicaid, managed care has a two-thirds market share and is growing. Cost savings associated with Medicaid managed care are mostly associated with control of inpatient hospitalization and prescription drugs. Consumers have pushed away from the stricter forms of decision making exhibited by commercial managed-care organizations, but voters and regulators permit such decision making for government-sponsored programs.

Lacking the cost control of managed care, decision making by insurance companies has reverted to consumer- or patient-oriented cost control, largely through increased cost sharing. Proponents of what are now called consumer-directed health plans suggest that it is the consumer who is in the best position to decide what services he or she wants and it should be the consumer making these decisions at point-of-care. Going beyond simply having higher coinsurance rates and copayments, the development of health savings account and health reimbursement arrangements to accompany high-deductible health plans has given consumers the opportunity to make decisions about specific services and decisions about how they allocate healthcare dollars over time.

The Bases for Insurance Design

Early designs in health insurance were largely provider based. Blue Cross for hospital services and Blue Shield for physician services played an important role in shaping commercial health insurance and were the models for Medicare. Given all the changes in the ownership and structures of Blue Cross and Blue Shield plans, the provider focus of insurance design is less apparent today. Some provider-owned managed-care organizations may also have organizational support as a motivation for insurance design decisions.

An extension of provider-based insurance design is clinical-service-based benefit design. When a clinician develops a new service, an ideal response might be expanded insurance coverage to include the new service. The rationale for coverage of many physician services and procedures is the recognition of a service as having an associated CPT-4® code.

For some insurance plans, approval by the Food and Drug Administration (FDA) is a sufficient basis for coverage of a new drug—a clear example of a clinical-service-based design. Interestingly, few persons with employer-sponsored or government-sponsored insurance have insurance plans for prescription drugs that are completely clinical service based. Merely having FDA approval is no longer sufficient to garner coverage. Rather, decisions about which drug to cover (on a tier in a formulary) and how much the patient will be required to pay (typically a higher amount for drugs on higher tiers) are made on the basis of employer demands and provider costs. The complexity of provider cost structures (with rebates and other arrangements) might imply that there is a provider (manufacturer) basis to insurance design, but the plans that appeared to be more directly provider based (e.g., owned by pharmaceutical companies) have largely been disbanded.

An advance in clinical-service-based design is clinical-benefit-based design, which takes into consideration the clinical benefit to the patient. Just because a service is possible and has demonstrated safety doesn't mean that it is clinically beneficial to one or many patients. Most health plans would likely argue that their decision making is based on clinical benefits. However, costs are a component of most health plans' decision-making process, and clinical benefit is not always defined in a very specific manner. Just because a service is of clinical value to one person, or even on average, doesn't mean that it is of clinical benefit to all persons. Going beyond making coverage decisions on services, an effective clinical benefit design would facilitate making benefit design decisions on combinations of services and patient characteristics.

Counter to the foundations of health insurance rooted in provider- and clinical-based designs is a cost-based insurance design. Indeed, one of the motivations for the managed-care backlash was the concern that the oversight

imposed by managed-care organizations was not just on clinical grounds but was also focused on costs. Going with the least expensive option may not be in the best, long-run interest of everyone involved in healthcare.

As an alternative to, or the rational evolution of, provider-, clinical-service-, clinical-benefit-, and cost-based designs, some are suggesting value-based insurance designs. The key ideas behind a value-based insurance design are as follows: (a) Selection of services and payment for services are decided by evaluating clinical benefit as well as cost, (b) decision making should be condition specific if not patient specific, and (c) evaluation processes should take into consideration multiple time periods—permitting consideration of some services as investments in health that may pay off in the future. It is not the value of the service to the provider, the clinical developer, and the payers but the clinical benefit and costs to the patient and society that matter.

Yet one more step in the evolution of benefit design would be to take into consideration the site of care in the purchasing decision. So-called *value-based purchasing* is now being widely discussed in the purchaser community. Just as employers seek the best vendors for all other inputs, they are increasingly seeking the best prices from the best providers of healthcare services. Still, seeking best pricing without having first defined which services one really values and wants to purchase may lead to suboptimal purchasing behavior.

The Evidence for Benefit Design

Making benefit design decisions based on clinical value and costs at the condition or patient level has substantial intuitive appeal. Beyond the initial appeal, however, decision making based on value requires both a process for assigning value to a service for an individual and a body of evidence to support the process. Both processes and evidence are in the developmental stage. There are two processes in current practice that seek to apply value-based insurance design. The first process targets services that are well-known to be clinically valuable and reduce consumer costs (deductibles or copayments) for these services. Although these services may provide substantial benefit for some users and provide less value for other patients, the process does not attempt to differentiate between these patients.

The second process, requiring more sophisticated data systems to implement, creates differential coverage based on patients' characteristics. Programs using this approach typically identify patients with specific diseases, such as diabetes or coronary heart disease, for which there is a good body of evidence, and reduce copayments for only high-value services for patients having these diseases. The targeting of high-value services and provision of improved coverage for specific groups of patients are elements of value-based insurance design, lacking only a clear consideration of the time dimension.

Several firms are experimenting with one of these two processes for value-based insurance design. Pitney Bowes is well known for using the first process and reducing copayments for all users of drugs commonly prescribed for diabetes, asthma, and hypertension. Companies such as ActiveHealth Management have created systems that go further by identifying patients for whom benefits would be greatest and creating communications that target these patients.

The University of Michigan Center for Value-Based Insurance Design was established in 2005 to develop, evaluate, and promote value-based insurance initiatives that achieve improvements in health outcomes and contain healthcare costs. The center is the first academic venue in which faculty with both clinical and economic expertise conduct empirical research to determine the health and economic impact of innovative benefit designs.

A current experiment created and being evaluated by the center involves targeting persons with diabetes for reduced cost sharing to increase compliance with physician recommendations for treatment. This experiment has both a study and a control group, permitting a formal mechanism to evaluate the effectiveness of a value-based insurance design. This experiment is the first prospective, controlled evaluation of these more nuanced benefit designs. The currently ongoing experiment is the first of its kind designed to improve the quality of care for persons while allowing for a rigorous evaluation. Many more such experiments will be required to provide evidence of the appropriateness of value-based insurance designs.

Dean G. Smith

See also Consumer-Directed Health Plans; Decisions Faced by Nongovernment Payers of Healthcare: Indemnity Products; Decisions Faced by Nongovernment Payers of Healthcare: Managed Care; Evidence-Based Medicine

Further Readings

Chernew, M. E., & Fendrick, A. M. (2008). Value and increased cost sharing in the American health care system. *HSR: Health Services Research, 43,* 451–457.

Chernew, M. E., Rosen, A. B., & Fendrick, A. M. (2007). Value-based insurance design. *Health Affairs, 26,* w195–w203.

Fendrick, A. M., Smith, D. G., Chernew, M. E., & Shah, S. N. (2001). A benefit-based copay for prescription drugs: Patient contribution based on total benefits, not drug acquisition cost. *American Journal of Managed Care, 7,* 861–867.

Rosen, A. B., Hamel, M. B., Weinstein, M. C., Cutler, D., Fendrick, A. M., & Vijan, S. (2005). Cost effectiveness of full Medicare coverage of angiotensin-converting enzyme inhibitors for beneficiaries with diabetes. *Annals of Internal Medicine, 143,* 89–99.

Value Functions in Domains of Gains and Losses

Many medical situations incorporating gains and losses involve exchange of present-day costs for future benefits. In our era of chronic illness, with antecedents often distant in time, decision making by physicians or patients is frequently influenced in a significant manner by the principles and practice of discounting. *Positive discounting* is the term for diminishing value over time. A more diversified use becomes important as we attempt to refine decision models. Temporal effects in the medical preventive context include overeating, smoking, and drinking. Another example is investment in calcium intake or vitamin D to prevent fractures because of osteoporosis.

These are situations where there is a choice between immediate pleasures and the future benefits of good health. The benefits of behavioral change often occur so far in the future that they seem of little value relative to the immediate costs. If individuals prefer to live for the present rather than save for the future, such a preference may not yield the best lifetime outcomes for the individual.

Studies of gains and losses and time preference in medical decision making have dealt with either life saving or health change. Different methods have been used in the two areas to find individuals' time preference rate, however, and this complicates comparisons between these areas.

Research on decision making under uncertainty has focused largely on expected utility and behaviors that violate this axiom. Intertemporal choices—that is, choices between something now and something later—have been investigated less. Discounted utility, however, is in need of a model that takes into account anomalies and framing factors in medical decision making. These effects are a challenge to normative theory. In spite of the difference between economic and medical decisions, both domains have revealed discounting biases such as magnitude effect, dynamic inconsistency, instant endowment and status quo bias, sequence effect, sign effect, and loss aversion.

Discounting Biases

Magnitude Effect

Magnitude effect signifies that discount rates are proportionally lower for large magnitude outcomes. If a magnitude effect occurs also in medical decisions, this suggests that the value of future health outcomes would increase if these outcomes were seen as important or large. When it comes to health behavior, the value range that is discounted is often not defined in an absolute sense. Value discounting is therefore uncertain and unpredictable. In a study by Gretchen Chapman and colleagues, a magnitude by domain interaction indicated a larger magnitude effect on health than on money. The range of health outcomes (1–8 years of full health) had a greater effect on discount rates than did the range of monetary outcomes. Discounting of lives saved in future generations has in the earlier study obtained estimated annual discount rates of 25% for the 20-year time horizon, 12% for the 50-year horizon, and 8% for the 100-year horizon, and these results support a magnitude effect.

Dynamic-Inconsistency Effect

Discount rates have also been found to be proportionally lower for time intervals that are relatively more distant, challenging the assumption that discount rates are constant over time. Framing health messages as long-term outcome might, therefore, diminish the discounting effect. The pattern of dynamic inconsistency implies that people will reverse their preference over time. An interaction between domain and delay had earlier been found, with delay decreasing the differences between health and money. With long delays (e.g., 12 years), discount rates for health and money were found to be equivalent.

Instant Endowment and Status Quo Bias

Instant endowment is a concept meaning that the loss of utility associated with giving up a valued good is greater than the utility gain associated with receiving the good. Status quo bias means a preference for no changes in decision making. One example is court decisions where compensations are more likely for out-of-pocket costs than for unrealized profits. One example of a status quo bias is that new medical plans are more often chosen by new employees than by employees who have started that work before the plan was introduced and that small changes are preferred to large ones.

Sequence Effect

Health behavior could be regarded as a series of short-term choices where each one seems to be a good optimal choice but together they give an undesirable result. Most analyses of judgments and decisions are based on single-period models or models that assume time-separable utility. Models have so far been applicable mostly to short-range decisions with simple outcomes in terms of gains or losses. Judgments involving long-term planning seem not to have been investigated to the same extent.

Many medical decisions could be regarded as sequences of choices where the whole picture is not perceived and when, at a certain point in the sequence, there could be a choice between two alternatives: smoking or not smoking, for instance, or taking the drug or not. The optimal sequencing of drugs in treating a disease such as rheumatoid arthritis is one example where this issue is of importance. Expected values for response might vary with position in the sequence. A theoretical point should thus be made about the possible relevance of the time sequence of interventions when it alters the context of expected gains and losses.

Sign Effect

Differences have been found between decisions related to losses and to gains; such differences are termed sign effect. Many projects have studied the efficacy and the toxicity of drugs; in treating a disease, efficacy can be regarded as a gain and toxicity as a loss. Good health is regarded as a gain, and poor health is generally regarded as a loss. Discount rates for losses are generally lower than discount rates for gains. Consequently, research on the sign effect in medical judgments and decisions would suggest that discount rates for negative health states would be lower than those for positive health states.

Discount rates have also been found to vary between different health settings, thus violating the assumption of identical discounting in all settings. This finding can be related to whether the health state is presented as a gain or a loss. Usually, gains (full health) have been studied. Discount rates for adverse health states such as blindness and depression have been found to be quite low, averaging 3% annually. One practical consequence is that a strong asymmetry between gains and losses will give a low willingness to undertake actions to improve health. This highlights a very important aspect of health-related behavior.

Diminishing sensitivity has been noted, as the marginal value of both gains and losses decreases with their size. Furthermore, the marginal value has been found to decrease with distance from the reference point (Figure 1). Reference levels seem to be adopted in a context. Their impact on decisions can be arbitrary when the norm for decisions with long-term consequences is being set. Moreover, the initial response to these consequences might be of relatively small importance if adaptation induces a shift of reference. These changes in adaptation level and the shift of reference could be further investigated in studies involving gains and losses in medical decision making.

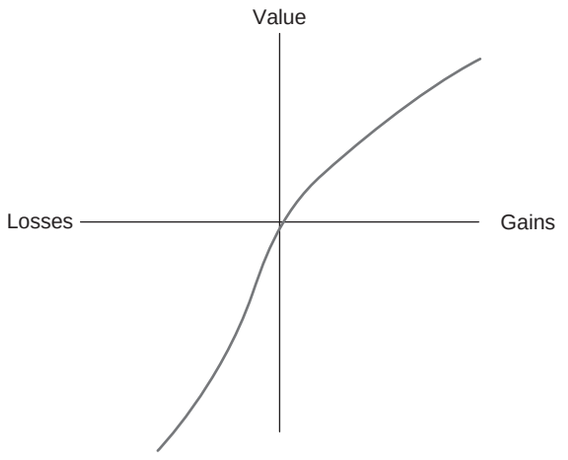


Figure 1 Value function of domains of gains and losses

Judgments and decisions should also be characterized by invariance of preference over response modes, with only the attributes and features determining preferences toward objects. Choice-choice reversals might be due to whether alternatives are presented as gains or losses relative to the reference. Responses to changes might be more intense for unfavorable situations such as losses than for changes leading to improvements. Therefore, when making medical decisions, it is advisable to give greater weight to negative than to positive consequences. Judgment reversals might also imply that with the same expected value, a safe bet is preferred to a long shot, but some people prefer the long shot. This lack of invariance causes a problem in prescribing procedures for decision analysis.

In the past, gains have usually been studied. Different behaviors might be experienced by the individual as a loss, especially related to a disease. Deteriorating in the disease is a loss. Recovering from a disease or avoiding progression in the disease is a gain. A speculative conclusion suggested here is that probabilities are perceived in different ways depending on the sign of the outcome—that is, whether it is a gain or a loss. An interaction between sign and perception of probability can be assumed. The perceived distribution of loss and gain can also vary by sign, with varying discounting effects over time. Future studies need to take this into consideration.

Loss Aversion

Loss aversion is related to the sign effect. Work in decision theory indicates that potential losses have a greater impact on preferences than do gains. Smoking, eating unhealthy food, and not exercising could give a later loss, such as getting cardiovascular disease, whereas sound health habits could give a gain. The effect of loss aversion applied to the addictive behavior of smoking means that it could be expected to vary with the gain of a smoke-free life and the loss of the immediate satisfaction of smoking. For overeating, the behavior would vary with the gain of a slim body and the loss of the immediate satisfaction of eating. Loss aversion has implications for medical decisions and has been found to be more pronounced for safety than for money. Speculatively, loss aversion also might imply a status quo bias.

Moreover, an asymmetry in risk aversion has been found by Daniel Kahneman and Amos Tversky in economics. Certain gains were preferred to uncertain gains; but for losses, uncertainty was preferred to certainty. Such asymmetry, when applied to medical decisions, may be important. In health, the immediate action might yield a loss with a high degree of certainty. For the individual, though, the gain in the future is uncertain despite its being predictable for the population as a whole.

Implications in Medical Decision Making

Many judgments and decisions are made daily in clinical work. The issues already described in this entry stress the importance of problems related to gains and losses in influencing medical decisions. These issues have relevance, for example, in treating patients with rheumatoid arthritis, which is a lifelong disease. What constitutes the gain and the loss from the treatment?

The probability of the patient enduring the adverse complications of a treatment will be greater if gains and losses are distributed in the best way over time. The gains of the treatment cannot be too distant in time, though, if discounting effects are considered. What is the optimal perceived distribution of different outcome variables such as pain, disability, and death? Does the patient prefer pain now or at some later time? All

these factors are potentially decisive for medical decisions and treatment choices.

Are treatments the choice of the physician, the patient, or both? How do the patients and doctors estimate the different variables? The doctor should follow expected value combinations more closely, while the patient might frequently be risk aversive. The importance of shared decision making, where patients and providers consider gains and losses and reach a healthcare decision based on mutual agreement, is increasingly stressed.

There may be differences in value functions of gains and losses for physical and mental health states and differences for different health states such as osteoporosis, diabetes, cancer, and myocardial infarct. The differing functions might potentially affect the treatment recommended by the physician. They might also affect the way the patient perceives and complies with the treatment. Doctors have also been found to exhibit both interindividual and intraindividual variation in judgments.

Time on a drug could also be related to the losses and gains affected by treatment. Some studies have found toxicity to be the most common reason to discontinue methotrexate for rheumatoid arthritis. The shorter duration with a drug found in a study for patients with the most negative initial health state might be explained by a difference between the negative effect of toxicity and the positive effect of efficacy on judgments performed in clinical decision making. Treatment choices could also be explained by this difference between efficacy and toxicity related to time. A drug's efficacy might have less impact than its toxicity on the length of time in a drug treatment and its discontinuation, regardless of constant improvement in disease variables such as disability. Furthermore, toxicity and efficacy may have different impacts at different periods. Toxicity might have a greater impact in the earlier periods, causing patients to leave the drug early.

One time-related explanation for change of treatment, in connection with gains and losses, is that patients expect not a constant level of disability but increasing improvement. This expectation disregards the failure of the disease process to impair, which is a positive outcome in itself. There might be an adaptation by time to the level of disability, with expectations and decisions for future treatment being based on the earlier experience of treatment. There might be a discounting effect operating, in which the present disability level matters more than a treatment's potential future positive effects on disability.

Diabetes mellitus, like rheumatoid arthritis, is a chronic disease requiring lifelong treatment, and there is a high risk of constantly deteriorating health in several respects. The patient with type 1 diabetes must inject insulin several times daily and maintain a strict diet. Despite strict treatment, the result might be only to decrease the slope of decrement, with a negative outcome in the disease process itself. The effects of the disease process might be greater than the effects of treatment, but without the treatment, the negative effects would be greater still. The medical necessity of the treatment motivates carrying on with it in spite of its adverse effects. Thus, over a very long period, the incentives might be in terms more of losses than of gains.

As the issues raised above indicate, medical judgments and decisions are complicated. However, discussions of value functions in clinical judgment and health policy have been fairly rare to date.

Monica Ortendahl

See also Discounting; Gain/Loss Framing Effects; Shared Decision Making

Further Readings

Baker, F., Johnson, M. W., & Bickel, W. K. (2003). Delay discounting in current and never-before cigarette smokers: Similarities and differences across commodity, sign, and magnitude. *Journal of Abnormal Psychology, 112,* 382–392.

Chapman, G. B., Brewer, N. T., Coups, E. J., Brownlee, S., Leventhal, H., & Leventhal, E. A. (2001). Value for the future and preventive health behavior. *Journal of Experimental Psychology Applied, 7,* 235–250.

Jan, S. (2003). Why does economic analysis in health care not get implemented more? Towards a greater understanding of the rules of the game and the costs of decision making. *Applied Health Economics and Health Policy, 2,* 17–24.

Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica, 47,* 263–292.

Khwaja, A., Silverman, D., & Sloan, F. (2007). Time preference, time discounting, and smoking decisions. *Journal of Health Economics, 26,* 927–949.

Ortendahl, M. (2007). Shared decision-making based on different features of risk in the context of diabetes mellitus and rheumatoid arthritis. *Therapeutics and Clinical Risk Management, 3,* 1–6.

Ortendahl, M., & Fries, J. F. (2005). Framing health messages based on anomalies in time preference. *Medical Science Monitor, 11,* 253–256.

Ortendahl, M., Schettler, J. D., & Fries, J. F. (2000). Factors influencing length of time taking methotrexate in rheumatoid arthritis. *Journal of Rheumatology, 27,* 1139–1147.

Roelofsma, P. H., & van der Pligt, J. (2001). On the psychology of time preference and human decisions: Introduction to the special issue. *Acta Psychologica, 108,* 91–93.

Values

See Utility Assessment Techniques

Variance and Covariance

The variance of a random variable X quantifies the spread or dispersion of the distribution of X. A random variable is essentially a function or rule that associates a number with the outcome of an experiment. Examples of random variables are an individual's weight, the number of inpatient admissions at a hospital on a single day, and a binary indicator (e.g., assigning values of 0 = *no* or 1 = *yes*) of whether or not a patient participating in a study had his or her conditions improve after taking a new drug.

The variance and its square root, the standard deviation, are important summary measures of the distribution of X and are essential for providing a complete description of X. Suppose that μ_X denotes the mean of X. (The mean is also sometimes referred to as the expected value or average value of X and is typically estimated by taking the average—i.e., summing over all observed values and then dividing this sum by the total number of observations.) While μ_X, one of the primary quantities used to characterize a distribution, gives an important description of X by quantifying an average of the possible values of X, it does not impart any information about the potential variation in these values. Consider that two random variables may have the same average value but their distributions may differ greatly in the degree of concentration of the values, highlighting the need to summarize this concentration or spread.

In addition to evaluating the variation in a single random variable, when two random variables X and Y are measured, there is often interest in evaluating their relationship. The covariance quantifies the tendency of two random variables to vary together. For instance, it might be of interest to know how weight and height measurements change together. It is an important component when characterizing the joint distribution of X and Y.

This entry provides formal definitions of variance and covariance and describes how to estimate these two quantities when analyzing data.

Variance

Let $E[g(X)]$ denote the mean, or expected value, of the random variable $g(X)$, a function of X. The variance of X is defined as

$$\begin{aligned}\text{var}(X) = \sigma^2 &= E[(X - \mu_X)^2] \\ &= E[X^2] - (E[X])^2,\end{aligned}$$

the expected value of the squared difference between X and its mean, μ_X. It is the second moment about the mean of X. Observed values for X tend to fall around the center of the distribution, μ_X. The variance quantifies dispersion by looking at how far apart these values are on average from the mean. The more concentrated the values are around the mean, the smaller the variance. The less concentrated the values are, the larger the variance. By definition, the variance of X is always positive; that is, $\text{var}(X) \geq 0$. If the variance of X equals 0, then there is no variation in X and every observation is identical.

Sometimes the phrase *noisiness of the data* is used when describing the spread of observations. Referring to data as being less noisy suggests a distribution with a small variance, whereas referring to data as being more noisy suggests a distribution with a larger variance.

The standard deviation of X, usually written as σ, is simply the positive square root of the variance of X. Like the variance, it is also a measure of

spread of the distribution of X. Unlike the variance, it quantifies spread using the same scale, or units, on which X is measured. For example, the variance of weight measurements in a population would be reported in terms of squared kilograms or squared pounds. In contrast, the standard deviation would be reported simply in terms of kilograms or pounds.

Alternate expressions for the variance can be written based on whether X is discrete or continuous. Discrete random variables have only a finite or countable number of distinct values. For instance, a binary indicator of whether or not patients on a study had their conditions improve after taking a new drug is a discrete random variable because it has only two possible values (0 or 1). Continuous random variables can take any value in an interval. Weight is an example of a continuous random variable because weight measurements may take any value greater than 0.

If X is discrete, the variance can be written as

$$\text{var}(X) = \sum_i (x_i - \mu_X)^2 P(X = x_i),$$

where $x_1, x_2, x_3, \text{K}, x_i, \text{K}$ are the observed values of X and $P(X = x_i)$, the probability mass function, denotes the probability that the random variable X takes on the observed value x_i. If X is instead continuous, the variance can be written as

$$\text{var}(X) = \int_{\infty}^{\infty} (x - \mu_x)^2 f(x) dx,$$

where $f(x)$ is the probability density function of X.

If a and b are constants, then the variance of the linear transformation $a + bX$ is

$$\text{var}(a + bX) = b^2 X$$

because a constant has no variance (var(a) = 0).

In practice, one typically has data on a random sample of individuals from a population and is interested in estimating the population variance using data available from this random sample. Suppose that $x_1, x_2, x_3, \ldots, x_n$ are the observed values of X from a random sample of size n. In this case, the sample mean of the observations is calculated as $\bar{x} = \sum_{i=1}^{n} x_i / n$, and the variance is estimated using the sample variance s_x^2, defined by

$$s_x^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1} = \frac{\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}}{n-1},$$

where s_x^2 is the sample sum of squares divided by its degrees of freedom, $n - 1$. The sample standard deviation, s_x is the positive square root of s_x^2. It is the most frequently reported measure of variation.

Covariance

The covariance of two random variables X and Y is defined as

$$\begin{aligned}\text{cov}(X, Y) &= E[(X - \mu_x)(Y - \mu_Y)] \\ &= E[XY] - E[X]E[Y],\end{aligned}$$

where μ_Y denotes the expected value of Y. If X and Y tend to vary together, so that large values of X are usually observed with large values of Y and small values of X are observed with small values of Y, then their covariance is positive. In contrast, if X and Y tend to have an inverse relationship, so that large values of X are usually observed with small values of Y and small values of X are observed with large values of Y, then the covariance is negative. If X and Y are identically the same over all sample points, then cov(X,Y) = var(X) = var(Y).

If X and Y are independent random variables (i.e., X is not influenced by Y), then their covariance is zero: cov(X,Y) = 0. The converse, however, is not necessarily true. It may be the case that cov(X,Y) = 0 but X and Y are not independent.

If X and Y are not independent, the covariance is an important part of the variance of their sum. The variance of the sum $X + Y$ is the sum of the variances of X and Y plus twice their covariance:

$$\text{var}(x + y) = \text{var}(x) + \text{var}(y) + 2\text{cov}(x, y).$$

The covariance is closely related to the correlation coefficient, $\rho(X,Y)$, through the equation

$$\text{cov}(x, y) = \rho(x + y)\sigma_x \sigma_y,$$

where σ_Y is the standard deviation of the distribution of Y. The correlation coefficient assesses the degree of linearity between X and Y by

measuring how close their relationship is to a straight line.

The covariance may be difficult to interpret alone for two reasons. First, it quantifies the association between X and Y on a scale whose units of measurements are X times Y. Second, two components contribute to the covariance, as shown in the above equation, one component measuring the linear relationship between the two variables and one component measuring their individual variability. The covariance will be large if either X or Y has a large variance. It may be small if either X or Y has a small variance or if the two variables are unrelated. In contrast, the correlation coefficient is reported on a scale ranging from –1 to +1 regardless of the units of measurement of X and Y. The correlation coefficient takes the values of –1 or +1 only when there is an exact linear relationship between X and Y. If $\rho(X,Y) = 1$, then $Y = a + bX$, where $b > 0$. If $\rho(X,Y) = -1$, then $Y = a + bX$, where $b < 0$. In practice, values of $\rho(X,Y)$ will fall somewhere in the middle of this range. If $\rho(X,Y) = 0$, then X and Y are said to be uncorrelated.

The population covariance is estimated in practice with the sample covariance. Suppose that $(x_1, y_1), (x_2, y_2), (x_3, y_3), \ldots, (x_n, y_n)$ are the observed values of the distribution (X, Y) from a random sample of size n so that (x_i, y_i) represents the ith person's measurements (e.g., their height and weight measurements). The sample covariance, s_{xy}, is defined by

$$s_{xy} \quad \frac{\sum_{i \quad 1}^{n} (x_i \quad x)(y_i \quad y)}{n \quad 1} \quad \frac{\sum_{i \quad 1}^{n} x_i y_i \quad \frac{\left(\sum_{i \quad 1}^{n} x_i\right)\left(\sum_{i \quad 1}^{n} y_i\right)}{n}}{n \quad 1}.$$

The sample Pearson product moment correlation coefficient is then

$$r = \frac{s_{xy}}{s_x s_y},$$

where s_x and s_y are the sample standard deviations of X and Y, respectively.

Chaya S. Moskowitz

See also Analysis of Covariance (ANCOVA); Analysis of Variance (ANOVA); Intraclass Correlation Coefficient; Measures of Variability

Further Readings

Fisher, L. D., & van Belle, G. (1993). *Biostatistics: A methodology for the health sciences*. New York: Wiley.

Rosner, B. (2006). *Fundamentals of biostatistics* (6th ed.). Pacific Grove, CA: Thomson Brooks/Cole.

Violations of Probability Theory

Medical decision making often involves measures of uncertainty, including the explicit use of probability. To the degree uncertainty is present, the quality of a medical decision clearly depends on avoiding violation of the laws governing probability. Adherence to the laws of probability is analogous to following the laws of geometry when computing distances or the characteristics of an object, such as its volume. The basic laws of probability are quite simple, but their application can be subtle, and they are easily violated. Intuition about uncertainty is often at odds with those laws, especially when probabilities are small, conditional, or must be combined.

Violations of the laws of probability arise in other ways. For example, not using probability to measure uncertainty can be problematic in some methodologies, such as those based on fuzzy set theory. Even when a decision-making methodology is inherently probabilistic, misinterpretations and violations of the laws can occur. For example, in Bayesian statistical analyses, medical professionals are sometimes asked to provide input in the construction of prior distributions. This is a relatively new context in which there is great potential for violating the laws of probability.

The Laws of Probability

The axiomatic foundation of probability, and the multitude of theorems derived from it, constitutes the formal—and vast—theory of probability. These, more technical, results are not the subject of this entry. The violations considered here concern the basic laws of probability, which are as follows. Suppose T is an event about which there is uncertainty, such as whether a subject will respond to a

treatment. Let A be another event, such as whether the subject will experience an adverse event while responding to treatment. Denote by $P(\cdot)$ the probability of an event. Thus, $P(A)$ is the probability of an adverse event. If one knows that event T has occurred, then the conditional probability of A given T is written as $P(A|T)$, where the vertical line means "given." All probabilities, conditional or otherwise, for any events E and F, must conform to the following laws:

1. $0 \le P(E) \le 1$; this is the *convexity law*.
2. If E and F are mutually exclusive (the occurrence of one precludes the occurrence of the other), then $P(E \text{ or } F) = P(E) + P(F)$; this is the *addition law*.
3. $P(E \text{ and } F) = P(E|F)P(F) = P(F|E)P(E)$; this is the *multiplication law* for the *conjunction* of E and F.

These laws easily extend to more than two events.

Basic Consequences of the Laws

A useful consequence of the convexity law (1) is that the probability of the opposite or complement of an event E, denoted by E^c, is $1 - P(E)$. In the addition law (2), note that the "or" in the event "E or F" is not exclusive. The statement of the law proceeds from the assumption that E and F cannot occur simultaneously; that is, they are mutually exclusive. This would hardly be reasonable in the illustration given. One cannot typically preclude the occurrence of an adverse event in a treated subject. In such a case, where E and F are not mutually exclusive, one adjusts by subtracting the probability of their conjunction, a result called the *addition theorem*:

$$P(E \text{ or } F) = P(E) + P(F) - P(E \text{ and } F).$$

The multiplication law (3) involves the conditional probabilities $P(E|F)$ and $P(F|E)$. From the example above, $A|T$ is the event that the subject experiences an adverse event given that he or she responds to treatment. Note that the events $A|T$ and $T|A$ are very different. They can be related, however, by using a consequence of the multiplication law known as Bayes's theorem:

$$P(A|T) = P(T|A)P(A)/P(T).$$

Two events are defined to be independent when the occurrence of one has no effect on the probability of the other. Thus, events E and F are independent if $P(E|F) = P(E)$ or, equivalently, if $P(F|E) = P(F)$. If E and F are independent, then, using the multiplication law, $P(E \text{ and } F) = P(E|F)P(F) = P(E)P(F)$.

Another basic consequence of the laws is the conjunction inequality: For any two events E and F,

$$P(E \text{ and } F) \le \min\{P(E), P(F)\}.$$

That is, the probability of the conjunction cannot be larger than the probability of either component event.

Violating the Laws

There are axiomatic and operational justifications for the use of probability in the measurement of uncertainty. In a sense, the most common "violation" is measuring uncertainty with something other than probability. Such problems are a source of controversy in statistical inference (e.g., Bayesian vs. non-Bayesian methods) and in other areas such as expert system development (e.g., probabilistic expert systems vs. those based on fuzzy set theory). The focus below is on five common violations of the basic laws and their interpretations.

Conjunction Fallacy

This mistake is the result of ignoring the conjunction inequality. For example, let D be the event that an individual has type 2 diabetes. Let O be the event that the individual is obese. Suppose one asks, "Which has the greater probability, that one is Type 2 diabetic or that one is Type 2 diabetic and obese?" It is common for the conjunction in this sentence, D and O, to be mistaken for the conditional event $D|O$, in which case the conjunction may be assigned the higher probability, thus violating the conjunction inequality.

Transposed Conditionals

Also known as the base-rate fallacy, the prosecutor's fallacy, confusion of the inverse, and the inversion error, this is the mistake of confusing the conditional event $E|F$ with $F|E$. To illustrate this

error, let R be the event that a person in a population has a certain disease. Suppose from extensive study one estimates that $P(R) = .01$. Let D be the event that a diagnostic test for the disease is positive for such a person. Assume that this test has well-known properties, such as its sensitivity, $P(D|R) = .99$, and specificity, $P(D^c|R^c) = .95$. Now, suppose a subject tests positive for the disease. When presented with scenarios similar to this, it is very common for medical professionals to conclude that the person is "likely" to have the disease. In fact, because the prevalence or base rate is so low, the probability that the person has the disease given a positive diagnostic test is just .167 under this sensitivity and specificity. To see this, one can use Bayes's theorem: $P(R|D) = P(D|R)P(R)/P(D)$. Only $P(D)$ is unknown on the right-hand side of this equation. This can be obtained by "extending the conversation" to include the ways in which a positive diagnostic test can occur. It is possible for a patient to test positive whether or not the disease is present. The unknown probability can thus be written as

$$P(D) = P(D \text{ and } R \text{ or } D \text{ and } R^c),$$

where, again, R^c means that the disease is not present. Using the multiplication, addition, and convexity laws, it is not hard to show that this can be written as

$$P(D) = P(D|R)P(R) + P(D|R^c)[1 - P(R)].$$

Given the sensitivity and specificity of the test, one can compute all the components on the right-hand side of the equation. Thus, $P(D|R) = .99$, $P(D|R^c) = 1 - P(D^c|R^c) = .05$, and $P(R|D) = (.99)(.01)/[(.99)(.01) + (.05)(1 - .01)] \approx .167$.

Incoherence

For a set of numbers to be probabilities, they must be consistent with all the laws of probability—they must "cohere." Such coherence avoids contradictions in statements about uncertainty. Combining rates calculated from different surveys can easily lead to incoherent probability estimates. Such violations can be subtle. For example, suppose for some events A and B, one claims that $P(A) = .89$, $P(B) = .30$, $P(B|A) = .15$, and $P(A \text{ or } B) = .94$. These numbers are not obviously contradictory. However, from the addition theorem, $P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$. For this to equal .94, one must have $P(A \text{ and } B) = .25$. But this is incoherent since, using the multiplication law, $P(B|A)P(A) = (.15)(.89) = .1335$. Detecting incoherence can be difficult.

Fallacious Causality

This is the interpretive mistake wherein one claims that if $P(F|E)$ is large, then E must be causing F. While this is not strictly a violation of the laws of probability, it certainly overreaches them and is similar to the caveat regarding causation and correlation. Returning to the adverse event example, suppose a drug is indicated for the treatment of extreme agitation, perhaps in an emergency room setting, where covariates may be hard to assess. Let T be the event that a patient responds to this treatment. Let E be the event that a patient treated with this drug has an elongated QT interval on the electrocardiogram. Suppose further that $P(E|T)$ has been estimated in observational studies of agitated patients in emergency room settings. No matter how large $P(E|T)$ is, one cannot conclude that the treatment has caused the elongated QT interval. For example, an agitated patient taking insulin for type I diabetes has an increased risk of QT interval elongation whether treated for agitation or not. In observational studies, and even in poorly designed clinical trials, *lurking variables*, such as the patient's normal treatment regimen in this case, can be the source of mistaken causal inferences based on estimated conditional probabilities.

A closely related problem concerns confounding. For example, suppose a disease can affect both males and females but females have a higher recovery rate without treatment than males. In a clinical trial, if a treatment indicated for this disease goes more frequently to females than to males, then the resulting estimate of the probability of recovery under treatment may be inflated. Sex and treatment are potentially confounded here. If the differential assignment of the treatment in the clinical trial is not recognized, then sex becomes in effect a lurking variable. This leads to Simpson's paradox, which can make contingency table analyses problematic.

Fallacious Independence

Suppose in a certain population, males above a given age are obese (event O) with probability .25 and are hypertensive (event H) with probability .15. Since obesity and hypertension are both risk factors for heart disease, one may be interested in the probability that a male in this population has both factors. That is, one may wish to know $P(O \text{ and } H)$. It is tempting to compute $P(O \text{ and } H) = P(O)P(H) = (.25)(.15) = .0375$, but this implies that O and H are independent. That is, one can only multiply the probabilities of O and H if $P(H|O) = P(H)$ or, equivalently, if $P(O|H) = P(O)$. Clearly, it is unlikely that O and H are independent. What happens if one ignores this? Suppose, for example, that $P(H|O) = .35$ for males in this population. Then, using the multiplication law, $P(O \text{ and } H) = P(O|H)P(H) = (.35)(.15) = .0525$, instead of .0375 as it would be if independence were obtained.

John W. Seaman Jr.

See also Bayes's Theorem; Causal Inference in Medical Decision Making; Conditional Probability; Conjunction Probability Error; Decision Analyses, Common Errors Made in Conducting; Subjective Probability; Uncertainty in Medical Decisions

Further Readings

Cooke, R. M. (1991). *Experts in uncertainty*. New York: Oxford University Press.

Gilovich, T. (1991). *How we know what isn't so*. New York: The Free Press.

Lindley, D. V. (2006). *Understanding uncertainty*. Hoboken, NJ: Wiley.

W

Weighted Least Squares

One of the most used statistical techniques is regression analysis. In its simplest (and most used) form, regression analysis is used to describe and analyze the relationship between two continuous variables (e.g., weight and systolic blood pressure). In such instances, each observation (i.e., patient) can be represented with a point in a two-dimensional plot. The analysis is done by assuming that the relationship follows a known mathematical model. The data and the model are displayed in a graph called *scatterplot*, where x, or explanatory variable, is plotted in the horizontal axis, and y, or response, in the vertical axis. When the model is a straight line (defined by a slope and an intercept) and it is fitted to the data using the method of least squares, it is known as ordinary least squares (OLS), and the analysis is called simple linear regression (SLR). Weighted least squares (WLS) is a modification of OLS.

Least Squares Methods

The OLS method consists of finding the equation (i.e., slope and intercept) of the line that minimizes the sum of the squared vertical distances between each point and the line, hence the name least squares. These distances are called residuals. A picture helps understand these concepts. In 2001, Frank Bengel and colleagues studied the relationship between percent change in ejection fraction from rest to exercise (y variable) and hydroxyephedrine retention (x variable) in heart transplant patients. Each point in Figure 1 represents a patient and the line was computed using the OLS method. As is apparent from the figure, some points are farther from the line than others. The residuals are the distances from the line to each point in the y direction. By convention, points below the line have a negative residual and those above the line, a positive residual. The well-known correlation coefficient is a measure of the strength of the linear association and depends on the magnitude of the residuals. In this case, $r = .61$, indicating a moderately positive association that is statistically significant (i.e., not due to chance, since $p < .001$). An equivalent statement is that the slope of the line is significantly different from zero.

Mathematically, the OLS method minimizes the quantity

$$\sum_{i=1}^{n} e_i^2,$$

where e_i is the residual of the ith point and n is the number of points. In this sum, each residual has the same weight, and thus, each point (or residual) contributes equally to the sum. This is consistent with the assumption that every point contains the same amount of information, which is one of the key statistical assumptions of SLR. The (theoretical) assumption is that for each x measure, there exists a distribution of y measures with a mean that depends on the x measure and a variance that is the same for all values of x.

In statistical terms, the uniformity of variances is called *homoscedasticity*. Frequently, it is assumed that the distribution is normal. Figure 2a illustrates

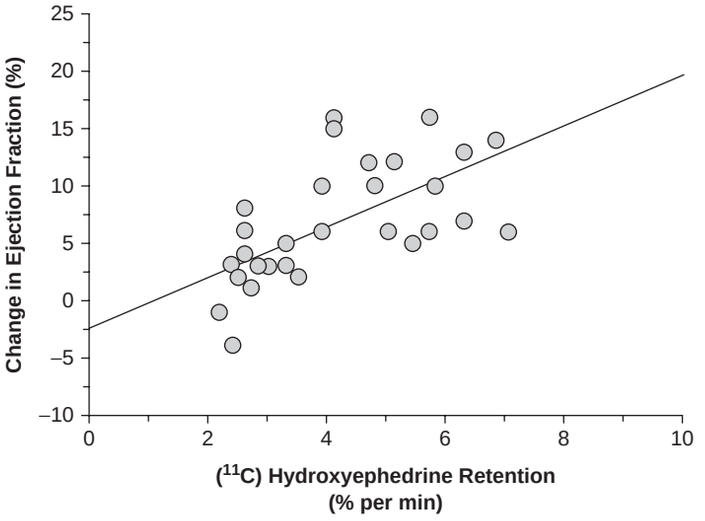


Figure 1 Relationship between hydroxyephedrine retention and percent change in ejection fraction

Source: Bengel et al. (2001).

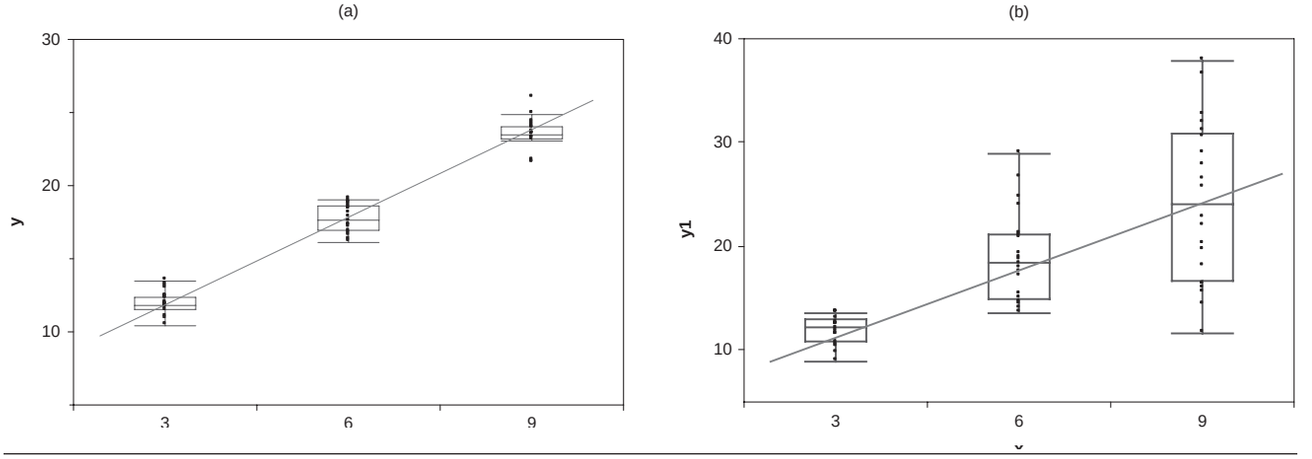


Figure 2 Plots displaying (a) homoscedasticity and (b) heteroscedasticity

this assumption. The popularity of OLS is due, in large part, to the fact that, under homoscedasticity (together with the independence of the observations), the fitted line has optimal statistical properties (i.e., the slope and intercept have smallest variance).

In many cases, homoscedasticity is a reasonable assumption, but sometimes the data or other theoretical reasons do not support it. Some biological relationships tend to have variation that depends on the average (e.g., larger average producing larger variation). This creates a "funnel" effect like the one illustrated in Figure 2b. In the example discussed previously, this would mean larger variation in ejection fraction among patients with higher hydroxyephedrine retention. The condition of unequal variances is called *heteroscedasticity,* and when it is present, the OLS method is not optimal.

The variance can be thought of as a measure of the information that a single observation conveys.

This is because as the variance increases, the likelihood that any observation is far from the mean increases (that is why larger samples are needed as the variance increases). Consequently, heteroscedasticity means that observations have different amounts of the information, and thus, it is reasonable to assign different weights to different points.

A common way to assign weights to observations is using the reciprocal of their variance. That is, if σ_i^2 is the variance at the ith point, the weight for the point would be $w_i = 1/\sigma_i^2$ These weights are smaller for observations with larger variances. For instance, in Figure 2b the weights assigned to the observations with $x = 9$ would be considerably smaller than those for $x = 3$ (and $x = 6$).

Assuming that the weights (i.e., the variances) are known, the problem is then to compute the straight line that minimizes the expression as

$$\sum_{i=1}^{n} w_i \cdot e_i^2 \cdot .$$

The mathematical technique is a straightforward modification of OLS and is called, appropriately, weighted least squares. Unfortunately, more often than not, the variances are not known and large samples are needed to estimate them with reasonable precision.

In practice, heteroscedasticity is difficult to ascertain and frequently goes undetected. If OLS is used when heteroscedasticity is present, the calculated slope and intercept will have larger variation than the ones obtained with WLS. Furthermore, the OLS will underestimate the true variances, and consequently, the confidence intervals will be too short and have a lower confidence than the one stated. The p values will also tend to be too small, increasing the Type I error rate.

Multiple linear regression (MLR) is the extension of SLR to the case of multiple explanatory variables. The model is more complicated than a straight line, but the fitting method is the same. In other words, OLS or WLS is used to fit the model depending on whether or not there is heteroscedasticity.

Application

WLS is used frequently to analyze public health information consisting of a large number of records. In 2007, Elyse Olshen and colleagues investigated the impact of state policies on vaccine coverage by age 13. The data consisted of percent vaccine coverage for more than 300 national insurance plans on 29 states. In this case, the response was the percent state coverage and the explanatory variable the type of state policy regarding vaccination. Olshen and colleagues used a WLS approach with weights that were inversely proportional to the variance in coverage across insurance plans within each state. Using MLR to adjust for other variables, the article found a significant association between coverage and middle school vaccination mandates.

WLS is commonly used in particular areas. In calibration, the objective is to estimate the unknown concentration of a substance given the observed response. The data are obtained by testing known concentrations and measuring the response. In these cases, the assumption of homoscedasticity is frequently not realistic. Estimates of the variance for each concentration of interest are obtained by running multiple replicates of the same concentration. Using the variance estimates, the weights can be calculated.

Another area in which WLS is used is meta-analysis. The goal of meta-analysis is to combine the results from multiple studies to produce a unified and more reliable conclusion. One of the challenges in meta-analysis is the pooling of results from studies that are heterogeneous in many ways, including sample size. One way this is done is by using weights that are proportional to the sample size.

Actually, any general regression technique such as logistic regression and proportional hazards (Cox) regression is based on some version of WLS.

Esteban Walker

See also Cox Proportional Hazards Regression; Logistic Regression; Ordinary Least Squares Regression

Further Readings

Bengel, F. M., Ueberfurh, P., Schiepel, N., Nekolla, S. G., Reichart, B., & Schwaiger, M. (2001). Effect of sympathetic reinnervation on cardiac performance after heart transplantation. *New England Journal of Medicine, 345,* 731–738.

Olshen, E., Mahon, B. E., Wang, S., & Woods, E. R. (2007). The impact of state policies on vaccine coverage by age 13 in an insured population. *Journal of Adolescent Health, 40,* 405–411.

Tellinghuisen, J. (2007). Weighted least-squares in calibration: What difference does it make? *Analyst, 132*(6), 536–543.

Welfare, Welfarism, and Extrawelfarism

Decisions need to be made when there are at least two alternatives to choose from. Like other decision making, medical decision making is about making such choices between alternatives. To do so in a rational, systematic, and optimal way, a decision-making framework must be developed, which stipulates how we should evaluate the different options. This first of all requires a definition of what exactly it is that we are seeking to improve by making optimal decisions, that is, the underlying goal of our decisions, and what it is that will be lost if we fail to do so. In economics, this underlying maximand has normally been labeled *welfare*, hence the term *welfare economics*. In essence, welfare economics is concerned with defining optimal allocations of scarce resources, where optimality is defined by the allocation of resources that maximizes (social) welfare. A dominant coherent normative framework has been developed to judge whether changes from state of the world A to B would improve welfare, either for an individual or a society.

Welfare

If welfare is the appropriate maximand of individuals and societies, in its most comprehensive form, it needs to represent some ultimate objective of individuals and societies. It should be an overall representation (index) of well-being of individuals or societies, in which all relevant underlying components of welfare are embraced. In this case, attaining higher levels of welfare is equivalent to saying that the involved individual or society is better off. Note that this is also possible when some underlying components of welfare decrease while others improve, as long as the latter compensate, in an acceptable way, for the former. Welfare is therefore also the sphere in which changes in individual underlying components of welfare can be traded off. Sometimes, it has been suggested that the components of welfare would be limited to goods and services only, and a few economists have even equated welfare, at either the individual or the societal level, with income or wealth in applied work. This, however, seems unnecessarily restrictive. Rather, the components of welfare can be diverse and may include aspects such as friendship, leisure, and marriage. For instance, an individual may trade off a loss of leisure against a gain of income and decide to be better off, that is, improve his or her welfare, by working more. A society may trade off the loss of an ancient forest against reduced travel time and decide to be worse off, that is, reduce welfare, when sacrificing the forest for more roads. It is especially such trade-offs of losses and gains that form the heart of welfare economics, as typically observed in economic evaluations.

Utility

In economics, a commonly used term for welfare is *utility*. Its meaning has always been ambiguous. In the days of Jeremy Bentham and John Stuart Mill, utility was often equated with happiness (as in the utilitarian motto "the greatest happiness for the greatest number") or with life satisfaction, although this equation always has been the topic of debate. The interpretation of utility as happiness or life satisfaction appears to become more popular again more recently, also in applied economics. However, the dominant view on utility moved away from this interpretation, also due to the problems of measurability and interpersonal comparability of utility (discussed below). The focus on the emotional evaluation of states of the world was rejected and utility is now usually taken to represent simply preference orderings or an index of choice. Within health economics and medical decision making, the term *utility* is indeed often used to refer to preference weights for different health states used in quality-adjusted life years (QALYs) calculations.

These two distinct interpretations of utility, that is, (1) *hedonic welfare* and (2) (anticipated) *preference*

satisfaction, still are the most prominent ones and sometimes are labeled as *experienced utility* and *decision utility* (or anticipated utility), respectively. In the first interpretation, an improvement in welfare for an individual is attained if a person actually experiences more utility (some might say "is happier") due to some change; and in the second interpretation, welfare is improved for that individual when moving a person to a state he or she prefers, for whatever reason. (A person may, for instance, prefer one health state to another, because one believes that one will be happier in that health state, simply because one will be healthier in that state, or because one feels that one would be less of a burden to others in that health state.) The reasons for preferring some state to another are normally not considered or evaluated by economists. Note that ignoring the quality of utility (i.e., the normative evaluation of the individuals' preferences, such as whether they are disgusting or cruel) has been criticized by some economists.

Measurability and Comparability

The measurability and interpersonal comparability of utility has been a matter of controversy as well and has importantly contributed to the second interpretation of utility mentioned above. While utility was believed to be measurable cardinally and interpersonally comparable in the old days (with the concept of the *hedonimeter*, which would interpersonally comparably measure happiness and pain in individuals, of Francis Edgeworth as the illustrative highlight of this school of thought), nowadays the idea of interpersonally incomparable ordinal utility is more prominent. This change has had a profound impact on welfare economics, especially in the context of social welfare. Under the assumption of cardinal and comparable utility, one may simply maximize social welfare by comparing utility gains and utility losses of some change. For instance, in a two-person world, if a change causes a loss of utility of 3 units in Person A but a utility gain of 5 units in Person B, it is immediately clear that total happiness increases by 2 units due to that change. Social welfare would then be some (weighted) function of individual utilities—like under utilitarianism it is assumed to be a simple unweighted summation of all individual utilities. Without comparability (and cardinality), it is impossible to directly trade off welfare gains and losses between individuals. Then, in the above example, we only know that Person A decreases in utility while Person B increases, but any quantitative comparison of gains and losses is impossible and meaningless. This has resulted in the popularity of the Pareto criterion, which states that social welfare increases unambiguously only when the utility of at least one person increases and that of no one falls. This criterion is not only very restrictive (as it does not allow a decrease in utility for anyone, regardless of the gains for others), but it has also been criticized for ignoring equity considerations. For example, even if in our two-person world, Person A would be extremely rich, while Person B is extremely poor, any redistribution of wealth from A to B making Person A even slightly worse off while dramatically improving B's position does not satisfy the Pareto criterion. Such equity considerations can be included in approaches using (Bergson-Samuelson) social welfare functions by including equity weights (or inequality aversion) attached to different levels of welfare, but then, again, interpersonal comparisons in terms of utility are made, not only to come to social welfare from individual welfares but also to determine who should receive more (equity) weight in this process.

Welfarism

Welfarism has two distinct, though related, meanings. First of all, it is often used to describe the dominant ethical framework used in welfare economics to judge states of the world with. This dominant framework is built on four key tenets. The first tenet is *the utility principle*. This means that individuals are assumed to maximize their welfare in a rational way. They are assumed to do so by ordering the options open to them and subsequently choosing the preferred option. This rational maximization behavior has importantly been questioned in the "behavioral economics" school, exposing a gap between the behavioral assumptions in the neoclassical economic framework and real-world decision making. The second tenet is *individual sovereignty*. Individuals themselves are considered to be the best judges of what is good for them, that is, what maximizes their welfare. They themselves therefore (should) decide on what the preferred option is. The third tenet is

consequentialism. This indicates that individuals derive utility from the outcomes of their choices, as opposed to, for instance, merely intentions. Fourth is the tenet of *welfarism*. Welfarism implies that the goodness or desirability of a state of the world is evaluated only by the level of utility (experienced or anticipated to be had) in that state of the world. This tenet, therefore, effectively reduces the information to be considered in any individual or social ranking of states of the world (i.e., the *evaluative space*) to only utility. Thus, only individual preferences are relevant in social rankings of states of the world. Together, these four tenets build the dominant welfare economic framework, sometimes labeled *welfarist economics*, with the distinct feature that only individual utilities are allowed to determine the desirability of different states of the world.

The second meaning of the term *welfarism* relates only to the fourth tenet of welfarist economics, that is, the one that restricts what can be considered in an evaluation (i.e., the evaluative space) to individual utility only, ignoring all other possible outcomes. These other outcomes may be diverse and are often generically labeled as *nonutility information*. Especially in the context of health economics and medical decision making, it has been questioned whether the welfarist restriction on the evaluative space would be appropriate and useful; hence the attention for an alternative stream, called *extrawelfarism* (although the label *nonwelfarism* has also been used).

Extrawelfarism

Extrawelfarism is a stream within welfare economics, which, while like welfarist economics seeking to make meaningful statements about the relative desirability of different states of the world, rejects the tenet of welfarism in doing so. It therefore rejects the key characteristic of welfarist economics, that is, that it only uses utility information in a welfare evaluation. Thus, extrawelfarism opens up the evaluative space in such evaluations to include a range of possible outcomes as well as or instead of individual utilities. The underlying reason for this relative permissiveness under extrawelfarism especially appears to be that the welfarist notion that only individuals' experienced or anticipated utilities are important in deciding whether one state of the world is preferable to another is considered unsatisfactory as a normative underpinning of welfare judgments. There is more to welfare than just individual preference-based utility, according to extrawelfarists. Extrawelfarism may include in the evaluative space other aspects and characteristics of human beings, such as their capabilities, their potential attainments, their health, or their share of such things in society. These outcomes are considered important in welfare judgments in their own right, not merely because they produce utility nor just to the extent that they do so.

The extrawelfarist stream, therefore, requires the purposeful selection of relevant outcomes to be included in welfare judgments and, therefore, the development of other normative (but still welfare economic) frameworks for deciding on the relative desirability of different states of the world. While the extrawelfarist framework allows a context-dependent definition of the relevant and appropriate outcomes to be considered in the evaluation, one important predefined alternative framework has been proposed by Amartya Sen. This is his capabilities approach. Therein, rather than focusing on the (anticipated) emotional reaction of people to what they do and are, welfare should be judged according to what people are *capable* of doing, being, or becoming. One may draw a parallel here to striving for *equal opportunity for welfare*.

In health economics and medical decision making, where extrawelfarism was introduced especially by Anthony Culyer, the focus has especially been on the *health* of individuals as an important human characteristic (or capability). Health, measured in some meaningful way (perhaps but not necessarily entailing utility measurement), may be considered an appropriate maximand in the context of healthcare decisions. QALYs, which seek to measure and combine the main characteristics of what is usually entailed by the idea of physical and mental health, have been proposed as such a meaningful measure of health. While QALYs are often used in the context of cost-utility analysis, extrawelfarists have claimed that they do not correspond to the traditional idea of utility, for instance, due to the rescaling on a fixed scale (so that everyone with perfect health gets a utility score of 1 regardless of other characteristics) and due to the way they are used (e.g., simply adding, subtracting, and averaging QALY scores across people). Note that a focus on health is not a general feature of the

extrawelfarist stream but merely an application of it in the context of healthcare. And there also, other relevant outcomes (such as relief of burden to caregivers or access to healthcare) may be selected and included in the evaluation, next to relevant health measures.

By allowing outcomes other than utility in the evaluative space, extrawelfarist analysis can circumvent some of the traditional problems of measurement and interpersonal comparability of utility. Individuals may be compared, for instance, in terms of their health, and one person may be deemed to be more in need of care than another on the grounds of being in a worse health state (while not claiming anything about these individuals' utility or preferences). This also allows weighting of relevant outcomes, for instance, to reflect equity considerations, but these equity weights then do not need to pertain to the utility of the affected individuals (but to some other relevant outcome, such as their health status), nor do the weights themselves necessarily need to be preference based. Moreover, unlike when eliciting utility, the source of valuation under extrawelfarism does not necessarily have to be the affected individual (e.g., to avoid the consequences of adaptation). This permissiveness of extrawelfarism depends on its key feature, that is, that it allows outcomes other than utility to be considered in a welfare economic evaluation.

Werner Brouwer

See also Economics, Health Economics; Equity; Expected Utility Theory; Patient Satisfaction; Quality-Adjusted Life Years (QALYs); Social Judgment Theory; Utility Assessment Techniques

Further Readings

Boadway, R., & Bruce, N. (1984). *Welfare economics*. Oxford, UK: Blackwell.

Brouwer, W. B. F., Culyer, A. J., van Exel, N. J. A., & Rutten, F. F. H. (2008). Welfarism vs. extra-welfarism. *Journal of Health Economics, 27*(2), 325–338.

Cohen, G. A. (1993). Equality of what? On welfare, goods and capabilities. In M. C. Nussbaum & A. K. Sen (Eds.), *The quality of life* (pp. 9–29). Oxford, UK: Clarendon Press.

Culyer, A. J. (1990). Commodities, characteristics of commodities, characteristics of people, utilities and the quality of life. In S. Baldwin, C. Godfrey, & C. Propper (Eds.), *The quality of life: Perspectives and policies* (pp. 9–27). London: Routledge.

Hurley, J. (1998). *Welfarism, extra-welfarism and evaluative economic analysis in the health care sector*. In M. L. Barer, T. E. Getzen, & G. L. Stoddard (Eds.), *Health, health care and health economics: Perspectives on distribution* (pp. 373–395). Chichester, UK: Wiley.

Kahneman, D., Wakker, P. P., & Sarin, R. K. (1997). Back to Bentham? Explorations of experienced utility. *Quarterly Journal of Economics, 112*, 375–405.

Sen, A. K. (1993). Capability and well-being. In M. C. Nussbaum & A. K. Sen (Eds.), *The quality of life* (pp. 30–53). Oxford, UK: Clarendon Press.

Willingness to Pay

Willingness to pay (WTP) is the maximum amount of money an individual is willing to pay to ensure that a proposed service or good is available. Determining WTP through the contingent valuation method (CVM) is increasingly being used to generate information on the benefits of, and demand for, healthcare programs. The method is called *contingent valuation* because the respondent is asked to consider the contingency of a market existing for the thing being valued. The WTP could be for the availability of a resource to the individual (own use); for others, for example, poor people (altruism); needed by others or the individual in the future (option or nonuse); or a combination of any of these measures. CVM is accepted as a theoretically correct method to estimate the value of goods and services to consumers and provides the monetary measure of benefit in cost-benefit analysis.

Question Formats

There is still paucity of knowledge about the relative validity of WTP elicited using different CVM question formats. There are many CVM question formats for eliciting WTP, which in turn might determine the level of validity of elicited responses. The various question formats (or elicitation methods) for eliciting WTP include open-ended, the bidding game, payment cards (or categorical scales), dichotomous choice (binary, close-ended, take-it-or-leave-it), and the dichotomous with follow-through

question formats. There are also variants of all these question formats. However, it is potentially better to use the question formats that are more context specific as this would yield more valid estimates. Some question formats have been developed to improve the realism of the WTP elicitation for the hopeful elicitation of more valid WTP estimates. Two CVM question formats that mimic market transactions (price taking) are the bidding game and the structured haggling technique, which are described below.

Bidding Game

The bidding game operates by the respondent being presented with an amount and asked whether he or she would be willing to pay that amount. However, depending on the response, the respondent is bid up or down using a predetermined bidding iteration until the maximum number of predetermined bidding iteration is reached. The final amount is open-ended and represents the maximum amount the respondent is willing to pay.

Structured Haggling

The structured haggling technique was developed to mimic the usual haggling price-taking characteristics in open markets (e.g., those found in southeast Nigeria). It is structured so that the steps are standardized for use by different interviewers in the same study. The seller or interviewer initiates the haggling process, by offering the good to the buyer at a price that is well above the expected sale price. This is to give adequate room for the respondent to bargain so that in the end the seller will end up selling at a price that is equal or more than the sale price of the net. As it does often happen, the buyers who feel that sellers normally inflate the first offer will say no to the first price offered. However, the price-naive ones may agree to the first offer, whereby the sale is effected and goods and money exchange hands. The structured haggling technique has an additional step where the respondent is asked his or her WTP in case the price of the good rises in the future due to unforeseen circumstances. This step is aimed at eliciting the maximum WTP. Hence, the final amount is also open-ended and represents the maximum amount the respondent is willing to pay.

Validity

Validity tests are divided into content, criterion, and construct validity. Content validity determines whether the contingent valuation scenario presents a realistic description of the good or service under valuation. Construct validity refers to whether the measurement corresponds to theoretical concepts, and criterion (or criterion related) validity is the correlation of hypothetical or stated WTP with actual behavior or actual WTP. Construct validity is usually investigated if there is no immediate available gold standard with which to investigate criterion validity, which is usually the case in most CVM studies. In testing construct validity, hypothetical constructs are used to understand elicited WTP, usually through econometric analysis, by using an appropriate regression analyses model depending on how WTP was measured. Usually, in testing for construct validity, independent (explanatory) variables are used to explain the determinants of elicited WTP.

Illustrative Study: Malaria Treatment in Southeast Nigeria

In this study, the structured haggling and bidding game techniques were used to elicit WTP for different malaria treatment strategies in the Oji-river town in Oji-River LGA of Enugu State. Multistaged sampling was used to select the respondents. A subsample of the households had a bidding game, while the other had structured haggling question formats for eliciting WTP. Using the household lists as the sampling frame, 370 households were selected from each subgroup (total of 740 households) using simple random sampling. In each selected household, an adult was interviewed.

Pretested questionnaires were used to elicit respondents' WTP per episode of malaria. In eliciting WTP, scenarios that explained the different malaria treatment strategies were first presented to the respondents. Then, either the bidding game or structured haggling technique was used to elicit WTP from a subsample of the respondents. The interventions presented were treatment in public hospitals, primary healthcare centers, private hospitals, and by community health workers.

The average WTP amounts that the respondents were willing to pay to receive treatment for an episode of malaria from four different treatment

providers are shown in Table 1. The four rows with the results compare the mean (average) WTP for treatment of malaria from different healthcare providers. Consistently higher WTP amounts were elicited using the structured haggling technique for the four valuations and all were statistically significant between structured haggling and the bidding game. The highest level of WTP using the structured haggling was for improved malaria treatment in private hospitals ($5.03), while it was for treatment in public hospitals ($3.11) in the case of the bidding game. However, in both the bidding game and structured haggling, the least WTP was for treatment by community health workers.

The relationships between WTP and many of the independent variables were consistent with a priori hypothesized relationships suggested by demand theory, hence supporting the construct validity of the two CVM question formats. The results of a comparative test of construct validity of the estimates of WTP from the bidding game and structured haggling show that having formal education was positively and statistically significantly related to WTP in five out of the eight regression models. Also, whether or not a respondent stated that he or she was willing to pay for malaria treatment was positively related to WTP in all eight valuations ($p < .05$). However, WTP for primary healthcare centers and WTP for private hospitals were weakly construct valid for the bidding game, while WTP for private hospitals was weakly construct valid for the structured haggling. All the regression analyses were statistically significant, and on average, they explained more than 40% of the variations (with the exception of WTP for private hospitals in the structured haggling) in WTP.

Obinna Onwujekwe

See also Contingent Valuation; Cost-Benefit Analysis; Health Status Measurement, Construct Validity; Human Capital Approach; Monetary Value; Ordinary Least Squares Regression

Further Readings

Carmines, E. G., & Zeller, R. A. (1979). *Reliability and validity assessment*. London: Sage.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52,* 281–302. (Quoted in Carmines and Zeller [1979])

Donaldson, C., Jones, A. M., Mapp, T. J., & Olson, J. A. (1998). Limited dependent variables in willingness to pay studies: Applications in health care. *Applied Economics, 30,* 667–677.

Frykblom, P. (1997). Hypothetical question modes and real willingness to pay. *Journal of Environmental Economics and Management, 34,* 275–287.

Johannesson, M., Jonsson, B., & Borgquist, L. (1991). Willingness to pay for anti-hypertensive therapy: Results of a Swedish pilot study. *Journal of Health Economics, 10,* 461–474.

Klose, T. (1999). The contingent valuation method in health care. *Health Policy, 47,* 97–123.

O'Brien, B., & Gafni, A. (1996). When do the "dollars" make sense? Toward a conceptual framework for contingent valuation studies in health care. *Medical Decision Making, 16*(3), 288–299.

Table 1 Willingness to pay for different providers in the four communities

	BG *N = 346* *Naira (US$)*	*SH* *N = 352* *Naira (US$)*	*Chi-Square (p value)*
WTP for improved quality of malaria treatment in public hospitals	373.6 ($3.11)	479.5 ($4.0)	29.4 ($p < .0001$)
WTP for improved quality of malaria treatment in public PHC centers	331.4 ($2.8)	436.2 ($3.6)	45.0 ($p < .0001$)
WTP to pay for treatment in private hospitals	317.0 ($2.6)	603.2 ($5.03)	179.8 ($p < .0001$)
WTP to pay for treatment by community health workers	245.2 ($2.04)	354.3 ($2.95)	38.8 ($p < .0001$)

Note: 120 Naira = US$1; WTP, willingness to pay.

Onwujekwe, O. (2004). Criterion and content validity of a novel structured haggling contingent valuation question format versus the bidding game and binary with follow-up questions. *Social Science and Medicine, 58,* 525–537.

Onwujekwe, O., Fox-Rushby, J., & Hanson, K. (2004). Valuing the benefits of a health intervention using three different approaches to contingent valuation: Re-treatment of mosquito bed-nets in Nigeria. *Journal of Health Services Research and Policy, 9,* 67–75.

Onwujekwe, O., Hanson, K., & Fox-Rushby, J. (2008). Construct validity of the bidding game, binary with follow-up, and a novel structured haggling question format in determining willingness to pay for insecticide-treated mosquito nets. *Medical Decision Making, 28,* 90–101.

Worldviews

Worldview is the overarching framework through which individuals perceive themselves and their world. It dictates (a) what information individuals seek and (b) how they organize, process, and determine the meaning of information in their social world. Thus, this construct allows for the concurrent examination of both (a) intra-individual-level beliefs and values and (b) external factors that affect behavior. It explores how individuals' unique experience of their world interacts with their self-perceptions to guide the decision-making process. An improved understanding of individuals' worldviews could be used to match interventions to targeted populations at risk to improve health outcomes.

Worldview and Self-Fulfilling Prophesy

The construct of worldview examines how beliefs about the self and the world interact to affect behavior. The following are two examples. Individuals with a high sense of self-efficacy may be more likely to engage in preventive behaviors because they believe that their actions will result in reduced harm (the self-controllability dimension). This may be amplified by the belief that the world is inherently just and that misfortune will not come to those who are good, decent, and try hard (the justice world dimension). However, this behavior may be undermined if individuals believe that they are not worthy of receiving good things (the self-deservingness dimension) and/or that bad things arbitrarily befall on people in the world (the randomness world dimension).

This notion is akin to the psychological theory of attribution, which studies individuals' tendency to see the world in a habitual way and how it can lead to habitual behavioral responses. Attribution theory takes this notion one step further to suggest that if one *believes* the world to be uncontrollable, as in the example above, it is likely that one will actually *experience* an uncontrollable world, in part due to the self-fulfilling prophesy. One example in medical decision making is the impact of misperception. If individuals hold the view that cancer is a death sentence, they are more likely to abstain from efforts to prevent or treat cancer; therefore, if diagnosed with cancer, it is likely to be at a later stage when overt symptoms develop (such as acute pain, uncontrolled bleeding) and medical treatment becomes necessary. Because the cancer was not caught early, treatment options become limited and the chance of survival dramatically decreases, tragically confirming their initial belief.

Imagine that this misperception was prevalent in a community at high risk for developing cancer, and as a result, community members engaged in little to no prevention. If policy makers failed to identify their worldview, they might assume that low cancer-screening rates were due to structural factors, such as limited access to screening centers. This mismatch in worldviews could lead policy makers to invest their limited healthcare dollars in the development of a screening center easily reached by public transportation rather than the more appropriate, and less costly, intervention aimed at addressing the community's misperceptions.

Therefore, a more informed understanding of how individuals perceive themselves, combined with specific beliefs about the world, may help (a) identify those at increased vulnerability for experiencing poor health outcomes and (b) inform the development of appropriate, targeted interventions.

Julie Goldberg

See also Cognitive Psychology and Processes; Context Effects; Cultural Issues; Decision Psychology; Decisions

Faced by Patients: Primary Care; Judgment; Protected Values; Risk Communication; Risk Perception

Further Readings

Janoff-Bulman, R. (1989). Assumptive worlds and the stress of traumatic events: Applications of the schema construct [Special issue]. *Social Cognition, 7*(2), 113–136.

Kagee, A., & Dixon, D. N. (2000). Worldview and health promoting behavior: A causal model. *Journal of Behavioral Medicine, 23*(2), 163–179.

Pepper, S. (1961). *World hypotheses: A study in evidence*. Berkeley: University of California Press.

Index

Entry titles and their pages are in **bold**.