OPERATIONS RESEARCH AND HEALTH CARE

A HANDBOOK OF METHODS AND APPLICATIONS

Recent titles in the

INTERNATIONAL SERIES IN
OPERATIONS RESEARCH & MANAGEMENT SCIENCE
Frederick S. Hillier, Series Editor, *Stanford University*

Zhu, J. / *QUANTITATIVE MODELS FOR PERFORMANCE EVALUATION AND BENCHMARKING*

Ehrgott, M. & Gandibleux, X. / *MULTIPLE CRITERIA OPTIMIZATION: State of the Art Annotated Bibliographical Surveys*

Bienstock, D. / *Potential Function Methods for Approx. Solving Linear Programming Problems*

Matsatsinis, N.F. & Siskos, Y. / *INTELLIGENT SUPPORT SYSTEMS FOR MARKETING DECISIONS*

Alpern, S. & Gal, S. / *THE THEORY OF SEARCH GAMES AND RENDEZVOUS*

Hall, R.W./ *HANDBOOK OF TRANSPORTATION SCIENCE* - **2nd** *Ed.*

Glover, F. & Kochenberger, G.A. / *HANDBOOK OF METAHEURISTICS*

Graves, S.B. & Ringuest, J.L. / *MODELS AND METHODS FOR PROJECT SELECTION: Concepts from Management Science, Finance and Information Technology*

Hassin, R. & Haviv, M./ *TO QUEUE OR NOT TO QUEUE: Equilibrium Behavior in Queueing Systems*

Gershwin, S.B. et al/ *ANALYSIS & MODELING OF MANUFACTURING SYSTEMS*

Maros, I./ *COMPUTATIONAL TECHNIQUES OF THE SIMPLEX METHOD*

Harrison, T., Lee, H. & Neale, J./ *THE PRACTICE OF SUPPLY CHAIN MANAGEMENT: Where Theory And Application Converge*

Shanthikumar, J.G., Yao, D. & Zijm, W.H./ *STOCHASTIC MODELING AND OPTIMIZATION OF MANUFACTURING SYSTEMS AND SUPPLY CHAINS*

Nabrzyski, J., Schopf, J.M., **Węglarz, J.**/ *GRID RESOURCE MANAGEMENT: State of the Art and Future Trends*

Thissen, W. A.H. & Herder, P.M./ *CRITICAL INFRASTRUCTURES: State of the Art in Research and Application*

Carlsson, C., Fedrizzi, M., & Fullér, R./ *FUZZY LOGIC IN MANAGEMENT*

Soyer, R., Mazzuchi, T.A., & Singpurwalla, N.D./ *MATHEMATICAL RELIABILITY: An Expository Perspective*

Chakravarty, A.K. & Eliashberg, J./ *MANAGING BUSINESS INTERFACES: Marketing, Engineering, and Manufacturing Perspectives*

Talluri, K. & van Ryzin, G./ *THE THEORY AND PRACTICE OF REVENUE MANAGEMENT*

Kavadias, S. & Loch, C.H./*PROJECT SELECTION UNDER UNCERTAINTY: Dynamically Allocating Resources to Maximize Value*

Brandeau, M.L., Sainfort, F., Pierskalla, W.P./ *OPERATIONS RESEARCH AND HEALTH CARE: A Handbook of Methods and Applications*

Cooper, W.W., Seiford, L.M., Zhu, J./ *HANDBOOK OF DATA ENVELOPMENT ANALYSIS: Models and Methods*

Luenberger, D.G./ *LINEAR AND NONLINEAR PROGRAMMING,* **2nd** *Ed.*

Sherbrooke, C.C./ *OPTIMAL INVENTORY MODELING OF SYSTEMS: Multi-Echelon Techniques, Second Edition*

Chu, S.-C., Leung, L.C., Hui, Y. V., Cheung, W./ *4th PARTY CYBER LOGISTICS FOR AIR CARGO*

Simchi-Levi, Wu, Shen/ *HANDBOOK OF QUANTITATIVE SUPPLY CHAIN ANALYSIS: Modeling in the E-Business Era*

Gass, S.I. & Assad, A.A./ *AN ANNOTATED TIMELINE OF OPERATIONS RESEARCH: An Informal History*

**** A list of the early publications in the series is at the end of the book ****

OPERATIONS RESEARCH AND HEALTH CARE

A HANDBOOK OF METHODS AND APPLICATIONS

Edited by

Margaret L. Brandeau
Stanford University

François Sainfort
Georgia Institute of Technology

William P. Pierskalla
University of California at Los Angeles

KLUWER ACADEMIC PUBLISHERS
NEW YORK, BOSTON, DORDRECHT, LONDON, MOSCOW

eBook ISBN: 1-4020-8066-2
Print ISBN: 1-4020-7629-0

Created in the United States of America

Visit Springer's eBookstore at: http://ebooks.kluweronline.com
and the Springer Global Website Online at: http://www.springeronline.com

CONTENTS

INTRODUCTION

HEALTH CARE OPERATIONS MANAGEMENT

PUBLIC POLICY AND ECONOMIC ANALYSIS

CLINICAL APPLICATIONS

1 HEALTH CARE DELIVERY: CURRENT PROBLEMS AND FUTURE CHALLENGES

Margaret L. Brandeau[1], François Sainfort[2]
and William P. Pierskalla[3]

[1]Department of Management Science and Engineering
Stanford University
Stanford, CA 94305

[2] Department of Industrial and Systems Engineering
Georgia Institute of Technology
Atlanta, GA 30332

[3]Anderson Graduate School of Management
University of California at Los Angeles
Los Angeles, CA 90095

SUMMARY

In both rich and poor nations, public resources for health care are inadequate to meet demand. Policy makers and health care providers must determine how to provide the most effective health care to citizens using the limited resources that are available. This chapter describes current and future challenges in the delivery of health care, and outlines the role that operations research (OR) models can play in helping to solve those problems. The chapter concludes with an overview of this book – its intended audience, the areas covered, and a description of the subsequent chapters.

KEY WORDS

Health care delivery, Health care planning

1.1 WORLDWIDE HEALTH: THE PAST 50 YEARS

Human health has improved significantly in the last 50 years. In 1950, global life expectancy was 46 years [1]. That figure rose to 61 years by 1980 and to 67 years by 1998 [2]. Much of these gains occurred in low- and middle-income countries, and were due in large part to improved nutrition and sanitation, medical innovations, and improvements in public health infrastructure.

However, not all countries have experienced an increase in life expectancy in recent years. In countries of the former Soviet Union, life expectancy dropped from 70 years in 1986 to 64 years in 1994, with an even more marked drop among men [3]. Factors contributing to this decline include economic and social instability, high rates of tobacco and alcohol consumption, poor nutrition, depression, and deterioration of the health care system [4]. In many African nations, life expectancy has been significantly diminished by HIV/AIDS. In seven African countries life expectancy is now less than 40 years and falling [5].

Worldwide, infectious diseases kill 13 million people per year [6]. In 1999, 2.8 million people died from AIDS alone [7]. Infectious diseases once confined to specific geographic regions have spread across country borders as a result of increasing global travel. New infectious diseases continue to emerge [8]. Noncommunicable diseases such as heart disease, cerebrovascular disease (stroke), cancer, and diabetes are the primary cause of death in high-income countries. Such diseases currently account for less than half of all deaths in low-income countries, but in the next 20 years are expected to account for 70% of deaths [9]. In low-income countries, malnutrition remains a serious health problem, whereas in high-income countries, obesity is increasingly becoming a health problem. Tobacco, alcohol, and drug use have led to significant health problems worldwide. Tobacco use currently accounts for almost 5 million premature deaths per year. This figure is projected to rise to more than 8 million deaths per year by 2020, with many of these in low- and middle-income countries [10]. Food and water contamination cause at least 2 million premature deaths per year, primarily in low- and middle-income countries. Environmental agents (e.g., arsenic, lead, silica) also pose significant health risks in some areas. In addition, manmade chemical and biological weapons are a potential threat to public health.

1.2 HEALTH CARE DELIVERY CHALLENGES

Governments and health care providers face a variety of challenges in the delivery of health care. Below we describe current and future health care challenges. Because low- and middle-income countries face significantly different challenges in health provision than do high-income countries, we describe their current health care challenges separately. We then describe the future challenges in health care delivery that are common to all countries.

1.2.1 Current health care delivery challenges in low-income and middle-income countries

In low- and middle-income countries, where 80% of the world's population lives, malnutrition and infectious diseases account for significant numbers of premature deaths. Half of young child deaths in low-income countries are caused by malnutrition [11]. Although vaccines are available for a number of infectious diseases that cause childhood deaths, 25% of the world's children have not received these vaccines [12]. Many people in low- and middle-income countries do not receive even basic health care. Health facilities are often located in urban areas, far from rural areas and frequently difficult to access by public transportation. The care that is provided can be costly and substandard. In recent years, low- and middle-income countries have seen a significant shift in population from rural to urban areas, but have had no commensurate increase in urban health services. Inadequate infrastructure (e.g., inadequate roads, storage and distribution systems, electricity, clean water) and poorly functioning public health systems also impede the provision of health care.

Resources for health care in low-income countries are quite limited. Among the world's 60 poorest nations, annual per capita health spending in the year 2000 was less than $15 [13], and approximately one third of this funding came from international aid. Such an amount is insufficient to provide even the most basic health services. In contrast, annual per capita health spending in the industrialized world was on the order of $2,000, and was $4,500 in the United States [13]. Even if low-income countries were to devote more of their scarce public funds to health care, as recently recommended by the World Health Organization [13], per capita spending would still be at levels far below that in the industrialized world.

The lack of health care funding in low- and middle-income countries is exacerbated by rising health care needs and costs. Countries severely affected by HIV/AIDS are facing far greater demands for health care than they can meet. In other low- and middle-income countries, aging populations have increased overall demand for health care. Health care costs have risen as a result of new health care technologies and procedures. Moreover, many medicines that are routinely used in high-income countries (so-called "essential medicines") are not affordable in low-income countries. In some cases, even basic vaccines are too expensive.

1.2.2 Current health care delivery challenges in high-income countries

In high-income countries, resources for health care are orders of magnitude greater than in low-income countries. However, high-income countries face their own health care challenges. Although such countries spend much more on health than low-income countries, performance of health care systems varies markedly among high-income countries. For example, the United States spends almost twice as much per capita on annual health care as many other high-income nations, without achieving any greater life expectancy or any lower "burden of disease" (measured in terms of life years lived, adjusted for health disabilities) [14]. A recent report by the World Health Organization ranked the U.S. 37th in overall health systems performance among 191 Member States [14]. France, which spends half as much as the U.S. on per capita annual health care, was ranked first in overall health systems performance [14]. (Health systems performance as measured in the report included not only measures of health, but also measures of health system fairness and responsiveness.)

Inequities in health care provision exist within high-income countries. In countries with no national health system, such as the U.S., a significant fraction of individuals have no health insurance coverage and thus have only limited access to health care. Poor people and those in rural areas also often have only limited access to health care. In some high-income countries, including the U.S., the gap in life expectancy between rich and poor people is as great as the gap in life expectancy between high- and low-income countries [15].

Like low-income countries, high-income countries have experienced significant increases in demand for and cost of health care. Aging populations are making disproportionately heavy demands on health systems in high-income countries. Chronic conditions have become more prevalent. While new health care technologies and procedures have improved health, they have also increased costs. Patients are not only consuming more health services, but are consuming more intensive health services. Prescription drugs have also become increasingly expensive. In many high-income countries, health care spending has significantly outpaced economic growth. In the U.S., for example, health care spending accounted for 5% the Gross Domestic Product (GDP) in 1960 (or $143 per person); by 2001, health care spending accounted for 14.1% of the GDP (or $5,035 per person) [16]. As a result of these increases in demand for and cost of health care in high-income countries, national health systems, insurers, and health care providers are all under strain.

1.2.3 Future health care delivery challenges

As we begin the 21st century, many of the health care challenges described above will continue. Preventable diseases will persist. Inequities in access to health care within and across countries will persist. Health care costs will continue to increase, as will demands for health care. Advances in medical knowledge will continue, along with costly new technologies and medicines. Aging populations will consume increasing amounts of health care services. Patients will have increasing expectations for cures and treatments of more health problems. New means of delivering health care (e.g., telemedicine) will continue to emerge, creating a need for improved communication and information management systems.

In both rich and poor nations, public resources for health care will remain inadequate to meet the demand. Policy makers and health care providers must determine how to provide the most effective health care to citizens using the limited resources that are available. Governments and health care providers must strive to meet basic health needs for all their citizens. Moreover, they must work to improve health and health-related quality of life for citizens in all stages of their lengthening life span. They must set health care priorities (e.g., between disease prevention and treatment or between alternate means of health care delivery) and develop health care systems that can deliver the needed health care in the most effective and efficient manner possible. Worldwide health improved dramatically during the 20th century. The challenge of the 21st century will be to continue this improvement.

1.3 PROVIDING EFFECTIVE AND EFFICIENT HEALTH CARE

To provide the best health care given the limited resources that are available, policy makers need effective methods for planning, prioritization, and decision making, as well as effective methods for management and improvement of health care systems. The planning and management decisions facing policy makers and planners can be grouped into two broad areas: health care planning and organizing, and health care delivery.

1.3.1 Health care planning and organizing

Health care planning and organizing involves relatively high-level policy decisions about the economics of health care systems (e.g., health care resources, pricing, and financing), the structure of health care systems, and other aspects of public policy regarding health care.

Economics of health care systems At the highest level of planning, governments and other health care providers must determine the level of resources they will devote to health care, and how much they will spend on individual patients. Governments must decide which goods and services are to be paid for through public funding and who will receive those goods and services. Because funds are not available to meet all health care needs, governments must set priorities and determine how they will ration the health services they pay for. Health care providers must determine the cost of services and set prices. Government agencies and other large insurers must negotiate prices for drugs and vaccines. Insurers, including governments, must determine who will receive health insurance coverage and what that coverage will consist of. They must develop affordable, workable payment schemes for physicians and other health care providers, and must determine what fees patients must pay for health care services. Such financing schemes must provide proper incentives for health system efficiency.

Structure of health care systems Another set of high-level decisions concerns the structure and organization of health care delivery systems. Health care providers must determine which goods and services they will provide and how to allocate resources among them. Governments must decide to whom the goods and services will be provided. Resources must be allocated among different levels of the health service – for example, among primary care and public health programs versus hospital services. Resources must be allocated between capital development and operating costs, and between salary and nonsalary expenditures. Resources must be allocated among geographic areas – for example, different regions of a country, or urban versus rural areas. Resources must be allocated among specific programs – for example, programs for control of specific diseases, immunization programs, or reproductive health programs. Resources must also be allocated among specific health care goods and services – for example, doctor visits, procedures, or medications.

Other public policy issues In addition to economic and structural issues, decision makers face a variety of other policy decisions that have a broad effect on the delivery of health care. Policy makers must develop strategic plans for national and regional health improvement. These include identifying risks to public health (e.g., environmental contaminants, infectious disease epidemics, or unhealthy lifestyles) and developing plans for mitigating such risks. Such plans may include, for example, national or regional disease screening and prevention programs, health promotion programs, mass vaccination programs, programs to control biological pests (e.g., spraying against malaria-transmitting mosquitoes), programs for the control of illicit drugs, or programs for response to potential bioterrorist attacks. Policy makers must develop plans for the provision of health care that address the availability of and access to health care among those whom the health care system serves, with consideration given to the impact of insurance and regulatory policies on such access. Other population-level policy issues include policies for the allocation of transplant organs among potential recipients, for managing national blood supplies, and for managing national vaccine and pharmaceutical stockpiles.

1.3.2 Health care delivery

Planning and managing health care delivery involves decisions about the management of health care operations and about clinical practice.

Operations management for health care delivery Operations management problems that arise in the delivery of health care are similar in many ways to traditional problems in operations management. These include strategic planning problems such as design of services (e.g., inclusion of neonatal intensive care units in some hospitals, or provision of free-standing urgent care clinics or rural health workers), design of the health care supply chain (e.g., design of a network of hospitals, outpatient clinics, and laboratory services), facility planning and design (e.g., location and layout of hospitals and outpatient clinics, or design of material handling systems), equipment evaluation and selection, process selection, and capacity planning. Other planning problems include demand and capacity forecasting, capacity management, scheduling and workforce planning, job design, and management of the health care supply chain. Managers of health care systems must manage inventory (e.g., drugs, supplies, or blood), measure and manage system performance and quality, and assess the performance of health care technologies. Decision support systems must be designed and implemented to support all of these activities.

Clinical Practice Clinicians face a number of important planning and management problems in the delivery of health care. These include assessing health risks and diagnosing diseases and conditions of individual patients. Clinicians must design and plan treatment for their patients. For example, they must assess how disease is likely to progress in a patient and then they must select appropriate drugs and dosages and design other aspects of a treatment regimen (e.g., surgery, radiation, rehabilitation). Clinicians must determine appropriate disease prevention strategies for individual patients (e.g., vaccination, disease screening, drug treatment, lifestyle changes). The goal of these clinical activities is to provide the highest quality care given the resources that are available. Doing so requires ongoing assessment of clinical quality and well as assessment of the cost and effectiveness of different health care interventions. A recent innovation in clinical practice has been the development of broad-based practice guidelines that specify the recommended standard of care for various diseases and conditions. Such guidelines are developed based on cost-effectiveness analysis of alternative interventions, and vary according to the population and setting (e.g., guidelines for treating a disease in a low-income country will differ from guidelines for treating the same disease in a high-income country). Finally, given the explosion of new medical knowledge, information management and decision support systems can play a crucial role in supporting effective and efficient clinical practice.

1.4 OVERVIEW OF THIS BOOK

Operations research techniques, tools, and theories have long been applied to a wide range of issues and problems in health care. However, to date, no single handbook has synthesized the wide applicability of such techniques and presented future challenges and avenues for research. In fact, practitioners, students, and researchers in this field have had difficulty finding a comprehensive reference that can help them improve their ability to apply such techniques, learn new techniques, explore new issues and challenges, and pursue new research avenues. This handbook aims to fill that need.

This book covers applications of operations research in health care, with particular emphasis on health care delivery. The book is geared toward a multidisciplinary audience that includes OR practitioners, students, scientists and researchers with interest in health care (either new interest or existing expertise), as well as health practitioners (such as clinicians, administrators, and managers), students, scientists, and researchers in health sciences, health administration, public health, health care delivery, and health policy.

Three main areas are covered: (1) health care operations management, (2) public policy and economic analysis, and (3) clinical applications. Within each area, a broad range of topics is addressed. Each chapter details a problem area, a state-of-the-art application, the methodology employed, and research issues raised. Each topic is structured and addressed in such a way that a wide audience – with varying levels of knowledge of the subject area or the methodology employed – will be able to access and use the material presented.

This book covers topics as diverse as hospital capacity planning and management, supply chain management for blood banking, evaluation of hospital efficiency, vaccine pricing policies, national drug control policy, decision making for bioterror preparedness, breast cancer diagnosis, optimal design of radiation treatments, and analysis of asthma treatments. Although they cover diverse topics, all of the chapters show how operations research can be applied to help make health care delivery more effective and efficient.

1.4.1 Health care operations management

The first main section of the book comprises chapters describing the application of OR models to problems in health care operations management. In Chapter 2, Linda Green describes how OR models have been and can be used for hospital capacity planning. In Chapter 3, Mark Daskin and Latoya Dean review the application of facility location models in health care. They also present a novel application of the classical set covering model to the analysis of cytological samples. In Chapter 4, Shane Henderson and Andrew Mason discuss the application of a customized simulation model to assist in decision making by a New Zealand ambulance service. In Chapter 5, William Pierskalla discusses the management of blood bank supply chains. In Chapter 6, Liam O'Neill and Franklin Dexter present a method to identify best practices among hospitals' perioperative services using data envelopment analysis (DEA). In Chapter 7, Yasar Ozcan, Elizabeth Merwin, Kwangsoo Lee, and Joseph Morrissey describe the application of DEA to develop a methodology for analyzing organizational performance of community mental health centers. They also present measures of efficiency that can be used as a basis for improving productivity in behavioral health care. In Chapter 8, Michael Carter and John Blake describe four case studies of simulation applied to problems in hospital operations management. They describe the obstacles encountered in these applications, and the lessons learned.

1.4.2 Public policy and economic analysis

The second main section of the book comprises chapters that illustrate the application of OR to problems of health care policy and economic analysis. In Chapter 9, Rose Baker describes applications of conditional likelihood methods for estimating risks to public health. In Chapter 10, Thitima Kongnakorn and François Sainfort describe how medical outcomes can be modeled in order to facilitate economic analysis of health care policy problems. In Chapter 11, Anke Richter presents three case studies of the application of OR techniques to evaluate the economic consequences and health benefits of new medications and treatments. In Chapter 12, Jonathan

Caulkins provides an overview of the ways in which OR models have been applied to evaluate policies for the control of illicit drugs. In Chapter 13, Gregory Zaric reviews recent OR advances in modeling maintenance treatment programs for opiate addicts. In Chapter 14, Harold Pollack describes how OR models have been used to evaluate syringe exchange programs and substance abuse treatment programs for injection drug users, and how such models can assist policy makers. In Chapter 15, Douglas Owens, Donna Edwards, John Cavallaro, and Ross Shachter apply a simulation model and economic analysis to evaluate the cost effectiveness of potential vaccines against HIV, the virus that causes AIDS. In Chapter 16, Sheldon Jacobson and Edward Sewell review the application of linear programming models to address a variety of economic issues surrounding pediatric vaccine formulary design and pricing. In Chapter 17, Margaret Brandeau reviews OR models that have been developed to assist in the allocation of resources to control infectious diseases. In Chapter 18, Stephen Chick, Sada Soorapanth, and James Koopman evaluate the public health benefits of two interventions for controlling infectious microbes in the water supply – improvements to centralized water treatment facilities, and localized point-of-use treatments in the homes of particularly susceptible individuals. In Chapter 19, Ruth Davies and Sally Brailsford present a model that evaluates policies for public health screening to detect diabetic retinopathy (which is early indications of eye disease caused by diabetes). In Chapter 20, Edward Kaplan and Lawrence Wein review the recent smallpox vaccination policy debate in the U.S., and describe the successful use of OR methods to influence policy in this arena. In Chapter 21, Stefanos Zenios reviews OR models that have been used to evaluate policies for allocating donor kidneys to transplant recipients. In Chapter 22, Mike Cushman and Jonathan Rosenhead describe the application of a model-based approach to the redesign of children's health services in inner London.

1.4.3 Clinical applications

The third main section of the book comprises chapters that describe the application of OR techniques to clinical problems. In Chapter 23, Andrew Schaefer, Matthew Bailey, Steven Shechter, and Mark Roberts review the application of Markov decision process models to guide medical treatment decisions. In Chapter 24, Gordon Hazen describes how dynamic influence diagrams can be applied to model clinical decision problems. In Chapter 25, Elisabeth Paté-Cornell describes the application of risk analysis to evaluate policies for reducing risk during anesthesia procedures. In Chapter 26, David Paltiel, Karen Kuntz, Scott Weiss, and Anne Fuhlbrigge present a model that simulates health and economic outcomes among patients with asthma, and they illustrate the application of the model to assess the cost effectiveness of inhaled corticosterioids among certain adult patients. In

Chapter 27, Daniel Rubin, Elizabeth Burnside, and Ross Shachter present a Bayesian network model that can help radiologists interpret mammograms and determine appropriate followup. In Chapter 28, Eva Lee and Marco Zaider describe an optimization model and decision support system to help plan radiation treatment for patients with cancer. In Chapter 29, Allen Holder describes linear optimization models that can be used to help design radiation treatments. In Chapter 30, Michael Ferris, Jinho Lim, and David Shepard describe the application of Matlab for radiation treatment planning. In Chapter 31, James Koopman, Ximin Lin, Stephen Chick, and Janet Gilsdorf present a transmission model of a common bacteria that colonizes the human nose and throat, and they show how the model can be used to evaluate the relative effectiveness of different vaccines (in particular, vaccines that reduce transmission of the bacteria versus vaccines that prevent disease once a person's throat has been colonized). Finally, in Chapter 32, David Craft, Lawrence Wein, and Dennis Selkoe present a model of the accumulation of amyloid, β-protein (Aβ) in the brain during the course of treatment for Alzheimer's disease, and show how the model can be used to determine appropriate treatments.

1.4.4 Conclusion

In a recent report [6], the World Health Organization stated that, "One of the most important roles of the World Health Organization is to assist countries in making optimum use of scarce health resources." This, too, is a role for operations researchers, as this book demonstrates.

Acknowledgments

M.L.B. was supported by the National Institute on Drug Abuse (Grant Number R-01-DA-15612).

References

[1] United Nations Population Division (1996). *Demographic Indicators, 1950-2050 (The 1996 Revision),* United Nations, New York.

[2] World Bank (2001). *Life expectancy learning module,* World Bank Development Education Program Web, http://www.worldbank.org/depweb/english/modules/social/life/, Accessed February 12, 2003.

[3] Mereu, F. (2002). *Russia: Life expectancy declining,* Radio Free Europe/Radio Liberty, http://www.rferl.org/nca/features/2002/07/05072002141441.asp, Accessed February 12, 2003.

[4] Notzon, F.C., *et al.* (1998). Causes of declining life expectancy in Russia. *Journal of the American Medical Association,* 279, 793-800.

[5] Stanecki, K.A. (2002). *The AIDS Pandemic in the 21st Century,* US Census Bureau, Washington, DC.

[6] World Health Organization (1999). *Removing Obstacles to Healthy Development,* http://www.who.int/infectious_disease_report, Accessed November 21, 2002.

[7] World Health Organization (2000). *Report on the Global HIV/AIDS Epidemic - June 2000,* http://www.unaids.org/epidemic_updated/report, Accessed November 21, 2002.

[8] World Health Organization (2003). *Severe Acute Respiratory Syndrome (SARS) - multi-country outbreak - Update 26,* World Health Organization, http://www.who.int/csr/don/2003_04_10/en/, Accessed April 10, 2003.

[9] Murray, C.J.L. and A. Lopez, 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.* Harvard School of Public Health, Boston, Mass.

[10] World Health Organization (2002). *WHO atlas maps global tobacco epidemic,* World Health Organization, http://www.who.int/mediacentre/releases/pr82/en/, Accessed February 12, 2003.

[11] World Health Organization (2002). *Unfinished business: Global push to save 11 million children,* World Health Organization, http://www.who.int/inf-fs/en/fact119.html, Accessed February 12, 2003.

[12] World Health Organization (2001). *Global Alliancefor Vaccines and Immunization (Fact Sheet No. 169),* World Health Organization, http://www.who.int/inf-fs/en/fact169.html, Accessed February 12, 2003.

[13] Brown, D. (2001). WHO calls for rise in health spending. *Washington Post,* Washington, DC, December 21, A3.

[14] World Health Organization (2000). *World Health Report 2000 - Health Systems: Improving Performance,* World Health Organization, Geneva, Switzerland.

[15] Rutter, T.L. (2003). *Study finds "life gap" in U.S.,* Harvard Public Health Review, Harvard University, http://www.hsph.harvard.edu/review/life_gap.html, Accessed February 28, 2003.

[16] Centers for Medicare and Medicaid Services (2003). *Report details national health care spending increases in 2001,* Centers for Medicare and Medicaid Services, http://cms.hhs.gov/media/press/release.asp?Counter=693, Accessed February 26, 2003.

2 CAPACITY PLANNING AND MANAGEMENT IN HOSPITALS

Linda V. Green

Graduate School of Business
Columbia University
New York, NY 10027

SUMMARY

Faced with diminishing government subsidies, competition, and the increasing influence of managed care, hospitals are under enormous pressure to cut costs. In response to these pressures, many hospitals have made drastic changes including downsizing beds, cutting staff, and merging with other hospitals. These critical capacity decisions generally have been made without the help of OR model-based analyses, routinely used in other service industries, to determine their impact. Not surprisingly, this has often resulted in diminished patient access without any significant reductions in costs. Moreover, payers and patients are increasingly demanding improved clinical outcomes and service quality. These factors, combined with their complex dynamics, make hospitals an important and rich area for the development and use of OR/MS tools and frameworks to help identify capacity needs and ways to use existing capacity more efficiently and effectively. In this chapter we describe the general background and issues involved in hospital capacity planning, provide examples of how OR models can be used to provide important insights into operational strategies and practices, and identify opportunities and challenges for future research.

KEY WORDS

Hospitals, Capacity management, Queueing theory

2.1 INTRODUCTION

2.1.1 Background

Hospitals are the locus of acute episodes of care for most serious illnesses and form the backbone of the emergency medical care system. Over the years, hospitals have been successful in using medical and technical innovations to deliver more effective clinical treatments while reducing patients' time spent in the hospital. However, hospitals are typically rife with inefficiencies and delays. Patients spend hours and sometimes days in emergency rooms and recovery rooms waiting for beds. Procedures and surgeries have to be cancelled and rescheduled. Inpatients are placed in inappropriate beds and transferred multiple times from one unit to another. Nurses and other staff are often in short supply to handle peak loads.

These inefficiencies have their roots in the regulatory and financing environment in which most hospitals existed until recently. Until the mid-1980's, U.S. hospitals were paid by insurers on a "fee for service" basis and capacity expansions were subsidized by state governments. With the increased prevalence of managed care and reduced government subsidies, hospital managers have been under increasing pressure to cut costs and have undertaken large-scale changes to do so. Hospitals have been merged, downsized, and in many cases, closed. Beds have been reorganized, units closed, and patients discharged earlier to increase utilization and throughput. Emergency rooms are getting more crowded and there are increasing reports of ambulance diversions due to a lack of beds. Yet, most hospitals struggle to operate in the black.

In this environment, it is more important than ever for hospital managers to identify ways to "right-size" their facilities and deploy their resources more effectively. Yet, hospitals do not generally use the kind of OR/MS methodologies used in many other service industries to help with capacity planning and management.

2.1.2 Capacity planning in hospitals: overview

The most fundamental measure of hospital capacity is the number of inpatient beds. Hospital bed capacity decisions have traditionally been made based on target occupancy levels – the average percentage of occupied beds. Historically, the most commonly used occupancy target has been 85%. Certain nursing units in the hospital, such as intensive care units (ICUs) are often run at much higher utilization levels because of their high costs.

Until recently, the number of hospital beds was regulated in most states under the Certificate of Need (CON) process, under which hospitals could not be built or expanded without state review and approval. (In the last few years, most of these states have either relaxed or totally eliminated CON bed requirements.) Target occupancy levels were the major basis for these approvals. Though there has been fairly extensive literature on the use of queueing, simulation, and optimization models to support hospital planning [1-6], occupancy targets have been and continue to be the primary measure for determining bed requirements at the individual hospital and even hospital unit level. Faced with increased pressure to be more cost efficient, some hospitals are now setting target levels that exceed 90% without understanding and addressing the issues of bottlenecks and congestion in what is usually a highly stochastic, interdependent system.

The other major component of capacity is personnel, particularly nurses. Nurses are the chief caregivers as well as managers of the clinical units. In recent studies, nursing has been found to have a significant impact on clinical outcomes [7]. In addition, nursing costs comprise a very substantial fraction of hospital budgets. In most hospitals, the number of nurses assigned to a unit is determined by a specified ratio of patients to nurses. The norm for most types of clinical units has been 8:1, while for intensive care units it could be as little as 1:1. Though most hospitals subscribe to these standards, cost pressures and a national nursing shortage have resulted in these ratios being exceeded in many cases. Sometimes, however, this is the result of a failure to adequately plan for the daily, weekly and sometimes seasonal variations in hospital census that are common in most clinical units of virtually every hospital. Though there have been many articles on the use of optimization models to determine nurse staffing (see references in [3, 8, 9]), hospitals often lack basic data, such as patient census by time of day, that would be needed to use such models [10].

Another significant component of capacity is operating rooms. Surgical procedures are usually a critical source of revenues for hospitals. The efficient use of operating rooms, which are often bottlenecks, can be central to the smooth functioning of the hospital as a whole. Substantial work on scheduling operating rooms has appeared in the OR literature (see references in [3, 11, 12]), though there is evidence that this resource is still a source of operational problems.

Major diagnostic equipment, such as magnetic resonance imaging devices (MRIs), comprise another important category of capacity. These machines are extremely expensive, so operating policies are usually oriented toward achieving 100% utilization. In order to avoid "excess" capacity and "unnecessary" usage, these purchases are regulated by the states under a certificate of need (CON) process. Hospital policies governing the use of MRIs are very varied. For example, in some hospitals, outpatients are scheduled on a dedicated facility while in others, inpatients, outpatients and emergency patients all use the same machine. Policies and priority rules are constructed and implemented without any OR analysis and often result in long lead times for outpatient appointments as well as on-site delays. See [13] for a dynamic programming approach to the allocation of capacity for a shared facility.

2.2 AN ILLUSTRATION OF THE ISSUES: EMERGENCY ROOM DELAYS

2.2.1 Understanding the problem

Newspapers, magazines and television have recently reported on severe overcrowding of emergency departments (EDs) and increases in the amount of time that ambulances are being turned away from hospitals [14-16]. Though troubling even on the surface, these reports are even more ominous given the current environment of terrorist threats. So what needs to be done to improve hospitals' ability to respond to emergencies?

Before looking for solutions, it is critical to first understand the nature of the problem. This should begin with the question: "How long should patients wait?" Reports of excessive delays and overcrowding can be very misleading unless there is an understanding of what performance standards should be applied. This, in turn, necessitates an understanding of the potential medical consequences of specific delays for each category of patients. Many patients who arrive to an ED are "non-urgent" and would not be harmed by significant delays in seeing a physician. Most, however, are either "emergent" (requiring "immediate" care) or "urgent" (requiring care within a "short" period of time). Within each of these broad categories, however, there is considerable variety in the exact nature of the illness or injury and extremely little clinical evidence supporting specific delay standards. Unlike, say, telephone call centers, there are no industry-wide standards for what constitutes excessive delays in an ED. Nor are there generally accepted standards for how long a patient requiring admission from the ED should wait for a bed. It is this latter delay that directors of EDs generally cite as most responsible for ED overcrowding and ambulance diversions.

2.2.2 Complexities of capacity planning

Even without specific standards, there is clearly a problem when patients wait for the better part of the day for a bed, when filled stretchers block walkways and hallways, or when a hospital must routinely turn ambulances away. What causes these problems? Though one likely cause (and the one most widely cited in the media) is the reduction of inpatient beds over the last ten years, many other factors must be considered. From a capacity planning perspective, the entire process from patient arrival in the ED to placement in a bed must be examined. Considering only the major steps, the process begins with the triage nurse, who determines the acuity of the patent's condition, and registration which is usually a clerical function. Next, the patient is seen by an ED physician. Often this results in a request for diagnostic testing such as blood analysis and x-rays. Laboratory specimens are generally collected by technicians or nurses and sent to a central testing facility of the hospital. If the patient needs to be taken to another location in the hospital for a diagnostic test, transport personnel are needed. When all tests are completed, the physician reviews them and determines whether the patient requires admission to the hospital. If so, a bed is requested in the appropriate nursing unit (e.g., medical, surgical, intensive care). The availability of a bed is affected not only by the capacity of the relevant unit, but also by the admission and scheduling policies of elective patients, particularly surgical patients who compete for the same beds as many emergency patients [17], and by transfer and discharge policies and procedures. Even if a suitable bed is vacant, it must be located and identified as empty, and then cleaned, if necessary. In addition, a floor nurse must be available to admit the patient. When everything is ready, a request is made for transport and when it is available, the patient is finally moved to the assigned bed. Clearly, there is the potential for a mismatch between the demand and availability of capacity in each step of the process.

This description of the ED admission process illustrates the complexities of hospital capacity planning and management. First, it demonstrates the interdependencies of the various parts of the hospital and the need to identify bottlenecks. These bottlenecks may change from hour to hour, shift to shift, daily, weekly and seasonally. Second, it shows the variety of both fixed capacity (e.g., inpatient beds, ED beds, diagnostic equipment) and variable capacity (e.g., nurses, physicians, technicians, housekeepers, transport staff) that must be managed. Third, much of the capacity required for ED admissions – such as inpatient beds, labs, diagnostic equipment and transport staff – is shared by other patients in the hospital, and thus policies and procedures are required to allocate these resources among the various patient groupings. Fourth, ED admissions are generally time-dependent with distinct time-of-day and day-of-week patterns as well as some seasonality. Therefore, it is imperative that managers develop appropriately flexible staffing policies as well as strategies for using fixed capacity to handle peak loads efficiently and effectively. Finally, in order to create a true emergency response system, capacity needs must be considered on a regional basis and ambulance dispatch and diversion policies developed to assure timely access to care for the most urgent patients. Given that hospitals within the same geographic area are likely to experience many of the same peaks in demand, this means that enough regional capacity should be available so that the probability of all hospitals within a given area being on ambulance diversion simultaneously is extremely small. This is well illustrated by the case of New York City which experienced a severe and protracted citywide shortage of inpatient hospital beds in 1987/1988 [18]. During this period, ambulances were routinely turned away from full hospitals and urgently sick patients experienced delays of days waiting for an open bed.

2.3 HOW MANY HOSPITAL BEDS?

2.3.1 The problem with occupancy levels

As mentioned previously, hospitals often rely on target occupancy levels to plan and evaluate bed capacity. Until recent reports on ED overcrowding and increased ambulance diversion started surfacing, the widespread perception among policymakers and hospital managers was that there were too many hospital beds in the U.S. This belief was primarily supported by the discrepancy between what has usually been considered the "optimal" occupancy figure of 85% (see, e.g., [19], p.55) and the actual average occupancy rate for nonprofit hospitals which has recently been about 64% [20]. This and other related target occupancy levels were originally developed at the federal government level in the 1970's as a response to accelerating health care costs and the perception that more hospital beds resulted in greater demand for hospital care (which was shown to occur under fee for service reimbursement). These occupancy targets were the result of analytical modeling for "typical" hospitals in various size categories and were based on estimates of "acceptable" delays [21].

What is wrong with using occupancy levels to manage capacity? First, reported occupancy levels are generally based on the average "midnight census". This refers to the time when hospitals count patients for billing purposes. However, the midnight census usually measures the lowest occupancy level of the day. One reason is the phenomenon known as the "23-hour patient" who is admitted in the morning and discharged in the evening. Managed care companies have encouraged this practice as a way of allowing evaluation of a patient while avoiding unnecessary hospitalization. More generally, most patients are admitted in the morning or early afternoon and are not discharged until after attending physicians have conducted examinations, so that the peak census is in the middle of the day and can easily be 20% higher than at midnight [22]. In addition, the utilization of hospital facilities is far from uniform across the week or across the year. Very few procedures are scheduled for weekends, so elective patients are not usually admitted on weekends when the average daily census is considerably lower. Summer and holiday periods are also slower [23] and other seasonal effects have been observed in specific hospitals and/or for specific units. Reported occupancy levels are yearly averages and hence do not reflect significantly higher levels that may exist for extensive periods of time. For all of these reasons, reported occupancy levels are not reliable measures of general bed utilization.

More importantly, bed occupancy levels do not measure or even indicate patients' delays for beds. Yet, hospitals do not typically measure bed delays nor do they use queueing or simulation models to estimate the delays that would result from changes in demand or the number or organization of beds.

2.3.2 Target occupancy levels, bed delays and size

Evaluating bed capacity based on a target probability of bed availability or other measure of delay can lead to very different conclusions than would be reached from the use of a target occupancy level. This can be illustrated in considering obstetrics units. Obstetrics is generally operated independently of other services, so its capacity needs can be determined without regard to other parts of the hospital. It is also one for which the use of a standard M/M/s queueing model is quite good. Most obstetrics patients are unscheduled and the assumption of Poisson arrivals has been shown to be a good one in studies of unscheduled hospital admissions [24]. In addition, the coefficient of variation (CV) of length of stay (LOS), which is defined as the ratio of the standard deviation to the mean, is typically very close to 1.0 [6] satisfying the service time assumption of the M/M/s model.

Since obstetrics patients are considered emergent, the American College of Obstetrics and Gynecology (ACOG) recommends that occupancy levels of obstetrics units not exceed 75% [25]. Many hospitals have obstetrics units operating below this level. For example, based on the 1997 Institutional Cost Reports (ICRs), 117 of the 148 or 79% of New York State hospitals had average occupancy levels below this standard. Some have eliminated beds to reduce "excess" capacity and costs [26]. Conversely, fewer than 20% of these hospitals had obstetrics units that would be considered over-utilized by this standard.

But evaluation of capacity based on a delay target leads to a very different conclusion. Though there is no standard delay target, Schneider [27] suggested that the probability of delay for an obstetrics bed should not exceed 1%. Applying this criterion and using the ICR data in an M/M/s model results in 40% of the hospitals having insufficient capacity by this standard. The major reason for this is size. From queueing theory, we know that larger service systems can operate at higher utilization levels than smaller ones while attaining the same level of delays [28]. While obstetrics units are usually not the smallest units in the hospital, there are many small hospitals, particularly in rural areas, and the units in these may only contain 5 to 10 beds. Of the New York State hospitals represented in this data, more than 50% had maternity units with 25 or fewer beds. How large would an obstetrics unit need to be to operate at a 75% occupancy level and have a probability of delay not exceeding 1%? The estimate based on the M/M/s model is that at least 67 beds are needed. Only 3 of the 148, or 2% of the New York hospitals represented in the 1997 ICR reports had units at least this large.

2.3.3 The impact of seasonality

The above discussion illustrates that policies based on target occupancy levels can result in less than desirable access to beds. Indeed, actual results are likely to be worse than described above. This is because the above analyses were based on average annual occupancy levels and obstetrics units typically experience a significant degree of seasonality in admissions. For example, data from Beth Israel Deaconess Hospital in Boston [6] revealed that the average occupancy levels varied from a low of about 68% in January to about 88% in July. With 56 beds, the probability of delay for an obstetrics bed, as estimated from the M/M/s model, for a patient giving birth in January is likely to be negligible, while in July, it would be about 25%. And if, as is likely, there are several days when actual arrivals exceed this latter monthly average by say 10%, this delay probability would shoot up to over 65%. The result of such substantial delays can vary from backups into the labor rooms and patients on stretchers in the hallways to the early discharge of patients. Clearly, hospitals need to plan for this type of predictable demand increase by keeping extra bed capacity that can be used during peak times, or by using "swing" beds that can be shared by clinical units that have countercyclical demand patterns.

2.3.4 The impact of clinical organization

Hospital beds are not all the same. In most general care hospitals, beds are organized into nursing units. A nursing unit generally corresponds to a specific physical location with a dedicated nursing staff headed by a general nurse manager. Each nursing unit is used for one or more clinical services, such as medicine, surgery, cardiology, neurology, and so forth. With the exception of a few services such as pediatrics, obstetrics and psychiatry, which are always operated as dedicated units, hospitals vary in the number and types of nursing units. For example, in some hospitals, nursing units may house both general medical and surgical patients, while others operate strictly dedicated units for each. In addition, hospitals generally have one or more intensive care units (ICUs). Some hospitals have many specific types of ICUs including neurological, surgical, medical and cardiac. One of the distinctive features of

ICUs is that all beds have telemetry so that vital functions can be continually monitored. However, other hospital beds may have telemetry as well and some patients who do not require care in an ICU may nevertheless require a telemetry bed.

Hospital managers are often aware that higher occupancy levels can be achieved if beds are used more flexibly. Hence some have engaged in efforts to cross-train nurses and/or invest in more telemetry in order to treat a greater variety of patient types within a single unit. In addition, small clinical services are often combined with other services because of physical constraints and overhead considerations. For example, cardiac and thoracic surgery patients are often treated in a single unit since thoracic patients are relatively few and require similar nursing skills as cardiac patients. From a strictly operational point of view, is it always beneficial to combine clinical services? What factors need to be considered in evaluating alternative clinical organizations?

As an example, consider the cardiac and thoracic surgery unit of Beth Israel Deaconess Hospital in Boston. Based on data collected for three years, the average arrival rate of cardiac patients in Beth Israel was 1.91 bed requests per day versus .42 for thoracic patients. Since no information was available on the pattern of admissions to these services, we assumed Poisson arrivals. Since most surgical patients are elective, this assumption could result in an overestimate of delays. However, as described in [6], other factors are likely to more than compensate for this. The CV of LOS was sufficiently close to one so that an M/M/s model produces estimates that are sufficiently reliable for examining the relative performance of alternative policies.

Table 2.1a shows the number of beds required to meet several performance targets by each of the two services operating independently as well as in a combined unit. Delay in this context usually measures the time a patient coming out of surgery spends waiting in a recovery unit or intensive care unit until a bed in the surgical unit is available. Long delays are problematic since they cause backups in the operating room and emergency room and can result in surgeries being cancelled and the hospital going on ambulance diversion. Table 2.1a shows that for each delay target, the combined unit results in a savings of one bed out of a total of about 20 beds.

However, this assumes that the admissions policy is the same for all patients. In Beth Israel Deaconess, as in other hospitals, cardiac patients have priority over thoracic patients. Table 2.1b shows the results of using a non-preemptive priority queueing model to estimate delays for both patient types [29]. Focusing on Beth Israel's target of expected delay of less than one day, we see again that 21 beds is the minimum that produces this result. However, the resulting expected delay for the low priority thoracic patients is now more than three days. This long delay is due to the fact that thoracic patients represent less than 20% of the total arrivals and thus will often be bumped in queue by the far more prevalent cardiac patients. Even worse, this predicted expected delay for thoracic patients of 3.2 days is actually an underestimate. This is because the model assumes the same (weighted) average service time for both customer classes while in reality, the higher priority cardiac patients have an average LOS of 7.7 days versus 3.8 for thoracic patients resulting in even longer delays than predicted for the thoracic patients. If an additional bed is added, the resulting delay for thoracic patients goes down to 1.5 days, a more reasonable level, but there will be no savings over operating the units separately. And to maintain a maximum expected delay of one day for each patient group, the combined unit would actually require one more bed than the separate units.

Table 2.1 Cardiac and thoracic surgery utilization and delays

A. Number of beds needed to achieve expected delay (E[D]) service targets

Target	*Cardiac*		*Thoracic*		*Combined*	
Maximum E[D] (Days)	*No. Beds*	*Util-ization*	*No. Beds*	*Util-ization*	*No. Beds*	*Util-ization*
.5	19	.84	4	.40	22	.81
1	19	.84	3	.53	21	.85
2	18	.88	3	.53	20	.89
3	18	.88	3	.53	20	.89

B. Delays when priority given to cardiac patients

	E[D] (Days)			
Number of Beds	*Cardiac*	*Thoracic*	*Overall*	*Utilization*
23	0.17	0.77	0.28	0.78
22	0.28	1.53	0.5	0.81
21	0.47	3.2	0.96	0.85
20	0.77	7.49	1.98	0.89

Therefore, the “increased efficiency” in terms of reduced beds (and thus higher occupancy level) is at best small and may actually be nonexistent. Of course, a unit of just three beds is likely to be inefficient from a physical space and overhead perspective. Therefore, it might be beneficial to operate the two services in one unit but employ a policy, such as a dynamic priority scheme, that would better balance the delays experienced by the two patient types. As a simple example, an admissions policy could give priority to cardiac patients as long as no thoracic patient has been waiting for T days. As soon as this threshold is reached, the policy reverts to first-come, first-served.

Another factor that needs to be considered in evaluating the benefits of a nursing unit with several clinical services is the degree of disparity in the LOS profile of the patients. Smith and Whitt [30] give examples of how combining customers who have different average service times can increase the variance of the service time in the combined queue and result in longer average delays. It is also possible that the average LOS could increase for one or more patient groups due to the reduced expertise that comes with a more generally trained staff.

2.3.5 The seven-day hospital?

In most hospitals, elective procedures and diagnostic testing come to a virtual stop on weekends. As a result, average bed occupancy levels are considerably lower and heavily demanded equipment such as MRIs are idle. Pressures to increase patient throughput are causing hospitals to think about the potential benefits of a "seven-day hospital". On the cost side, scheduling elective procedures and tests on weekends would require additional staffing, perhaps at overtime rates in some cases. What might be gained?

To illustrate the possible impact of a seven-day hospital on capacity needs, consider the case of a surgical intensive care unit (SICU). Most patients in an SICU are elective and therefore admissions drop significantly on the weekend. The data from one such unit, shown in Table 2.2, illustrate a typical pattern, with the average admission rate peaking at 4.42 patients per day on Tuesday and dropping to only 1.44 patients on Sunday. Given this demand profile and an average LOS of 3.05 days, how many beds are needed in this unit?

Using numerical integration to solve the differential equations that describe this nonstationary queueing process, we find that 17 beds are needed to assure that the daily probability of delay is below 10%. Now assume that the same number of admissions is smoothed over the entire seven-day week. Using the average daily arrival rate of 3.34 patients in an M/M/s model indicates that only 15 beds are now needed to meet this target performance. What if 15 beds are used but the demand is not smoothed over the week? Then the nonstationary model indicates that while the average probability of delay over the week would be about 11%,

Table 2.2 Surgical intensive care – Admissions

Day	Admissions/Day
Sunday	1.44
Monday	3.36
Tuesday	4.42
Wednesday	3.59
Thursday	3.92
Friday	4.40
Saturday	2.21
Average	3.34

the daily probability of delay would peak on Fridays at about 18% with an expected delay of over 13 hours for those who are delayed [6]. The result of this might be a backup of patients in the surgical recovery room which could result in the cancellation of some surgeries scheduled for the end of the week. The "optimal" capacity and operating policy could be determined by weighing the expected revenue loss against the alternatives of expanded bed capacity and the additional staffing costs associated with conducting a regular surgical schedule on weekends

2.4 STAFFING THE ED: HOW SHOULD LEVELS VARY ACROSS THE DAY?

2.4.1 Overview

Visits to emergency departments (EDs) have been increasing while the number of emergency departments has been decreasing. This has put a significant strain on the directors of emergency departments to keep patient delays in receiving treatment reasonable. The most critical resource for controlling delays is the physician staff. However, unlike hospital beds, the number of available physicians can be adjusted to accommodate varying arrival volumes.

Hospital managers are aware that arrivals to EDs are very variable with time-of-day, day-of-week and even seasonal patterns. Under federal law, emergency rooms are required to allow all patients access to care 24 hours a day, regardless of ability to pay. Therefore, people who lack health insurance (currently more than 44 million in the U.S.), as well as others who may have difficulty gaining access to primary care physicians, use hospital emergency rooms as their sole source of treatment.

Matching physician capacity to patient needs is critical to the ED's ability to provide timely care to urgently ill or injured patients. Given the substantial variability and unpredictability of demand, as well as the diversity of patients and their medical needs, determining physician staffing levels is very challenging. Yet, as in other areas of the hospital, decisions are not generally based on the use of OR models.

2.4.2 Using queueing models to determine physician staffing: an example

Figure 2.1 illustrates the arrival pattern for the busiest weekday of an ED in a mid-sized urban medical center, which shows a low of about .9 arrivals per hour in the middle of the night to over 5 per hour in the middle of the day. Also shown are physician staffing levels over the day based on the judgment of the ED directors. No explicit data was kept on the duration of physician examination times, and though no data was kept on patients' delays before seeing a physician, delays were observed to be very long, particularly during the late afternoon and evening hours. This resulted in a high rate of "walkouts" - patients who leave after registering but before being seen by a physician - a matter of great concern to the ED directors as well as senior management.

At the time of this study, a request for additional physician hours was under consideration by senior hospital officials. To determine the appropriateness of using queueing models to guide the allocation of any additional staffing, current performance was estimated by using the empirical demand data, the mean physician exam time (estimated to be 45 minutes) and the staffing levels shown in Figure 2.1, and solving the differential equations that describe the time-varying behavior of the system based on Poisson arrivals and exponential service times [31]. In order to represent the true workload in the system, the realized demand for physicians was derived from the arrival data shown in Figure 2.1 by adjusting for walkouts. The walkout rate was about 14.1% over the day. Based on a survey of U.S. ED directors [32] and discussions with ED managers, we adopted as our primary performance measure the probability of delay exceeding one hour, or $\Pr(D > 1)$. Figure 2.2 shows the time-varying behavior of this performance measure resulting from the staffing pattern shown in Figure 2.1 (see [33] for the derivation of this calculation). The results, showing $\Pr(D > 1)$ ranging from a low of .25 at 5 a.m. to a high of .87 at 1 la.m., were considered by the ED managers as consistent with empirical observations.

To help identify the number and scheduling of ED physicians that would yield more acceptable performance, we used a target of $\Pr (D > 1) < .10$.

Figure 2.1 Monday arrival rate and staffing levels

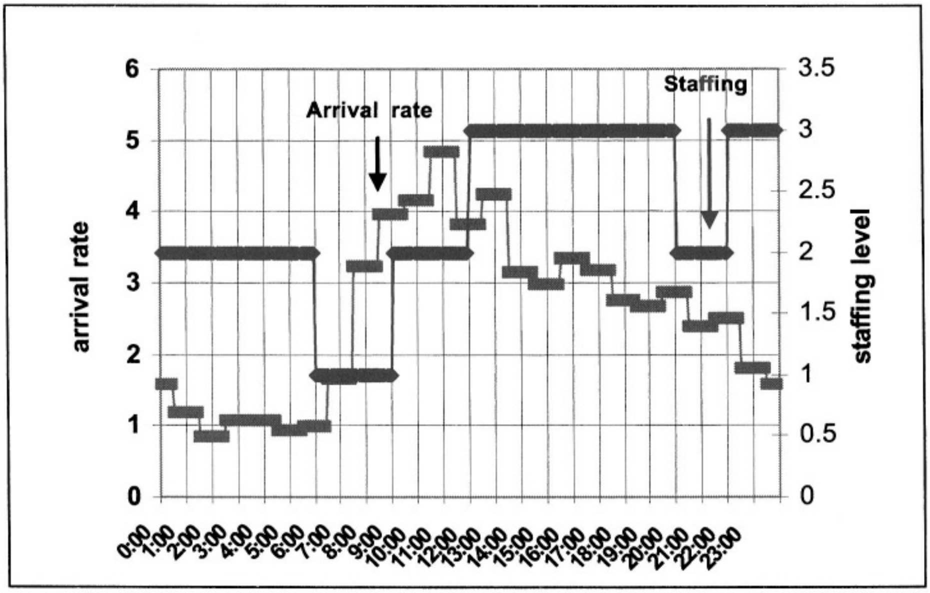
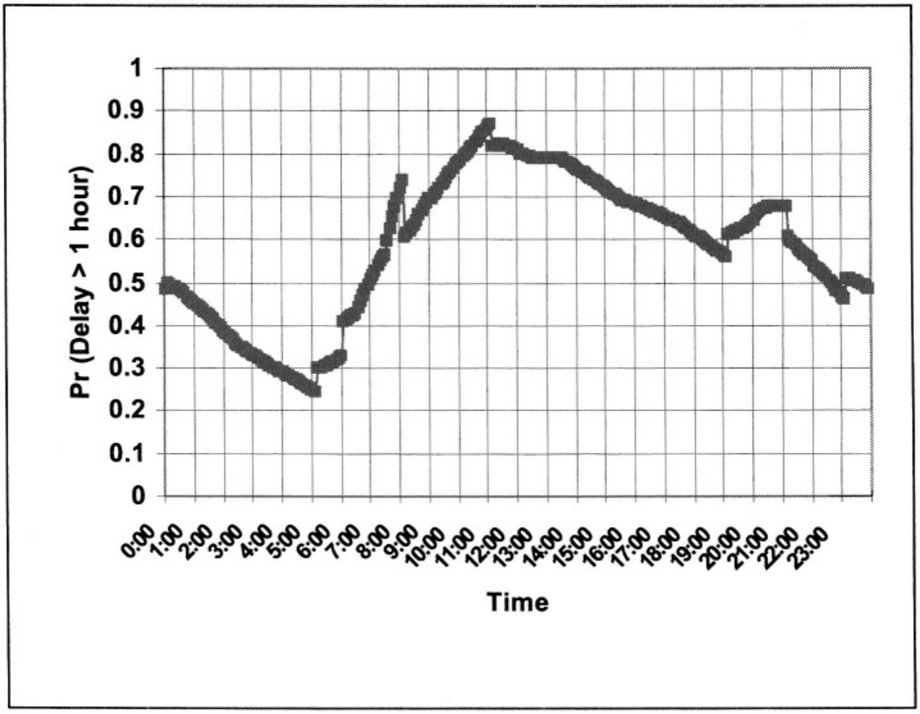


Figure 2.2 Actual staffing levels and estimated Pr (Delay > 1 hour)



Traditionally, in a service system with time-varying arrivals, the desired staffing levels would be determined by the *stationary independent period by period* or SIPP approach which begins by dividing the workday into planning periods, such as shifts, hours, half-hours, or quarter-hours. Then a series of stationary queueing models, most often M/M/s type models, are constructed, one for each planning period. Each of these period-specific models is independently solved for the minimum number of servers needed to meet the service target in that period. In a similar vein, Vassilacopoulos [34] used a dynamic programming model to determine physician staffing levels in an ED assuming that the allocation in each hour should be proportional to the corresponding arrival rate for that hour. In [31], the SIPP approach was shown in many cases to seriously underestimate the number of servers needed to meet a given delay performance target. This is particularly true when the mean service times are high (e.g., 30 minutes or more) and planning periods are long (two hours or more). In these situations, it was demonstrated that a simple variant of SIPP, called Lag SIPP, performs far better than the simple SIPP approach. The major reason is that in cyclical demand systems, there is a time lag between the peak in the arrival rate and the peak in system congestion. This lag is significant when the mean service time is long. Lag SIPP corrects for this factor.

We used both the SIPP and Lag SIPP approaches with the unadjusted empirical arrival data to compare the current staffing levels with the staffing levels the models suggest would be needed to serve the total demand at the targeted level of performance. As expected, both the SIPP and Lag SIPP approaches indicated that current staffing of 55 hours per day would need to increase substantially, by about 63% to meet this target. Though both the SIPP and Lag SIPP methods suggested a total of 90 physician hours per day, the staffing pattern suggested by the SIPP approach resulted in $\Pr(D > 1)$ exceeding the target of .10 by more than 10% for 4 hours of the day and attaining a maximum of .22 for one 2-hour period. In contrast, the Lag SIPP method yielded staffing estimates that met the target delay in every period. Figure 2.3 shows the Lag SIPP proposed staffing levels as well as the predicted $\Pr(D>1)$ curve.

Though the hospital was not in a position to hire this many new physicians, the ED director was interested in the staffing pattern suggested by Figure 2.3. One important insight was that the changes in staffing levels generally lag the changes in the arrival rate by one planning period.

The Lag SIPP model was also used to explore other alternatives. First, the performance target was relaxed so that $\Pr(D > 1) < .2$. In this case, the Lag

Figure 2.3 LAG SIPP Staffing, Pr (Delay> 1 hour) < .10

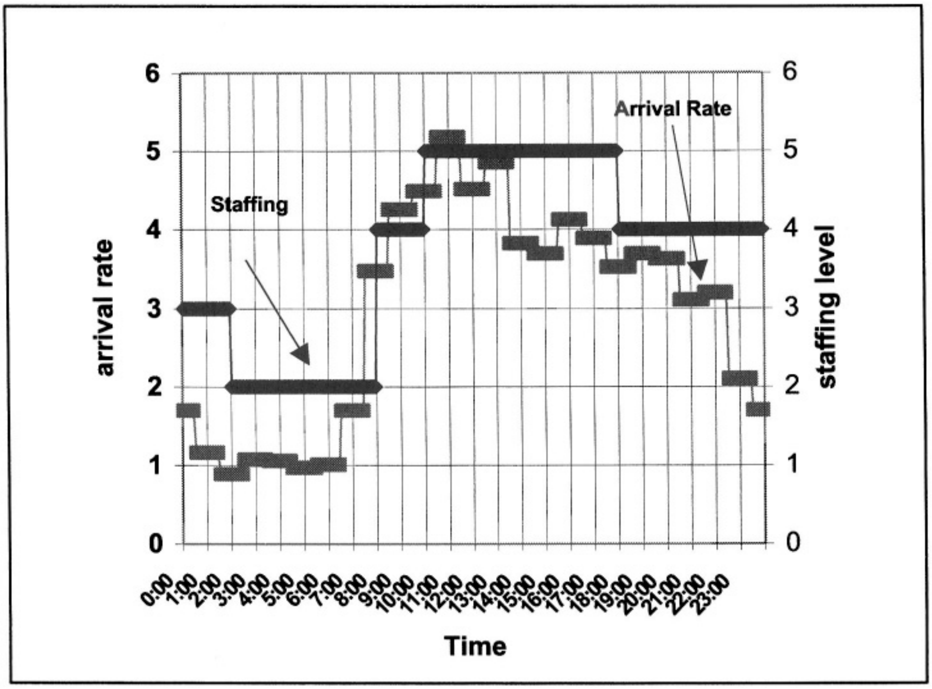
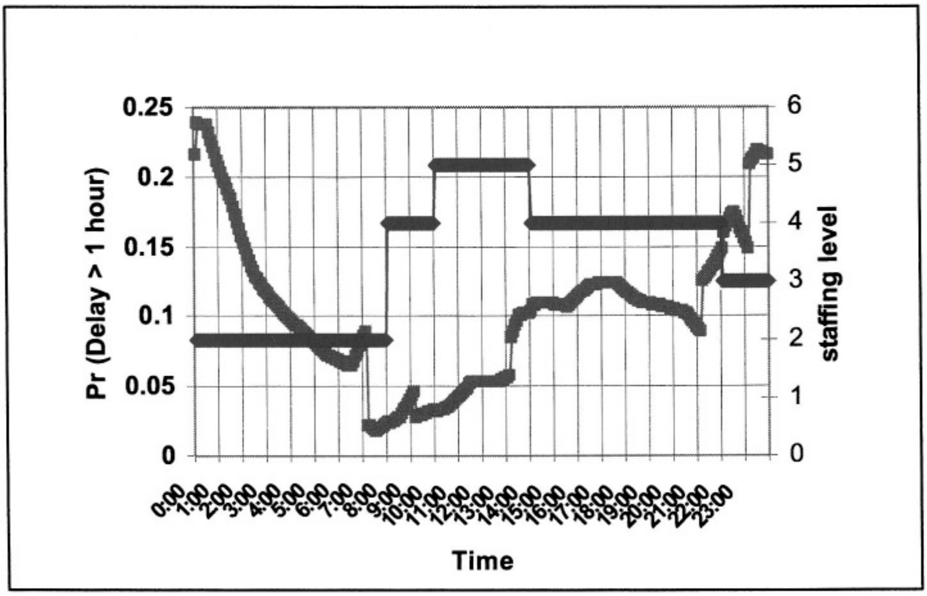


Figure 2.4 LAG SIPP staffing, Pr (Delay > 1 hour) < .2



SIPP results indicated that the staffing would need to increase by about 50% to 82. (Interestingly, the SIPP model suggested a total of 84 physician-hours for this case.) This was still considered too expensive. However, Lag SIPP does not necessarily result in an optimal allocation and looking at the predicted resulting curve for Pr(D>1) shown in Figure 2.4, we noticed that this probability dips significantly between 7 a.m. and 2 p.m. Therefore, we postulated that we could reduce the staffing by one physician in each of the 2-hour periods starting at 8 a.m. The result, shown in Figure 2.5, shows that the delay target is still never exceeded by more than 10% in any 2-hour period. This pattern was used by the ED directors as a guide to reallocating their current physician staff over the day.

To refine the model's recommendations, it would have been helpful to consider priority classes since it is most important that the emergent and urgent patients be seen by a physician within a given time window. However, no reliable data was kept on arrivals by priority class and the hospital had no immediate plans to do so.

2.4.3 Transport staffing: another potential source of delays

Though a lack of appropriate inpatient beds is usually cited as the major reason for ED overcrowding, patients often experience delays even when beds are available. In fact, as illustrated in Figure 2.6, which shows ambulance diversions by time of day for all hospitals in Manhattan from 1999 through 2001, one of the two most frequent times for ED overcrowding and hence diversions is from midnight to 2 a.m. However, this is the time period when hospital occupancy levels are lowest.

One reason for this seeming anomaly was identified in one large New York hospital where a data collection effort showed that the time between bed assignment and the patient leaving the ED peaked from an average of 2.1 hours to between 3 and 4 hours during the midnight to 4 a.m. time interval. Further analysis revealed three reasons for this. First, the demand for transports peaked to about 8 patients per hour starting at midnight from a daytime average of about 7. This counterintuitive finding was due to the combination of peak arrival rates that started at about noon and stayed high until early evening, and an average duration of 8.2 hours between arrival time and bed assignment. However, because ED arrival rates drop to near their lowest during this time, hospital managers had decided that transport staff should be reduced starting at midnight from two to one. In addition, it was found that while the average transport during daytime hours was about 20 minutes, this increased to 27 minutes starting at midnight. This was attributed to the fact that during the day, ED transport personnel were used for transporting patients to diagnostic facilities located within the ED as well

Figure 2.5 Modified lag SIPP staffing, Pr (Delay > 1 hour)

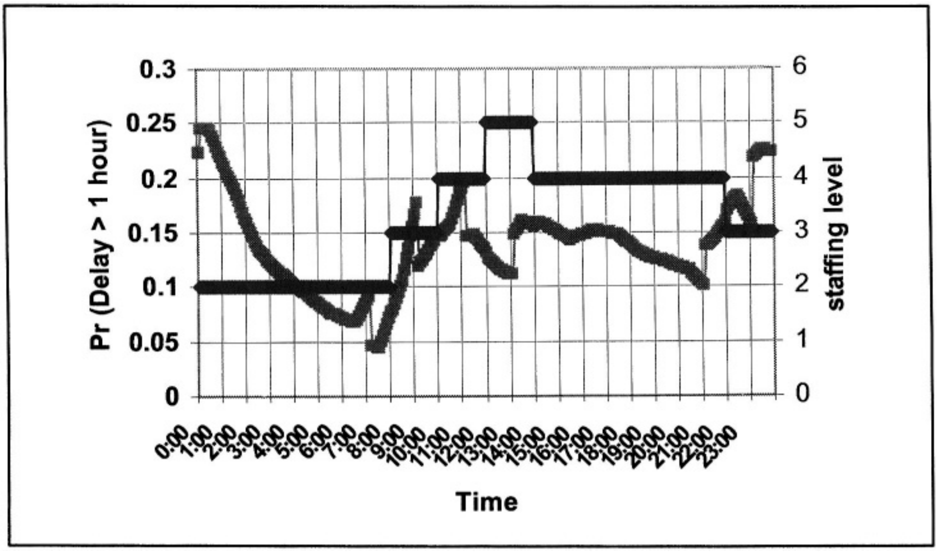
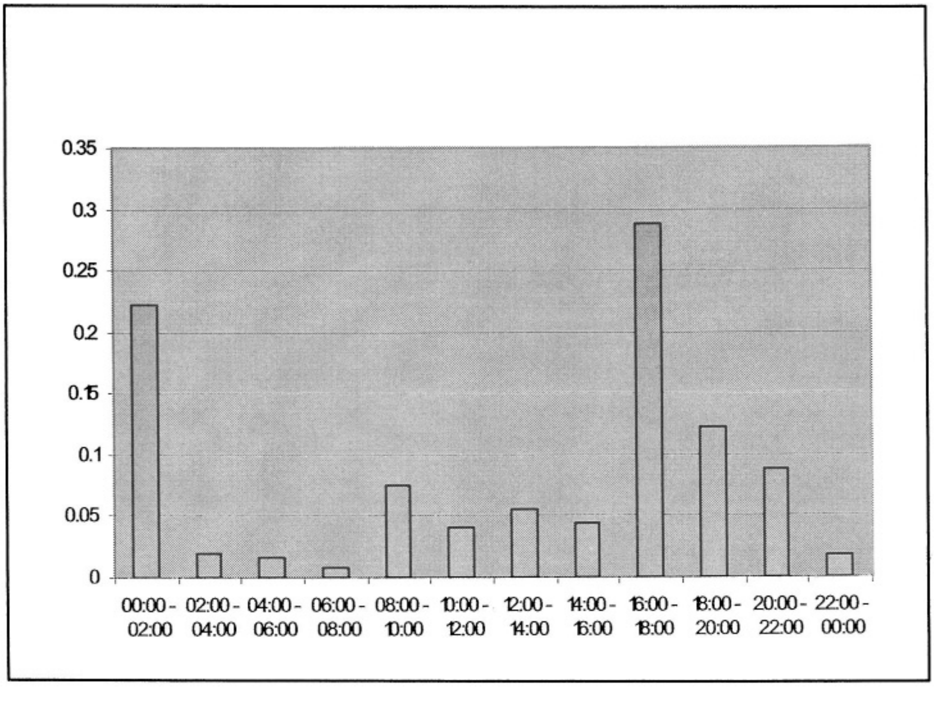


Figure 2.6 Manhattan ambulance diversions (1999-2001) by time of day



as to inpatient beds; while at night, when these facilities are closed, personnel were used only for transporting patients to beds. As a result of a queueing analysis that incorporated these factors, the hospital added a transporter during the midnight to 4 a.m. period with a subsequent average decrease of over an hour in transport delays.

In addition to transport personnel, most hospitals reduce other support staff at midnight. Many of these, such as nurses, who are needed to physically receive patients on the floors, housekeepers, who must make sure beds are prepared, and other personnel who are responsible for locating beds, impact ED delays. The above demonstrates the need to properly identify and analyze the impact of time-varying effects of both demands and processing times throughout the hospital in order to alleviate ED overcrowding

2.5 FUTURE RESEARCH OPPORTUNITIES AND CHALLENGES

2.5.1 Creating flexibility

As indicated in the examples above, patients often experience serious delays due to highly variable patient demands and capacity constraints. Yet, hospitals are often reluctant or unable to add capacity because of cost pressures, regulatory constraints, or a shortage of appropriate personnel. This makes it extremely important to use existing capacity most efficiently. Increasing bed flexibility can be a key strategy in alleviating congestion. However, no comprehensive analysis has evaluated alternatives or identified good policies regarding bed flexibility. Two approaches that have been used in some hospitals are worthy of comprehensive analysis.

As noted before, the degree to which inpatient beds are segregated into nursing units dedicated to one or more clinical services varies across hospitals. From a medical perspective, there may be benefits derived from having patients clustered by diagnostic categories in dedicated units managed and staffed by specialized nurses. These include shorter LOS, fewer adverse events and fewer readmits. Yet, many hospital managers believe that nurses can be successfully cross-trained and that increasing bed flexibility is ultimately in the best interests of patients by increasing speedy access to beds and minimizing the number of bed transfers. By incorporating waiting times, percentage of "off-placements" and the effects on LOS, OR models can be used to address some important research questions dealing with these tradeoffs including:

1. For a given predicted set of demands and a fixed number of nursing units of a given size, how should clinical services be clustered into nursing units?

a. What is the minimum amount of flexibility needed to assure timely access to beds? Can this best be achieved by assigning each clinical service to only one nursing unit, or by allowing some diagnostic categories to be served in multiple units?

b. Which services should be consolidated into a common unit? How should this be affected by LOS characteristics? By nursing requirements? By other resource requirements?

2. For a given nursing unit configuration, what is an optimal real-time bed allocation policy? For example, in the event that there is no appropriate bed available when needed by a new patient, should the patient be placed in another available bed or wait (e.g. in the emergency room or recovery room) until the "right" bed becomes available?

3. When services share a common nursing unit, what admissions policy should be used if there are differing levels of urgency associated with different patient types? For example, in the case of the cardiac and thoracic surgery unit described previously, what type of dynamic priority rule should be used to assure an appropriate level of bed availability for both patient types?

Another approach for increasing bed flexibility is the use of "overflow" units or "swing" beds. These often exist in hospitals that have downsized by closing units without converting them to another use. This results in beds that are not normally staffed but may be used when bed demand increases substantially. A related strategy is to use units that generally have more predictable demand and lower occupancy levels to serve as overflow units for those that frequently fill up. These practices raise several important planning and policy issues such as the following:

1. Given the associated fixed and variable costs, what are the optimal policies for opening and shutting a normally unused overflow unit?

2. How many swing beds should a hospital have and for which clinical services?

3. How should clinical units be used to "back up" each other so as to minimize overall off-service placements without jeopardizing the timely provision of care?

The above strategies increase "horizontal" bed flexibility. Some hospitals have increased "vertical" bed flexibility by reducing the number of different areas in which certain categories of patients reside during their stay. For example, the traditional patient flow model for maternity patients is to move from a labor room to a delivery room to a recovery room and then, finally, to a postpartum bed. Similarly, critically ill patients may spend time in an ICU followed by a "step-down" unit and finally a non-monitored inpatient bed before being discharged. Yet some maternity units have combined labor/delivery/recovery rooms, and some hospitals do not have "step-down" units. OR-based analyses could help shed light on which of these alternatives is more attractive and under what conditions.

2.5.2 Allocating capacity among competing patient groups

Many hospitals provide service to three distinct categories of patients: inpatients, outpatients and emergency patients. These patient groups have differing medical, financial and service requirement profiles, but often require the same set of resources including laboratories, imaging facilities and operating rooms. One important example is magnetic resonance imaging machines (MRIs). A hospital MRI or "magnet" is a very expensive piece of equipment and is critical in diagnosing a broad variety of illnesses, each of which may require a unique examination protocol and duration. For these reasons, utilization of MRIs tends to very heavy and unpredictable and, consequently, significant delays are common. Delays are compounded by late arrivals, cancellations and "no-shows. Operating rooms have very similar characteristics.

Research on operational policies for these types of shared resources could be very useful in increasing their efficiency and service performance. Important questions include:

1. How should outpatient (or elective patient) schedules be designed so as to allow for timely access by emergency patients and/or inpatients without resulting in unacceptable backups?

2. Given the costs of delay for each patient type, what dynamic priority rules are optimal for allocating time slots during the day when more than one type of patient is waiting? (See [13] for some work on this issue.)

3. Assuming that the likelihood of cancellations and "no-shows" increases with the duration of time between when an appointment is made and the scheduled examination date, what is the optimal length of the scheduling horizon?

4. When a hospital has multiple diagnostic or treatment facilities, how many and which patient categories should be assigned to each?

2.5.3 Regional capacity planning

The merger activity of the 1990's has resulted in networks of hospitals within certain geographical regions that have various sorts of contractual commitments to coordinate their planning and activities to some degree. Though these types of associations are often formed primarily to enhance hospitals' bargaining power with payers and suppliers, in some cases an important goal is to streamline and improve the delivery of health care. One possible means of increasing operational efficiency is through clinical consolidation or "regionalization" of one or more clinical services. In other words, it could be advantageous to offer a particular clinical service in a single location. One example of a service that has been considered for such treatment is obstetrics. As discussed above, most obstetrics patients require quick access to beds and most obstetrics units are relatively small. The result is that average occupancy levels must be quite low to provide timely provision of beds. Consolidating obstetrics units across two or more hospitals in a region would clearly result in bed savings and likely result in greater administrative and staffing efficiencies. Other candidates for regionalization are clinical services with small patient demands or those that involve unique technologies and/or skills such as burn units. However, in assessing the desirability of any clinical regionalization, patient travel distances and times must be considered. OR-based analyses could be very helpful in identifying candidate services for regionalization and in determining which hospitals in a given geographic region might be best able to provide a given clinical service.

Another dimension of regional planning is emergency responsiveness. Increasingly, hospitals are coordinating efforts to communicate and respond to unanticipated spikes in demand for emergency department services and inpatient capacity. This has become more of a priority since the events of September 11th, 2001, and the resulting increased concern with preparedness in the event of terrorist attacks. Initial efforts have focused on developing better communications and information systems to collect and disseminate relevant information quickly among hospitals and public agencies. Little attention has been given to identifying which hospitals, clinical units and resources might be vulnerable given sudden, unanticipated surges in demand within and across a given region. (See [26] for some initial work on this issue.) More fundamentally, there is no widely accepted definition of emergency room overcrowding nor agreement on hospital policies for ambulance diversion. Emergency planning is a complex, multi-dimensional issue involving a high degree of unpredictability. The following questions illustrate some broad areas of potential research:

1. How should hospital planning regions be defined? Should this definition differ by clinical service?

2. When should a hospital go on ambulance diversion? How should this be affected by conditions at the other hospitals in the region?

3. How should a hospital's "surge capacity" (the percentage increase in demand above normal levels that can be "adequately" accommodated), be defined and predicted?

2.5.4 Conclusion

Hospital managers are increasingly aware of the need to use their resources as efficiently as possible in order to continue to assure that their institutions survive and prosper. As this chapter has attempted to demonstrate, effective capacity management is critical to this objective as well as to improving patients' ability to receive the most appropriate care in a timely fashion. Yet, effective capacity management must deal with complexities such as tradeoffs between bed flexibility and quality of care, demands from competing sources and types of patients, time-varying demands, and the often differing perspectives of administrators, physicians, nurses and patients. All of these are chronic and pervasive challenges affecting the ability of hospital managers to control the cost and improve the quality of healthcare delivery.

From an analytical perspective, these capacity management issues involve complex dynamics that will require the development of new optimization, queueing and simulation models in order to gain insights to guide strategies and decisions. However, a major obstacle to developing and applying these much needed models is a lack of relevant operational data. Hopefully, as management information systems continue to be developed and enhanced, hospitals will prove to be an extremely rich area for using OR/MS models to improve the quality of healthcare delivery and, perhaps, ultimately save lives as well as money.

References

[1] Worthington, D.J. (1987). Queueing models for hospital waiting lists. *Journal of the Operations Research Society,* 38, 413-422.

[2] Pendergast, J.F. and W. B, Vogel (1988). A multistage model of hospital bed requirements. *Health Services Research,* 23, 381-399.

[3] Smith-Daniels, V.L., S.B. Schweikhart and D.E. Smith-Daniels (1988). Capacity management in health care services: Review and future research directions. *Decision Sciences,* 19, 889-919.

[4] Butler, T.W., K.R. Karwan and J. Sweigart (1992). Multi-level strategic evaluation of hospital plans and decisions. *Journal of the Operational Research Society,* 43, 665-675.

[5] Huang, X. (1995). A planning model for requirement of emergency beds. *IMA Journal of Mathematics Applied in Medicine & Biology,* 12, 345-353.

[6] Green, L.V. and V. Nguyen (2001). Strategies for cutting hospital beds: The impact on patient service. *Health Services Research,* 36, 421-442.

[7] Aiken, L.H., S.P. Clarke, D.M. Sloane, J. Sochalski, and J.H. Silber (2002). Hospital nurse staffing and patient mortality, nurse burnout, and job dissatisfaction. *Journal of the American Medical Association,* 288, 1987-1993.

[8] Kwak, N.K. and C. Lee (1997). A linear programming model for human resource allocation in a health-care organization. *Journal of Medical Systems,* 21, 129-140.

[9] Jaumard, B., F. Semet and T. Vovor, (1998). A generalized linear programming model for nurse scheduling. *European Journal of OperationalResearch,* 107, 1-18.

[10] Green, L.V. and J. Meissner (2002). Developing insights for nurse staffing. Columbia Business School, Working Paper.

[11] Weiss, E.N. (1990). Models for determining estimated start times and case ordering in hospital operating rooms. *IIE Transactions,* 22, 143-150.

[12] Gerchak, Y., D. Gupta and M. Henig (1996). Reservation planning for elective surgery under uncertain demand for emergency surgery. *Management Science,* 42, 321-334.

[13] Green, L.V., S. Savin and B. Wang (2003). Managing competing demands in a medical diagnostic facility. Columbia Business School, Working Paper.

[14] Goldberg, C. (2000). Emergency crews worry as hospitals say, 'no vacancy'. *New York Times,* Dec. 17, 39.

[15] Shute, N. and M.B. Marcus (2001). Code blue: Crisis in the ER. *U.S. News and World Report,* September 10, 55-61.

[16] Anonymous (2002). 1 in 3 hospitals say they divert ambulances. *New York Times,* April 9.

[17] Litvak, E. and M.C. Long (2000). Cost and quality under managed care: Irreconcilable differences? *American Journal of Managed Care,* 6, 305-312.

[18] Myers, L.P., K.S. Fox and B.C. Vladeck (1990). Health services research in a quick and dirty world: The New York City hospital occupancy crisis. *Health Services Research,* 25, 739-755.

[19] Brecher, C. and S. Speizio (1995). *Privatization and Public Hospitals.* Twentieth Century Fund Press, NY.

[20] American Hospital Association (2000). *Hospital Statistics 2000.* American Hospital Association, Chicago, IL.

[21] McClure, W. (1976). *Reducing Excess Hospital Capacity.* Bureau of Health Planning.

[22] LaPierre, S.D., D. Goldsman, R. Cochran, and J. Dubow (1999). Bed allocation techniques based on census data. *Socio-Economic Planning Sciences,* 33, 25-38.

[23] Baker, L., C. Phibbs, J. Reynolds, and D. Supina (2000). Within-year variation in hospital utilization and its implications for hospital costs. Unpublished Manuscript.

[24] Young, J.P. (1965). Stabilization of inpatient bed occupancy through control of admissions. *Journal of the American Hospital Association,* 39, 41-48.

[25] Freeman, R.K. and R.L. Poland (1992). *Guidelines for Perinatal Care.* American College of Obstetricians and Gynecologists, 3rd ed., 14.

[26] Green, L.V. (2003). How many hospital beds? *Inquiry,* 39, 400-412.

[27] Schneider, D. (1981). A methodology for the analysis of comparability of services and financial impact of closure of obstetrics services. *Medical Care,* 19, 395-409.

[28] Whitt, W. (1992). Understanding the efficiency of multi-server service systems. *Management Science,* 38, 708-723.

[29] Cobham, A. (1954). Priority assignment in waiting line problems. *Operations Research,* 2, 70-76.

[30] Smith, D.R. and W. Whitt (1981). Resource sharing for efficiency in traffic systems. *Bell System Technical Journal,* 60, 39-55.

[31] Green, L.V., P. J. Kolesar and A. Svoronos (1991). Some effects of nonstationarity on multi-server Markovian queueing systems. *Operations Research,* 39, 502-511.

[32] Derlet, R.W. and J.R. Richards (2000). Overcrowding in the nation's emergency departments: Complex causes and disturbing effects. *Annals of Emergency Medicine,* 35, 63-68.

[33] Green. L.V. and J. Soares (2003). "Computing tune-dependent waiting time probabilities in nonstationary Markovian queueing systems with variable server staffing. Columbia Business School, Working Paper.

[34] Vassilacopoulis, G. (1985). "Allocating doctors to shifts in an accident and emergency department. *Journal of the Operational Research Society,* 36, 517-523.

3 LOCATION OF HEALTH CARE FACILITIES

Mark S. Daskin[1] and Latoya K. Dean[1]

[1] Department of Industrial Engineering and Management Sciences
Northwestern University
Evanston, IL 60208

SUMMARY

This chapter reviews the location set covering model, maximal covering model and P-median model. These models form the heart of the models used in location planning in health care. The health care and related location literature is then classified into one of three broad areas: accessibility models, adaptability models and availability models. Each class is reviewed and selected formulations are presented. A novel application of the set covering model to the analysis of cytological samples is then discussed. The chapter concludes with directions for future work.

KEY WORDS

Facility location, Covering, Scenario planning

3.1 INTRODUCTION

The location of facilities is critical in both industry and in health care. In industry, poorly located facilities or the use of too many or too few facilities will result in increased expenses and/or degraded customer service. If too many facilities are deployed, capital costs and inventory carrying costs are likely to exceed the desirable value. If too few facilities are used, customer service can be severely degraded. Even if the correct number of facilities is used, poorly sited facilities will result in unnecessarily poor customer service.

In health care, the implications of poor location decisions extend well beyond cost and customer service considerations. If too few facilities are utilized and/or if they are not located well, increases in mortality (death) and morbidity (disease) can result. Thus, facility location modeling takes on an even greater importance when applied to the siting of health care facilities.

This chapter begins with a review of three basic facility location models from which most other models are derived: the set covering model, the maximal covering model, and the *P*-median model. Next, we discuss three major focal points of the location literature as it applies to health care facilities: accessibility, adaptability and availability. In the course of doing so, we review selected models and applications that have appeared in the literature. Our purpose is not to provide a comprehensive survey; rather our goal is to give the reader a feel for the models that have been proposed and the problems to which they have been applied. The reader interested in a more general introduction to facility location modeling should consult [1-4]. More recently Marianov and ReVelle [5] reviewed emergency siting models, Current, Daskin and Schilling [6] summarized general location models, Marianov and Serra [7] discussed the application of facility location models to problems in the public sector and Berman and Krass [8] summarized the state of the art in modeling problems with uncertainty and congestion, two issues we will return to below. We conclude the chapter by discussing an emerging health care application of facility location models that has nothing to do with the location of new physical facilities. We see such applications and adaptations of existing models as an important area for future research.

3.2 BASIC LOCATION MODELS

In this section we review three classic facility location models that form the basis for almost all of the facility location models that are used in health care applications. These are the set covering model, the maximal covering model, and the *P-median* model.

All three models are in the class of discrete facility location models, as opposed to the class of continuous location models. Discrete location models assume that demands can be aggregated to a finite number of discrete points. Thus, we might represent a city by several hundred or even several thousand points or nodes (e.g., census tracts or even census blocks). Similarly, discrete location models assume that there is a finite set of candidate locations or nodes at which facilities can be sited. Continuous location models assume that demands are distributed continuously across a region much the way peanut butter might be spread on a piece of bread. These models do not necessarily assume that demands are uniformly distributed, though this is a common assumption. Likewise, facilities can generally be located anywhere in the region in continuous location models. Throughout this chapter we restrict our attention to discrete location models since they have been used far more extensively in health care location problems.

At the heart of the set covering and maximal covering models is the notion of coverage. Demands at a node are generally said to be covered by a facility located at some other node if the distance between the two nodes is less than or equal to some exogenously specified coverage distance. Typically, the coverage distance is the same for all demand nodes, though additional restrictions on the set of candidate locations that can cover any particular demand node may be added. Such additional restrictions might reflect the ease of travel between population centers and a candidate site for a local clinic. For example, significant elevation changes might be penalized relative to flat terrain [9, 10]. Whether or not additional restrictions are placed on the cover sets, the mathematics is basically the same.

We define an indicator variable as follows:

$$a_{ij} = \begin{cases} 1 & \text{if demand node i can be covered by a facility at candidate site j} \\ 0 & \text{if not} \end{cases}$$

The set covering model [11] attempts to minimize the cost of the facilities that are selected so that all demand nodes are covered. To formulate this model, we need the following additional sets and inputs.

I = set of demand nodes

J = set of candidate facility sites

f_j = fixed cost of locating a facility at candidate site j

In addition, we need the following decision variable.

$$X_j = \begin{cases} 1 & \text{if we locate at candidate site j} \\ 0 & \text{if not} \end{cases}$$

With this notation, we write the set covering problem as follows:

$$\text{Minimize} \quad \sum_{j \in J} f_j X_j \qquad (1)$$

$$\text{Subject to} \quad \sum_{j \in J} a_{ij} X_j \geq 1 \qquad \forall i \in I \qquad (2)$$

$$X_j \in \{0,1\} \qquad \forall j \in J \qquad (3)$$

The objective function (1) minimizes the total cost of all selected facilities. Constraint (2) stipulates that each demand node must be covered by at least one of the selected facilities. The left hand side of (2) represents the total number of selected facilities that can cover demand node *i*. Finally, constraints (3) are standard integrality conditions.

In location problems, we are often interested in minimizing the number of facilities that are located, and not the cost of locating them. Such a situation might arise when the fixed facility costs are approximately equal and the dominant costs are operating costs that depend on the number of located facilities. In that case, the objective function becomes:

$$\text{Minimize} \quad \sum_{j \in J} X_j \qquad (4)$$

To distinguish between these two model variants, we will refer to the problem with (1) as the objective function as the set covering problem or model; when (4) is used, we will call the problem the *location* set covering problem. A number of row and column reduction rules can be applied to the location set covering problem to reduce the size of the problem. Daskin [4] discussed and illustrated these rules.

In practice, at least two major problems occur with the set covering model. First, if (1) is used as the objective function, the cost of covering *all* demands is often prohibitive. If (4) is used as the objective function, the number of facilities required to cover all demands is often too large.

Second, the model fails to distinguish between demand nodes that generate a lot of demand per unit time and those that generate relatively little demand. Clearly, if we cannot cover all demands because the cost of doing so is prohibitive, we would prefer to cover those demand nodes that generate a lot of demand rather than those that generate relatively little demand. These two concerns motivated Church and ReVelle [12] to formulate the maximal covering problem. This model requires the following two additional inputs

h_i = demand at node i

P = number of facilities to locate

as well as the following additional decision variable

$$Z_i = \begin{cases} 1 & \text{if demand node i is covered} \\ 0 & \text{if not} \end{cases}$$

With this additional notation, the maximal covering location problem can be formulated as follows:

$$\text{Maximize} \quad \sum_{i \in I} h_i Z_i \tag{5}$$

$$\text{Subject to} \quad Z_i - \sum_{j \in J} a_{ij} X_j \le 0 \qquad \forall i \in I \tag{6}$$

$$\sum_{j \in J} X_j = P \tag{7}$$

$$X_j \in \{0,1\} \qquad \forall j \in J \tag{8}$$

$$Z_i \in \{0,1\} \qquad \forall i \in I \tag{9}$$

The objective function (5) maximizes the number of covered demands. Again, it is important to note that this model maximizes *demands* that are covered and not simply *nodes*. Constraint (6) states that demand node *i* cannot be counted as covered unless we locate at least one facility that is able to cover the demand node. Constraint (7) states that exactly *P* facilities are to be located and constraints (8) and (9) are standard integrality constraints.

A variety of heuristic and exact algorithms have been proposed for this model. In our experience, Lagrangian relaxation [13, 14] provides the most effective means of solving the problem. When constraint (6) is relaxed, the problem decomposes into two separate problems: one for the coverage variables and one for the location variables. The subproblem for the coverage variables can be solved by inspection and the location variable subproblem requires only sorting. This approach can typically solve instances of the problem with hundreds of demand nodes and candidate sites to optimality in a few seconds or minutes on today's computers even though the problem is technically NP-hard [15, 16]. Schilling, Jayaraman and Barkhi [17] reviewed the general class of location covering models.

The *P*-center model addresses the problem of needing too many facilities to cover all demands by relaxing the service standard (i.e., by increasing the coverage distance). This model finds the location of *P* facilities to minimize the coverage distance subject to a requirement that all demands are covered. Daskin [4] provided a traditional formulation of this problem. More recently, Elloumi, Labbé and Pochet [18] presented an innovative formulation of the problem that exhibits improved computational characteristics when compared to the traditional formulation.

The three models outlined so far – the location set covering model, the maximal covering location model, and the *P*-center model – treat service as binary: a demand node is either covered or not covered. While the notion of coverage is well established in health care applications, in many cases we are interested in the average distance that a client has to travel to receive service or the average distance that a provider must travel to reach his/her patients. To address such problems we turn to the *P*-median problem [19, 20], which minimizes the demand weighted total (or average) distance. To formulate this problem, we need the following additional input

d_{ij} = distance from demand node i to candidate location j

as well as the following new decision variable

$$Y_{ij} = \begin{cases} 1 & \text{if demands at node i are assigned to a facility at candidate site j} \\ 0 & \text{if not} \end{cases}$$

With this notation, the *P*-median problem can be formulated as follows:

$$\text{Minimize} \quad \sum_{j\in J}\sum_{i\in I} h_i d_{ij} Y_{ij} \tag{10}$$

$$\text{Subject to} \quad \sum_{j\in J} Y_{ij} = 1 \qquad \forall i \in I \tag{11}$$

$$Y_{ij} - X_j \leq 0 \qquad \forall i \in I; \forall j \in J \tag{12}$$

$$\sum_{j\in J} X_j = P \tag{13}$$

$$X_j \in \{0,1\} \qquad \forall j \in J \tag{14}$$

$$Y_{ij} \in \{0,1\} \qquad \forall i \in I; \forall j \in J \tag{15}$$

The objective function (10) minimizes the demand weighted total distance. This is equivalent to minimizing the demand weighted average distance since the total demand is a constant. Constraint (11) states that each demand node must be assigned to exactly one facility site. Constraint (12) stipulates that demand nodes can only be assigned to open facility sites. Constraint (13) is identical to (7) above and states that we are to locate exactly *P* facilities. Constraints (14) and (15) are standard integrality constraints. Constraint (15) can be relaxed to a simple non-negativity constraint since each demand node will naturally be assigned to the closest open facility.

As in the case of the maximal covering problem, a variety of heuristic algorithms have been proposed for the *P*-median problem. The two best-known algorithms are the neighborhood search algorithm [21] and the exchange algorithm [22]. More recently, genetic algorithms [23], tabu search [24, 25] and a variable neighborhood search algorithm [26] have been proposed for this problem. Correa et al. [27] developed a genetic algorithm for a capacitated *P*-median problem in which each facility can serve a limited number of demands. They compared their algorithm to a tabu search algorithm and found that the genetic algorithm slightly outperformed the tabu search approach when the GA was accompanied by a heuristic hypermutation procedure. The latter simply performs an exchange algorithm on selected elements of the initial GA population and on population elements at a small number of randomly selected generations.

For moderate-sized problems, Lagrangian relaxation works quite well for the uncapacitated *P*-median problem. Constraint (11) is relaxed resulting in a set of subproblems for each candidate node that can easily be solved by inspection. Daskin [4] outlined the use of Lagrangian relaxation for both the *P*-median problem and the maximal covering model in detail. Daskin [28] reported solution times for a Lagrangian relaxation algorithm for the *P*-median and vertex *P*-center problems with up to 900 nodes.

Some authors have transformed the maximal covering problem into a *P*-median formulation. This can be done by replacing the distance between demand node *i* and candidate site *j* by the following modified distance:

$$\hat{d}_{ij} = \begin{cases} 1 & \text{if } d_{ij} > D_c \\ 0 & \text{if not} \end{cases}$$

where D_c denotes the coverage distance. This has the effect of minimizing the total uncovered demand which is equivalent to maximizing the covered demand.

The uncapacitated fixed charge location (UFL) problem is a close cousin of the *P*-median problem. The UFL problem is derived from the *P*-median problem by eliminating constraint (13) and adding the objective function (1) to objective function (10) multiplied by a suitable constant to convert demand-miles into cost units. The problem then becomes that of determining the optimal number of facilities as well as their locations and the allocation of demands to those facilities to minimize the combined fixed facility location costs and the transport costs.

3.3 LOCATION MODELS IN HEALTH CARE

Having formulated three basic location models (the set covering model, the maximal covering model and the *P*-median model) and having qualitatively discussed two other classical models (the *P*-center problem and the uncapacitated fixed charge model) we now turn to applications and extensions of these models in health care. The health care location literature has tended to address three major topics, which we refer to as *accessibility, adaptability* and *availability.*

By accessibility we mean the ability of patients or clients to reach the health care facility or, in the case of emergency services, the ability of the health care providers to reach patients. Papers that deal with accessibility tend to ignore the needs of the system to evolve in response to changing conditions as well as short-term fluctuations in the availability of service providers as a result of their being busy serving other patients. Papers that focus on availability tend to be direct applications of one or more of the models above or are minor extensions of these models.

It is relatively easy and straightforward to site facilities based on a snapshot of the current or recent past conditions. Unfortunately, there is no guarantee that the future will replicate the past. Predicting future demand rates and operating conditions is exceptionally difficult. Thus, some recent applications and modeling efforts have focused on identifying solutions that can be implemented in the short term but that can adapt to changing future conditions relatively easily.

For some health care systems, and for emergency services in particular, some portion of the nominal capacity is likely to be unusable by new demands at any point in time as it is already in use by current demands. Thus, an ambulance may be busy responding to one emergency when another call for service within its district arises. To handle such situations, a significant literature has focused on designing systems to maximize some measure of the availability of the servers.

In short, accessibility models tend to take a snapshot of the system and plan for those conditions. As such, they are static models. Adaptability models often consider multiple future conditions and try to find good compromise solutions. As such, they tend to take a long-term view of the world. Availability models focus on the short-term balance between the ever-changing demand for services and the supply of those services.

3.3.1 Accessibility models and applications

Accessibility models attempt to find facility locations that perform well with respect to static inputs. In particular, demand, cost and travel distance or travel time data are generally assumed to be fixed and non-random in this class of models. Thus, the models are often relatively straightforward extensions of the classic models outlined in section 1 above.

Indeed, federal legislation has encouraged the use of such models. The EMS (Emergency Medical Services) Act of 1973 stipulated that 95% of service requests had to be served within 30 minutes in a rural area and within 10 minutes in an urban area. This encouraged the use of models like the maximal covering model. Eaton et al. [29] used the maximal covering model to assist planners in Austin, TX in selecting permanent bases for their emergency medical service. The model was solved using the greedy adding and greedy adding and substitution algorithms. More recently, Adenso-Díaz and Rodríguez [30] also used the model to locate ambulances in Leon, Spain. They developed a tabu search algorithm to solve the problem.

Sinuany-Stern et al. [31] and Mehrez et al. [32] used two discrete models, the *P*-median model and a variant of the fixed charge location model in which they constrained the travel time to any hospital and also invoke penalties for the assignment of demand to a hospital in excess of the hospital's capacity. These models were used, along with qualitative techniques, to generate alternative locations, which were then analyzed using the analytic hierarchy method. It is worth noting that the sites that were ultimately preferred tended to be those that were identified using one or more of the analytic methods, as opposed to those identified using qualitative techniques.

Jacobs, Silan and Clemson [33] used a capacitated *P*-median model to optimize collection, testing and distribution of blood products in Virginia and North Carolina. McAleer and Naqvi [34] also used a *P*-median model, in this case to relocate ambulances in Belfast, Ireland. Their problem was to locate four facilities to serve 54 demand nodes. The authors used a heuristic approach that decomposed the demand nodes into four sectors and ranked the possible single facility locations within each sector. This led to a number of acceptable solutions in each sector. All combinations of acceptable locations were then evaluated using all 54 demand nodes. While such a heuristic decomposition approach may make intuitive sense, it is not guaranteed to result in an optimal solution. Modern algorithms (e.g., Lagrangian relaxation embedded in branch and bound as implemented in SITATION [35]) can readily solve such problems to optimality on today's computers in seconds. Practitioners can also use such models to identify near optimal solutions, particularly when the number of facilities being located is small.

In hierarchical location modeling, a number of different services are simultaneously located. These might be, for example, local clinics, community health centers and regional hospitals. Lower level facilities (e.g., clinics) are generally assigned lower numbers (e.g., 1), while the highest level facilities (e.g., regional hospitals) are assigned the top number (e.g., *3*). Another common application of hierarchical modeling is the location of basic life support vehicles (BLS or level 1 facilities) and advanced life support vehicles (ALS or level 2 facilities).

At least three factors need to be considered in hierarchical location problems [36]. The first is whether a level *m* facility can provide only level *m* service or whether or not it can also provide services at all lower levels (1, …, *m*). Clearly, an ALS vehicle can provide all levels of service that a BLS vehicle can provide. It is less clear that a regional hospital will be designed or staffed to provide all levels of support provided by a local clinic. For example, regional hospitals may not stock flu vaccines and, as such, may not be able to vaccinate individuals against the flu, while local clinics may be able to do so. A successively inclusive hierarchy is one in which a level *m* service can provide level *m* and all lower level services, while a successively exclusive hierarchy is one in which each level of service is provided by a unique facility. The second issue is, in a successively inclusive service, whether a level *m* facility can provide all *m* levels of service to *all* demand nodes, or a level *m* facility can provide all *m* levels of service only to demands at the node at which the facility is located and level *m* service only to other nodes. The former is referred to as a successively inclusive service hierarchy while the latter is termed a locally inclusive service hierarchy. A successively exclusive service hierarchy is one in which a level *m* facility provides only level *m* service to all nodes. Finally, there will generally be fewer high level facilities (e.g., regional hospitals) than low level facilities (e.g., local clinics). If high-level facilities can only be located at sites housing a lower level facility, the system is termed nested; otherwise it is not nested.

Finally, Price and Turcotte [37] used a center of gravity model to locate a blood donor clinic in Quebec. The model was used with a variety of inputs to identify a number of different locations from which a final choice was made. The center of gravity model minimizes the demand-weighted average distance between a facility that can be located anywhere in the plane and a discrete set of points. It is in the class of continuous location models (since the single facility location can be anywhere in the plane), which we are not explicitly reviewing and that have seen relatively little use in the health care location field. Nevertheless, Sinuany-Stern et al. [31] and Mehrez et al. [32] used two different continuous models in identifying candidate sites for a new hospital in the Negev. The first was the Weber model, which minimizes the demand weighted average Euclidean distance between a facility that can be anywhere in the plane and fixed demand locations, while the second was similar but used the square of the Weber objective function. (The reader interested in the Weber problem should consult the excellent review by Drezner et al.[38]).

3.3.2 Adaptability models

Location decisions must be robust with respect to uncertain future conditions, particularly for facilities such as hospitals that are difficult if not impossible to relocate as conditions change. A number of approaches have been developed to deal with future uncertainty. Scenario planning [39-41] is frequently used to handle future uncertainty. A number of future conditions are defined and plans are developed that do well in all (or most) scenarios.

In scenario planning, some decisions are made before the true scenario is revealed while others can be made after knowledge of the true scenario is gained. In location planning, the facility sites must generally be chosen before we know which scenario will evolve; the assignment of demand nodes to sites can generally be done after we know which scenario will occur.

Designing a robust system often entails compromises. The "best" compromise plan may not be optimal under any particular scenario but will do well across all scenarios. The *regret* associated with a compromise solution and a scenario measures the difference between the performance measure using the compromise solution for that scenario and the performance measure when the optimal solution is used for that scenario.

Three performance measures are often used in scenario planning: optimizing the expected performance, minimizing the worst case performance, and minimizing the worst case regret. Minimizing the expected regret is identical to optimizing the expected performance.

In what follows, we formulate scenario-based extensions to the P-median problem. We define the following additional set and input

S = set of scenarios

q_s = probability that scenario s will occur

With this additional notation, the problem of minimizing the expected demand weighted total distance is formulated as follows, where we have added the subscript s to the demands and distances as well as the allocation variables:

$$\text{Min} \quad \sum_{s \in S} q_s \left\{ \sum_{j \in J} \sum_{i \in I} h_{is} d_{ijs} Y_{ijs} \right\} \tag{16}$$

$$\text{Subject to} \quad \sum_{j \in J} Y_{ijs} = 1 \qquad \forall i \in I; \forall s \in S \tag{17}$$

$$Y_{ijs} - X_j \leq 0 \qquad \forall i \in I; \forall j \in J; \forall s \in S \tag{18}$$

$$\sum_{j \in J} X_j = P \tag{19}$$

$$X_j \in \{0,1\} \qquad \forall j \in J \tag{20}$$

$$Y_{ijs} \in \{0,1\} \qquad \forall i \in I; \forall j \in J; \forall s \in S \tag{21}$$

The objective function (16) minimizes the expected demand weighted total distance over all scenarios. Constraint (17) states that each demand node is assigned to a facility in each scenario. Constraint (18) stipulates that these assignments can only be made to open facilities. Constraints (19) and (20) are identical to (13) and (14), respectively, and (21) is a standard integrality constraint.

To minimize the worst-case performance, the problem is restructured as follows:

$$\text{Min} \quad W \tag{22}$$

$$\text{Subject to} \quad W - \sum_{j \in J} \sum_{i \in I} h_{is} d_{ijs} Y_{ijs} \geq 0 \qquad \forall s \in S \tag{23}$$

and (17) – (21)

where W is the maximum demand weighted total distance across all scenarios.

Finally, to minimize the maximum regret, we solve the following problem:

$$\text{Min} \quad V \tag{24}$$

$$\text{Subject to} \quad V - \left\{ \sum_{j \in J} \sum_{i \in I} h_{is} d_{ijs} Y_{ijs} - V_s^* \right\} \geq 0 \qquad \forall s \in S \tag{25}$$

and (17) – (21)

where V_s^* is the optimal objective function value (smallest demand weighted total distance) for scenario s.

Both the minimax model (22)-(23) and the minimax regret model (24)-(25) avoid the need for scenario probabilities, which can be difficult to estimate. However, these models suffer from the fact that an unlikely scenario can drive the entire solution. At the other extreme, the problem of minimizing the expected performance (or equivalently the expected regret) tends to undervalue scenarios in which the compromise solution performs poorly if those scenarios are low probability events. To handle these problems, Daskin, Hesse and ReVelle [42] introduced an α-reliable minimax regret model. The model minimizes the maximum regret over an endogenously determined subset of the scenarios whose total probability must be at least α.

Carson and Batta [43] considered the problem of locating an ambulance on the campus of the State University of New York at Buffalo in response to changing daily conditions. This is a particular problem on a large university campus since the center of gravity of the population shifts from dormitories to classrooms and offices over the course of the day. They determined that modeling four different time periods would suffice. By relocating the ambulance for each period, they were able to reduce the predicted average response time by 30% from 3.38 minutes (with a single static location) to 2.28 minutes (with four periods of unequal duration). The actual decrease in travel time when the solution was implemented was closer to 6% with the difference attributed to the non-linear nature of travel times. This work should not technically be viewed as part of the scenario planning literature since the decisions for each time period are unlinked. However, the work does highlight the value of being able to modify ambulance locations in response to changing daily conditions. The work also emphasized the need for careful modeling of travel time relationships, particularly when the average time is likely to be small.

ReVelle, Schweitzer and Snyder [44] proposed a number of variants of a conditional covering model in which demands at a node that houses a facility must be covered by a facility located elsewhere. In such models, the original demand nodes must be covered and each facility located by the model must be covered by a different facility. The rationale for such models is that if an emergency occurs at node *j* (e.g., an earthquake), then any emergency services at that location must be assumed to be damaged or unavailable for service at that node. Therefore, the node must be covered by some other facility.

In many important cases, the actual number of facilities that can be constructed in the long term is uncertain when the planning begins. Then, it is often important to be able to locate a known number of facilities now, accounting for the possibility that additional facilities could be built in future years. Current, Ratick and ReVelle [45] addressed this uncertainty with two models. In each model, the first stage decision entails locating $\boldsymbol{P_0}$ facilities now and $\boldsymbol{P_s}$ facilities in future state *s,* which occurs with probability $\boldsymbol{\pi_s}$. The objective of the first problem is to minimize the expected opportunity loss (or regret) while the second problem minimizes the maximum regret. They illustrated the results using a small problem with 20 nodes, of which 10 were candidate facilities, and 4 future states allowing for 0, 1, 2, or 3 additional facilities to be constructed. The models were solved using a standard LP/IP solver on a personal computer.

3.3.3 Models of facility availability

Adaptability reflects long-term uncertainty about the conditions under which a system will operate. Availability, in contrast, addresses very short-term changes in the condition of the system that result from facilities being busy. Such models are most applicable to emergency service systems (ambulances) in which a vehicle may be busy serving one demand at the time it is needed to respond to another emergency.

Deterministic models One simple, but somewhat crude, way of dealing with vehicle busy periods is to find solutions that cover demand nodes multiple times. The Hierarchical Objective Set Covering (HOSC) model [46] first minimizes the number of facilities needed to cover all demand nodes. Then, from among all the alternate optima to this problem – and there often are multiple alternate optima – the model selects the solution that maximizes the system-wide multiple coverage. The multiple coverage of a node is given by the total number of times the node is covered in addition to the one time needed to satisfy the set covering requirement. The system-wide multiple coverage is the sum of the nodal multiple coverage over all nodes. In essence, the model introduces an explicit surplus variable into constraint (2) and maximizes the sum of the surplus variables as a secondary objective to objective (4).

Benedict [47] modified the HOSC model to account for node demands and termed this **excess** coverage. To do so, he weighted the surplus variable by the node's demand. Eaton et al. [48] independently formulated and solved a similar model for locating ambulances in Santo Domingo. Hogan and ReVelle [49] considered a similar model that they termed **backup** coverage in which only a single additional cover of each node was counted and the additional cover of the node was weighted by the demand at the node.

Benedict also modified the maximal covering model to account for excess coverage. In this model the primary objective is to maximize the covered demand while the secondary objective is to maximize the excess coverage in the system. Benedict's third model was termed the hierarchical objective excess coverage model. In this model, the primary objective is to maximize excess coverage within T time units using the minimum number needed to cover all demand within T; the secondary objective is to maximize the demand that is covered within S, where S is less than T. Daskin, Hogan and ReVelle [50] reviewed a variety of models of multiple, excess and backup coverage as well as the expected covering model discussed below.

Gendreau, Laporte and Semet [51] considered the problem of maximizing the number of demands that are covered by (at least) two ambulances in a distance $r_1 < r_2$ while ensuring that each demand is covered within r_2 and that at least $\alpha\%$ of the demand is covered within r_1. A total of P ambulances are to be located. Like other multiple coverage models, this formulation is designed to increase the likelihood of there being an available ambulance within the coverage distance of a demand. Gendreau, Laporte and Semet solved the problem using tabu search for problem instances with up to 400 demand nodes and 70 candidate sites and 45 facilities.

Pirkul and Schilling [52] developed a model that minimizes the sum of the fixed facility costs, the costs of primary service and the costs of secondary service. Each demand node must be assigned to both a primary and a secondary facility. They developed a Lagrangian heuristic for solving the problem. The algorithm was embedded in a branch and bound algorithm to ensure optimality. They applied the algorithm to test problems ranging in size from 100 demand nodes and 10 candidate sites to 300 demand nodes and 30 candidate locations. They also tested the algorithm on a fire station location problem with 30 candidate sites and 625 demand nodes. By varying the weight on the fixed cost term of the objective function, the tradeoff between the number of facilities located and the average (primary and secondary) distance was identified for this larger problem.

Narasimhan, Pirkul and Schilling [53] considered the problem of locating a fixed number of facilities to maximize the amount of covered demand across a number of different levels of coverage, subject to a constraint that the total demand assigned to a facility across all levels of coverage cannot exceed a given value (the facility capacity). The model converts the maximal covering model into a *P*-median model and then introduces multiple levels of coverage and facility capacities. They argued that this "service level" can represent the order in which the facility providing service is called for service at a node. This is somewhat problematic since the order in which a facility at node *j* is called upon to respond to demands at node *i* depends on the location of other facilities, which is determined endogenously. Specifying this order exogenously seems extraordinarily difficult. They used a Lagrangian approach to solve the problem heuristically relaxing the assignment constraints. The authors solved the problem with up to 200 demand nodes, 30 candidate sites, 5 levels of service and 15 facilities being sited. Optimality gaps tended to be small, though for some (smaller) problems the maximum gap was 3 percent.

Probabilistic models The models discussed above take a deterministic approach to increasing the likelihood that a demand will be covered by an available vehicle or served adequately. Two different probabilistic approaches have been developed. The first approach is based on queuing theory while the second is based on Bernoulli trials.

Fitzsimmons [54] approximated the number of busy ambulances using an $M/G/\infty$ queuing model. The average service time in his model depends on the number of busy vehicles, which, in turn, depends on the average service time. Thus, the two quantities are jointly estimated using an iterative sampling procedure. This is embedded in a search routine for finding improved ambulance locations. Eaton [55] provided an introduction to the use of this model in siting ambulances. While Fitzsimmons' approach can readily be embedded in a heuristic facility location model, it does not fully account for spatial differences in the probability of a vehicle being busy.

To address this shortcoming, Larson [56] developed a hypercube queuing model that accounts for spatially distributed service systems. The hypercube model is essentially an M/M/N queuing model with distinguishable servers. A binary string whose length is equal to the number of servers represents each state of the queuing system. For a system with *n* servers (ambulances) the model requires the solution of 2^n simultaneous linear equations. Larson [57] proposed an approximation to the exact hypercube that entails solving *n* non-linear equations. Because of the difficulty in solving these models with *known* locations, they have tended not to be used in optimization modeling. Jarvis [58], however, embedded an approximation to the hypercube model in a heuristic search algorithm. Brandeau and Larson [59] used the hypercube model to locate ambulances in Boston.

An alternate, though less exact, approach involves representing the probability that a vehicle at any site j will be available as the outcome of a Bernoulli trial with probability of success (available) of *q*. Then, assuming that the probability *q* is the same throughout the system, the probability that all *k* vehicles that can cover a demand node *i* are busy is q^k. The probability that at least one of these *k* vehicles is available is $1-q^k$ and the incremental probability of at least one being available given that *k* vehicles can cover the demand node rather than just *k-1* vehicles is $\left(1-q^k\right)-\left(1-q^{k-1}\right)=q^{k-1}-q^k=q^{k-1}(1-q)$.

This argument is at the heart of the maximum expected covering location model proposed by Daskin [60, 61]. To formulate this model, we define the following decision variable:

$$Y_{ik}=\begin{cases}1 & \text{if demands at node i are covered by k or more vehicles}\\ 0 & \text{if not}\end{cases}$$

With this notation, the maximum expected covering model can be formulated as follows:

$$\text{Max} \qquad \sum_{k=1}^{P}\sum_{i\in I}h_i q^{k-1}(1-q)Y_{ik}=(1-q)\sum_{k=1}^{P}\sum_{i\in I}h_i q^{k-1}Y_{ik} \qquad (26)$$

$$\text{Subject to} \qquad \sum_{k=1}^{P}Y_{ik}-\sum_{j\in J}a_{ij}X_j\le 0 \qquad \forall i\in I \qquad (27)$$

$$\sum_{j \in J} X_j = P \qquad (28)$$

$$X_j \in \{0,1,\ldots,P\} \qquad \forall j \in J \qquad (29)$$

$$Y_{ik} \in \{0,1\} \qquad \forall i \in I; k = 1,\ldots,P \qquad (30)$$

Under the independence assumption implicit in the Bernoulli trials model and the assumption that a single system-wide probability of a vehicle being busy *(q)* can be estimated, the objective function (26) maximizes the expected demand covered by an available vehicle. Constraint (27) links the location variables to the coverage variables and states that a demand node cannot be counted as being covered *k* time unless there are at least *k* vehicles that can cover the node. Constraint (28) states that exactly *P* vehicles are to be located. Constraint (29) states that an integer number of vehicles must be located at any node, and constraint (30) states that the counting variables (Y_{ik}) are binary. Note that constraint (29) does not restrict the number of vehicles at any location to be either 0 or 1. Daskin [61] proposed an exchange-based heuristic that approximates the solution for all values of *q,* the probability of a vehicle being busy.

Repende and Bernardo [62] extended the maximal expected covering location model to incorporate different time periods. The model allowed planners to reduce ambulance response time in Louisville, Kentucky, by 36%. They used simulation to validate the results of the time-variant expected covering model and to get better approximations of the actual expected coverage.

The maximum expected covering location model has two major limitations. Batta, Dolan and Krishnamurthy [63] showed that the independence assumption does not generally hold. They propose a number of ways of handling this including a formulation of an adjusted maximum expected covering location model that uses a correction term similar to that used by Larson [57] in developing the hypercube queuing model approximation. The second limitation of the maximum expected covering model has to do with the computation of the system-wide busy probability. Daskin [61] suggested computing system-wide busy probability as

$q = \frac{\bar{t} \cdot \sum_i h_i}{24 \cdot P}$ where $\bar{t}$ = average service time (in hours).

ReVelle and Hogan [64, 65] extended the computation of the system-wide busy period to account for local conditions by approximating

$$q_i = \frac{\bar{t} \cdot \sum_{r \in M_i} h_r}{24 \sum_{j \in N_i} X_j} = \frac{F_i}{\sum_{j \in N_i} X_j} \tag{31}$$

where

M_i = set of demand nodes that are within the coverage distance of node i

N_i = set of candidate sites that can cover demand node i and

q_i = Probability that a vehicle located at i will be busy

With this local busy probability, ReVelle and Hogan [64] formulated the probabilistic set covering model as follows:

$$\text{Minimize} \qquad \sum_{j \in J} X_j \tag{32}$$

$$\text{Subject to} \qquad \sum_{j \in J} a_{ij} X_j = b_i \qquad \forall i \in I \tag{33}$$

$$X_j \in \{0,1,\ldots\} \qquad \forall j \in J \tag{34}$$

In this model, node *i* must be covered b_i times, where b_i is the is the smallest value satisfying

$$1 - \left(\frac{F_i}{b_i} \right)^{b_i} \geq \alpha$$

and α is the required probability of a node being covered by an available vehicle. Thus, this model is essentially a set covering model except that the right hand side of (33) is greater than 1 and we can locate more than one vehicle at a node. The model finds the minimum number of vehicles required to ensure that each demand node is covered by an available vehicle with probability α, using the local busy probability estimates given by (31).

ReVelle and Hogan [64] defined the α-reliable *P*-center problem and the maximum reliability location problem. The α-reliable *P*-center problem finds the smallest coverage distance such that all demands are covered with probability α by an available vehicle. This is solved by solving the problem above (32)-(34) for successively smaller values of the coverage distance until the objective function exceeds *P*. The maximum reliability location problem is to find the locations of *P* facilities such that the reliability α is maximized. This can be solved by fixing a feasible value of α and then solving the problem above. The value of α is then increased until the required number of vehicles increases above *P*.

Similarly, ReVelle and Hogan [65] formulated the maximum availability location problem (MALP) as the problem of locating *P* vehicles to maximize the number of demands that are covered by an available vehicle with probability at least α. Using the notation defined above, this model becomes:

$$\text{Maximize} \quad \sum_{i \in I} h_i Y_{ib_i} \tag{35}$$

$$\text{Subject to} \quad \sum_{k=1}^{b_i} Y_{ik} \leq \sum_{j \in J} a_{ij} X_j \qquad \forall i \in I \tag{36}$$

$$Y_{ik} \leq Y_{i,k-1} \qquad \forall i \in I; k = 2,\ldots,b_i \tag{37}$$

$$\sum_{j \in J} X_j = P \tag{38}$$

$$X_j \in \{0,1,\ldots,P\} \qquad \forall j \in J \tag{39}$$

$$y_{ik} \in \{0,1\} \qquad \forall i \in I; k = 1,\ldots,b_i \qquad (40)$$

The objective function (35) maximizes the total demand that is covered by an available vehicle with probability at least α. Constraint (36) links the coverage and location variables. Constraint (37) states that a node cannot be counted as being covered *k* times unless it is also counted as being covered *k-1* times. This constraint is not needed in the maximum expected covering problem since the decreasing value of the objective function coefficients for $0<q<1$ ensures that the coverage variables will enter the solution in this order. Constraint (38) states that *P* vehicles are to be located. Constraints (39) and (40) are integrality constraints. Again, we do not limit the number of vehicles located at a node to 0 or 1.

Ball and Lin [66] developed a model that is similar to the maximum availability location problem of ReVelle and Hogan [65], but do so from first principles. This helps identify the assumptions necessary for the development of the model. They then outlined a number of constraints that can be added to the formulation to tighten its linear programming relaxation, thereby facilitating the solution of the problem.

Goldberg et al. [67] developed a highly non-linear model that accounts for vehicle busy periods as a function of assignments. Assignments are for the k^{th} vehicle to respond to a demand in a region. The model was solved heuristically and was applied to the location of ambulances in Tucson, AZ. The model objectives include maximizing the number of calls responded to in 8 minutes (success rate), maximizing the worst node's success rate, and balancing workload. The approach was used primarily to evaluate a given set of sites though they did do some limited experimentation with an exchange algorithm.

Mandell [68] formulated a hierarchical ambulance location model in which demands are not covered unless either (1) a basic life support (BLS) unit can arrive at the scene within t^B *and* an advanced life support unit (ALS) can arrive within t^A with $t^A>t^B$ or (2) an ALS unit can arrive within t^B. The model was formulated in terms of the probability that a demand is served adequately given that there are *r* ALS units within t^A, *r'* ALS units within t^B and *s* BLS units within t^B. Mandel used a two-dimensional Markov model (with states representing the number of ALS units within t^A of a demand node and the number of BLS units within t^B of the node) to estimate the required probabilities. The Markov model used demand-area specific arrival rates. The model was tested on a 55-node network. Computation times were under 1.5 seconds in all cases for the IP problem as formulated.

In the models described above, the primary objective was to account for vehicle busy periods. Another source of randomness arises from the location of the demands. Recognizing that demands occur over a region and not at discrete points, Aly and White [69] considered a probabilistic extension of the set covering model and of the *P*-median model. In both models the location of demands is uncertain, making the travel times random variables. Demand locations are uniformly distributed in rectangular regions. The distribution of travel time to a random point from a base with given coordinates is derived. From this the probability of being able to cover demands in a region from the base within a given time limit is derived. This results in the probabilistic set covering model – minimize the number of facilities need to ensure that each region is covered with probability γ – becoming a standard set covering model. Similarly, once we have the distribution of travel times, we can compute the expected travel time from a base at *j* with known coordinates to a point that is randomly distributed in some rectangular region *i*. This makes the probabilistic *P*-median problem – minimize the demand weighted expected travel time – a standard *P*-median problem as well. They concluded that the probabilistic formulation requires more facilities than does the deterministic formulation. Specifically, they stated, “In summary, using an aggregate point to represent a densely populated area may yield a less expensive siting cost. However, by ignoring the probabilistic element the actual service level will be much less than the one anticipated by the decision-maker.” (p. 1176)

Whether it arises from uncertain demand locations, vehicle busy periods, or changing and uncertain underlying conditions, stochasticity will degrade the performance of the system for a fixed set of resources.

3.4 ANOTHER APPLICATION OF LOCATION MODELS IN HEALTH CARE

The location set covering model – objective function (4) subject to (2) and (3) – has recently been used in a new health care application. Laporte et al. [70] reported on the use of this model to determine the minimum number of fields of view (FOV) to read a cytological sample (PAP test). A field of view is the area that a microscope can see without moving the slide being analyzed. All areas of interest on a slide need to be examined (i.e., need to be in at least one FOV). At the same time, one would like to minimize the number of required FOVs so as to minimize the time needed to analyze each sample.

While the set covering model used by Laporte et al. is identical to that used in the location problems discussed above, there is an important difference. Typical location problems involve several hundred demand nodes and candidate locations. Solution time is not generally a problem in these instances because the problems are small and they do not have to be solved in real time. In the cytological example, the number of points to be covered can range from 2,500 to 55,000, approximately two orders of magnitude more than is typically found in a facility location example. Furthermore, the problems have to be solved very quickly as decisions about how to read a sample need to be made in real time. Furthermore, once appropriate FOVs have been identified, a routing problem needs to be solved to guide the microscope from one FOV to the next.

Laporte et al. [70] employed a series of heuristics to attack the problem. First, a mesh of FOVs was generated to cover all of the points of interest. Within each square, the smallest rectangle containing all of the points in the square was identified and up to four additional FOVs were generated, one located at each of the corners of this rectangle. A number of heuristics were then used to identify FOVs to include in the solution and others that could be excluded. Then a greedy heuristic proposed by Balas and Ho [71] was applied to solve the remaining problem. The routing heuristic was a straightforward adaptation of the strip heuristic proposed by Daganzo [72]. Solution times for the combined heuristic were typically under two minutes and thus were satisfactory for this application.

Brotcorne, Laporte and Semet [73] subsequently developed even faster heuristics for the tiling problem. It is worth noting that the best results in terms of a compromise between solution quality and execution time were generally not those that involve using the heuristic solution to the set covering model; instead, they used a variety of improvement heuristics.

3.5 SUMMARY AND DIRECTIONS FOR FUTURE WORK

In this chapter we have presented the formulations of three location models that underlie most of the facility location models used in health care. The set covering model finds the minimum number (or cost) of facilities needed to cover all demands within a specified time or distance. The maximal covering location model relaxes the condition that all demands must be served within the covering standard and maximizes the number of covered demands using a fixed number of facilities. Finally, the *P*-median model drops the notion of coverage and minimizes the demand-weighted total distance between demand nodes and the nearest facilities.

We identified three approaches to location modeling that have been used in health care applications. Accessibility models are typically straightforward extensions or applications of one of the basic location models. The goals of accessibility models are generally to maximize coverage or to minimize average distance. Adaptability models recognize that future conditions are difficult, if not impossible, to predict. These models attempt to find solutions that perform well across a range of future scenarios. Generally, a single set of locations must be identified for all scenarios, but the assignment of demands to facilities can be scenario-dependent. Typical objectives include optimizing the expected system performance, minimizing the worst-case performance, and minimizing the maximum regret. Regret measures the difference in the performance of the system for a given scenario between the compromise solution and the solution that would have been optimal for the specified scenario. Availability models attempt to account for the short-term unavailability of vehicles or facilities. Many such models have been applied to ambulance location problems. An ambulance might not be available when called upon for service because it is already serving another demand. A variety of deterministic, queuing-based and probabilistic availability models were reviewed.

We also outlined a health care application of the set covering model that results in problems that are approximately two orders of magnitude bigger than typical location problems and that has to be solved in real time. The application has to do with screening cytological samples and finding the minimum number of fields of view needed to read a sample.

In our view, the accessibility literature and the availability literature are quite mature, at least as applied to health care location problems. Considerably less work has been done on applying well-known concepts of scenario planning, or adaptability modeling, to health care problems. This seems to be a potentially fertile area for future work. Related to this is the area of reliability modeling. Reliability differs from adaptability in that adaptability (or robustness as it is sometimes termed) refers to the ability of a system to perform well in the face of uncertain future conditions. The uncertainty is typically in the input conditions including the costs and demands. Reliability, on the other hand, refers to the ability of a system to perform well when parts of the designed system fail [74]. Failures might result from capacity limitations or simply facility closures. Menezes, Berman and Krass [75] discussed reliability problems associated with Toronto hospitals. They noted that it is common for emergency rooms to be at capacity and to request that the citywide system redirect emergencies to some other facility. Also, some hospitals were actually closed due to the SARS outbreak. Daskin and Snyder [76] presented two extensions of the P-median model designed to consider reliability, while Snyder [74] formulated and solved a variety or reliability extensions to location models. We believe that adaptability, robustness and reliability will become increasingly important in future applications in health care.

Finally, the application of location constructs to problems that do not involve locating any facilities seems to be an exciting area for future research and development. The use of location models in improving the efficiency of cytological diagnostic procedures outlined above is but one example of this line of research. Another application involves locating radioactive sources or seeds in the treatment of prostate cancer [77]. Applications of facility location-like models in the diagnosis and treatment of medical conditions is likely to be an important area of future work.

Acknowledgments

This work was supported by grant number DMI-0121815 from the National Science Foundation as well as the NSF-MRSEC Program, grant number DMR # 00706097. This support is gratefully acknowledged. In addition, the authors thank Larry Snyder and Leyla Ozsen for their editorial comments.

References

[1] Handler, G.Y. and P.B. Mirchandani (1979). *Location on Networks: Theory andAlgorithms.* MIT Press, Cambridge, MA.

[2] Love, R.F., J.G. Morris, and G.O. Wesolowsky (1988). *Facilities Location: Models andMethods.* North Holland, New York.

[3] Francis, R.L., L.F. McGinnis, and J.A. White (1992). *Facility Layout and Location: An Analytical Approach.* Prentice Hall, Englewood Cliffs, NJ.

[4] Daskin, M.S. (1995). *Network and Discrete Location: Models, Algorithms andApplications.* John Wiley, New York.

[5] Marianov, V. and C. ReVelle (1995). Siting emergency services, in *Facility Location: A Survey ofApplications and Methods, Z.* Drezner, Ed., Springer, New York.

[6] Current, J., M. Daskin, and D. Schilling (2002). Discrete network location models, in *Facility Location: Applications and Theory, Z.* Drezner and H.W. Hamacher, Eds., Springer, Berlin.

[7] Marianov, V. and D. Serra (2002). Location problems in the public sector, in *Facility Location: Applications and Theory, Z.* Drezner and H.W. Hamacher, Eds., Springer, Berlin.

[8] Berman, O. and D. Krass (2002). Facility location problems with stochastic demands and congestion, in *Facility Location: Applications and Theory, Z.* Drezner and H.W. Hamacher, Eds., Springer, Berlin.

[9] Eaton, D., R. Church, V. Bennett, B. Hamon, and L.G. Valencia (1982). On deployment of health resources in rural Valle del Cauca, Colombia, in *Planning and Development Processes in the Third World,* W. Cook, Ed., Elsevier, Amsterdam.

[10] Bennett, V., D. Eaton, and R. Church (1982). Selecting sites for Rural health workers. *Social Science andMedicine,* 16, 63-72.

[11] Toregas, C.S.R., C. ReVelle, and L. Bergman (1971). The location of emergency service facilities. *Operations Research,* 19, 1363-1373.

[12] Church, R. and C. ReVelle (1974). The maximal covering location problem. *Papers of the Regional Science Association,* 32, 101-118.

[13] Fisher, M.L. (1981). The Lagrangian relaxation method for solving integer programming problems. *Management Science,* 27, 1-18.

[14] Fisher, M.L. (1985). An applications oriented guide to Lagrangian relaxation. *Interfaces,* 15, 2-21.

[15] Garey, M.R. and D.S. Johnson (1979). *Computers and Intractability: A Guide to the Theory of NP-Completeness.* W. H. Freeman, New York.

[16] Megiddo, N., E. Zemel, and S.L. Hakimi (1983). The maximal coverage location problem. *SIAM Journal of Algebra and Discrete Methods,* 4, 253-261.

[17] Schilling, D., V. Jayaraman, and R. Barkhi (1993). A review of covering problems in facility location. *Location Science,* 1, 25-56.

[18] Elloumi, S., M. Labbé, and Y. Pochet (2001). *New formulation and resolution method for the P-center problem,* Optimization Online, http://www.optimization-online.com/DB_HTML/2001/10/394.html.

[19] Hakimi, S.L. (1964). Optimum locations of switching centers and the absolute centers and medians of a graph. *Operations Research,* 12, 450-459.

[20] Hakimi, S.L. (1965). Optimum distribution of switching centers in a communication network and some related graph theoretic problems. *Operations Research,* 13, 462-475.

[21] Maranzana, F.E. (1964). On the location of supply points to minimize transport costs. *Operations Research Quarterly,* 15, 261-270.

[22] Tietz, M.B. and P. Bart (1968). Heuristic methods for estimating generalized vertex median of a weighted graph. *Operations Research,* 16, 955-961.

[23] Bozkaya, B., J. Zhang, and E. Erkut (2002). An efficient genetic algorithm for the P-median problem, in *Facility Location: Applications and Theory,* Z. Drezner and H. Hamacher, Eds., Springer, Berlin.

[24] Voss, S. (1996). A reverse elimination approach for the P-median problem. *Studies in Locational Analysis,* 8, 49-58.

[25] Rolland, E., J. Current, and D. Schilling (1996). An efficient tabu search procedure for the P-median problem. *European Journal of Operational Research,* 96, 329-342.

[26] Hansen, P. and N. Mladenovic (1997). Variable neighborhood Search for the P-median. *Location Science,* 5, 207-226.

[27] Correa, E.S., M.T. Steiner, A.A. Freitas, and C. Carnieri (2004). A genetic algorithm for solving a capacitated P-median problem. forthcoming in *Numerical Algorithms.*

[28] Daskin, M. (2000). A new approach to solving the vertex P-center problem to optimality: Algorithm and computational results. *Communications of the Operations Research Society of Japan,* 45, 428-436.

[29] Eaton, D., M. Daskin, D. Simmons, B. Bulloch, and G. Jansma (1985). Determining emergency medical service vehicle deployment in Austin, Texas. *Interfaces,* 15, 96-108.

[30] Adenso-Díaz, B. and F. Rodríguez (1997). A simple search heuristic for the MCLP: Application to the location of ambulance bases in a rural region. *Omega, International Journal of Management Science,* 25, 181-187.

[31] Sinuany-Stern, Z., A. Mehrez, A.-G. Tal, and B. Shemuel (1995). The location of a hospital in a rural region: The case of the Negev. *Location Science,* 3, 255-266.

[32] Mehrez, A., Z. Sinuany-Stern, A.-G. Tal, and B. Shemuel (1996). On the implementation of quantitative facility location models: The Case of a hospital in a rural region. *Journal of the Operational Research Society,* 47, 612-625.

[33] Jacobs, D.A., M.N. Silan, and B.A. Clemson (1996). An Analysis of alternative locations and service areas of American Red Cross blood facilities. *Interfaces,* 26, 40-50.

[34] McAleer, W.E. and I.A. Naqvi (1994). The relocation of ambulance stations: A successful case study. *European Journal of Operational Research,* 75, 582-588.

[35] Daskin, M. (2002). *SITATION Location Software,* Department of Industrial Engineering and Management Sciences, Northwestern University, http://users.iems.nwu.edu/~msdaskin/.

[36] Narula, S.C. (1986). Minisum hierarchical location-allocation problems on a network: A survey. *Annals of Operations Research,* 6, 257-272.

[37] Price, W.L. and M. Turcotte (1986). Locating a blood bank *Interfaces,* 16, 17-26.

[38] Drezner, Z., K. Klamroth, A. Schöbel, and G.O. Wesolowsky (2002). The Weber problem, in *Facility Location: Applications and Theory,* Z. Drezner and H.W. Hamacher, Eds., Springer, Berlin.

[39] Kouvelis, P. and G. Yu (1996). *Robust Discrete Optimization and its Applications.* Kluwer Academic Publishers, Dordrecht.

[40] Ringland, G. (1998). *Scenario Planning: Managing for the Future.* John Wiley, New York.

[41] Sheppard, E.S. (1974). A conceptual framework for dynamic location-allocation analysis. *Environment and Planning A,* 6, 547-564.

[42] Daskin, M., S.M. Hesse, and C.S. ReVelle (1997). α-Reliable P-minimax regret: A new model for strategic facility location modeling. *Location Science,* 5, 227-246.

[43] Carson, Y.M. and R. Batta (1990). Locating an ambulance on the Amherst campus of the State University of New York at Buffalo. *Interfaces,* 20, 43-49.

[44] ReVelle, C., J. Schweitzer, and S. Snyder (1994). The maximal conditional covering problem. *INFOR,* 34, 77-91.

[45] Current, J., S. Ratick, and C.S. ReVelle (1998). Dynamic facility location when the total number of facilities is uncertain: A decision analysis approach. *European Journal of Operational Research,* 110, 597-609.

[46] Daskin, M. and E. Stern (1981). A hierarchical objective set covering model for EMS vehicle deployment. *Transportation Science,* 15, 137-152.

[47] Benedict, J.M. (1983). *Three Hierarchical Objective Models Which Incorporate the Concept of Excess Coverage to Locate EMS Vehicles or Hospitals.* Department of Civil Engineering, Northwestern University, Evanston, IL.

[48] Eaton, D.J., H.M. Sanchez, R.R. Lantigua, and J. Morgan (1986). Determining ambulance deployment in Santo Domingo, Dominican Republic. *Journal of the Operational Research Society,* 37, 113-126.

[49] Hogan, K. and C. ReVelle (1986). Concepts and applications of backup coverage. *Management Science,* 32, 1434-1444.

[50] Daskin, M., K. Hogan, and C. ReVelle (1988). Integration of multiple, excess, backup and expected covering models. *Environment and Planning B: Planning and Design,* 15, 15-35.

[51] Gendreau, M., G. Laporte, and F. Semet (1997). Solving an ambulance location model by tabu search. *Location Science,* 5, 75-88.

[52] Pirkul, H. and D.A. Schilling (1988). The siting of emergency service facilities with workload capacities and backup service. *Management Science,* 34, 896-908.

[53] Narasimhan, S., H. Pirkul, and D.A. Schilling (1992). Capacitated emergency facility siting with multiple levels of backup. *Annals of Operations Research,* 40, 323-337.

[54] Fitzsimmons, J.A. (1973). A methodology for emergency ambulance deployment. *Management Science,* 19, 627-636.

[55] Eaton, D.J. (1979). *Location Techniques for Emergency Medical Service Vehicles: Volume I – An Analytical Framework for Austin, Texas,* University of Texas, Austin, TX.

[56] Larson, R.C. (1974). A hypercube queueing model to facility location and redistricting in urban emergency services. *Computers and Operations Research,* 1, 67-95.

[57] Larson, R.C. (1975). Approximating the performance of urban emergency service systems. *Operations Research,* 23, 845-868.

[58] Jarvis, J.P. (1975). *Optimization in Stochastic Service Systems with Distinguishable Servers,* Doctoral Dissertation, Massachusetts Institute of Technology, Cambridge, MA.

[59] Brandeau, M.L. and R.C. Larson (1986). Extending and applying the hypercube queueing model to deploy ambulances in Boston, in *Management Science and the Delivery of Urban Service,* E. Ignall and A.J. Swersey, Eds., TIMS Studies in the Management Sciences Series, North-Holland/Elsevier.

[60] Daskin, M.S. (1982). Application of an expected covering model to EMS system design. *Decision Sciences,* 13, 416-439.

[61] Daskin, M.S. (1983). A maximum expected covering location model: Formulation, properties and heuristic solution. *Transportation Science,* 17, 48-70.

[62] Repende, J.F. and J.J. Bernardo (1994). Developing and validating a decision support system for locating emergency medical vehicles in Louisville, Kentucky. *European Journal of Operational Research,* 75, 567-581.

[63] Batta, R., J.M. Dolan, and N.N. Krishnamurthy (1989). The maximal expected covering location problem: Revisited. *Transportation Science,* 23, 277-287.

[64] ReVelle, C. and K. Hogan (1989). The maximum reliability location problem and **α-reliable** P-center problem: Derivatives of the probabilistic location set covering problem. *Annals of Operations Research,* 18, 155-174.

[65] ReVelle, C. and K. Hogan (1989). The maximum availability location problem. *Transportation Science,* 23, 192-200.

[66] Ball, M.O. and F.L. Lin (1993). A reliability model applied to emergency service vehicle location. *Operations Research,* 41, 18-36.

[67] Goldberg, J., *et al.* (1990). Validating and applying a model for locating emergency medical vehicles in Tucson, AZ. *European Journal of Operational Research,* 49, 308-324.

[68] Mandell, M. (1998). Covering models for two-tiered emergency medical services systems. *Location Science,* 6, 355-368.

[69] Aly, A.A. and J.A. White (1978). Probabilistic formulation of the emergency service location problem. *Journal of the Operational Research Society,* 29, 1167-1179.

[70] Laporte, G., F. Semet, V.V. Dadeshidze, and L.J. Olsson (1998). A tiling and routing heuristic for the screening of cytological samples. *Journal of the Operational Research Society,* 49, 1233-1238.

[71] Balas, E. and A. Ho (1980). Set covering algorithms using cutting planes, heuristics, and subgradient optimization: a computational study. *Math Programming Study,* 12, 37-60.

[72] Daganzo, C.F. (1984). The length of tours in zones of different shapes. *Transportation Research B,* 18B, 135-145.

[73] Brotcorne, L., G. Laporte, and F. Semet (2002). Fast heuristics for large scale covering-location problems. *Computers and Operations Research,* 29, 651-665.

[74] Snyder, L. (2003). *Supply Chain Robustness and Reliability: Models and Algorithms.* Department of Industrial Engineering and Management Sciences, Northwestern University, Evanston, IL.

[75] Menezes, M., O. Berman, and D. Krass (2003). The Median problem with unreliable facilities. *EURO/Informs Meeting.* Istanbul, Turkey.

[76] Daskin, M. and L. Snyder (2003). The reliability P-median problem. *EURO/Informs Meeting.* Istanbul, Turkey.

[77] Lee, E.K., R.J. Gallagher, D. Silvern, C.-S. Wuu, and M. Zaider (1999). Treatment planning for brachytherapy: An integer programming model, two computational approaches and experiments with permanent prostate implant planning. *Physics in Medicine and Biology,* 44, 145-165.

4 AMBULANCE SERVICE PLANNING: SIMULATION AND DATA VISUALISATION

Shane G. Henderson[1] and Andrew J. Mason[2]

[1]Department of Operations Research and Industrial Engineering
Cornell University
Ithaca, NY 14853

[2]Department of Engineering Science
University of Auckland
Auckland, New Zealand

SUMMARY

The ambulance-planning problem includes operational decisions such as choice of dispatching policy, strategic decisions such as where ambulances should be stationed and at what times they should operate, and tactical decisions such as station location selection. Any solution to this problem requires careful balancing of political, economic and medical objectives. Quantitative decision processes are becoming increasingly important in providing public accountability for the resource decisions that have to be made. This chapter discusses a simulation and analysis software tool 'BARTSIM' that was developed as a decision support tool for use within the St. John Ambulance Service (Auckland Region) in New Zealand (St. Johns). The novel features incorporated within this study include

- the use of a detailed time-varying travel model for modelling travel times in the simulation,
- methods for reducing the computational overhead associated with computing time-dependent shortest paths in the travel model,
- the direct reuse of real data as recorded in a database (trace-driven simulation), and
- the development of a geographic information sub-system (GIS) within BARTSIM that provides spatial visualisation of both historical data and the results of what-if simulations.

Our experience with St. Johns, and discussions with emergency operators in Australia, North America, and Europe, suggest that emergency services do not have good tools to support their operations management at all levels (operational, strategic and tactical). Our experience has shown that a customized system such as BARTSIM can successfully combine GIS and simulation approaches to provide a quantitative decision support tool highly valued by management. Further evidence of the value of our system is provided by the recent selection of BARTSIM by the Metropolitan Ambulance Service for simulation of their operations in Melbourne, Australia. This work has led to the development of BARTSIM's successor, SIREN (Simulation for Improving Response times in Emergency Networks), which includes many enhancements to handle the greater complexities of the Melbourne operations.

KEY WORDS

Ambulance service planning, Simulation

4.1 INTRODUCTION

In 1997 we were contacted by the St. Johns Ambulance Service (Auckland region) in New Zealand, henceforth referred to as St. Johns. St. Johns wanted assistance in developing rosters for their ambulance personnel. This initial contact led to our study of ambulance service management, and to the development of a comprehensive simulation and analysis tool to assist in decision making. (We should emphasize that here the word "simulation" refers to a computer software tool, and not to the replication of realistic incident conditions where volunteers pretend to have certain injuries.) This chapter reviews some of the issues faced by St. Johns managers, and indeed ambulance service managers all over the world, and discusses the methods and tools that we developed to assist them.

The manager of an ambulance service faces a host of difficult policy questions related to operation of the service. The following list is only a sample.

- How many ambulances should be employed and where should they be stationed?

- What policies and procedures should be followed as calls for assistance are received in order to ensure rapid response to calls while obtaining quality information to allow appropriate dispatching?

- Should ambulances be used for non-urgent patient transfers in addition to the usual emergency response function?

- How should dispatching decisions be made when multiple vehicles are available for dispatch?

- How can one examine the tradeoffs associated with sharing a limited number of ambulances between a high-demand metropolitan area and a low-demand rural area? Here the issue is "fairness" in the sense of coverage, versus "efficiency" in the sense of placing ambulances where they will be in high demand.

This is a rather daunting list of problems, to which a great deal of research effort has been focused in the past. Swersey [1] provides a survey of work in emergency service planning that serves as an excellent entry point for the literature. There is a very large literature on such problems, so one might very well ask, what is the motivation for revisiting these problems?

A key difference between the ambulance-planning problem as faced before 1994 and the problem as faced today is the prevalence of data. Virtually all ambulance operations now employ some form of computer-aided dispatch (CAD) system that automatically logs the details of calls as they are received. This information is a veritable goldmine for planners! Without CAD data, ambulance studies typically relied on manual collection of data; see, for example, Swoveland et al. [2], where some of the required data was manually recorded over a period of two weeks.

A second factor that motivated much of the developments discussed in this chapter is the difference in the questions that are being asked. Much of the early development of ambulance theory focused on the questions of where and when ambulances should be operated. While this question is central to much of what we do, we are also motivated by "finer granularity" questions such as how call taking and dispatching should be performed.

To answer these and other questions at St. Johns, we developed a discrete-event simulation of ambulance operations. By manipulating the parameters of the simulation, it is possible to address, in a quantitative manner, many of the questions mentioned earlier. The flexibility of discrete-event simulation means that one can avoid simplifying assumptions that are otherwise needed to obtain performance measure predictions using other methods, such as queueing theory or Markov chain analysis. Perhaps the biggest advantage of simulation is that it is easy to explain as a decision tool to both managers and frontline personnel, so that after they understand the model, they place great store in its results. Obtaining this "buy-in" from decision makers and frontline personnel is crucial in moving from model predictions to decisions and implementation.

To reinforce these points, consider the hypercube model as surveyed in Larson and Odoni [3], and the specialization of this model to ambulance planning in Brandeau and Larson [4]. The hypercube model, while possessing great predictive power, also requires several assumptions with regard to the way that ambulances are dispatched, gives only steady-state results, and requires certain assumptions about the form of "service time" distributions, at least in the case where calls queue when all units are busy. Moreover, explaining how it works to managers is a somewhat daunting task, so that it is hard to instill a feeling of confidence in decision makers as to its predictions. In spite of these disadvantages, it seems to work very well in practice, so it remains a powerful modeling approach that, for a subset of the questions considered here, is a viable alternative to simulation.

Of course, simulation is not new to the ambulance-planning problem. Early examples are Savas [5] for ambulance operations in New York City, and Fitzsimmons [6, 7] for operations in the San Fernando Valley in Los Angeles. Swoveland et al. [2] used simulation to fit the parameters of a metamodel that predicts expected ambulance response time. The expected response time as predicted by the metamodel was then optimized using branch and bound. Simulation was used by Fujiwara et al. [8] to carefully examine a small number of alternative plans that were obtained from an optimization model developed in Daskin [9]. Lubicz and Mielczarek [10] developed a simulation model of rural ambulance operations in Poland. Ingolfsson, Erkut and Budge [11] used simulation to help in siting a "single-start station," i.e., a station from which multiple ambulances begin their shifts. In addition, the use of simulation as a tool to validate the selections of optimization models is almost universal in the literature, and continues to this day. For recent examples see Erkut et al. [12], Harewood [13] and Ingolfsson, Budge and Erkut [14]. For a recent survey of optimization methods in ambulance location problems see Brotcorne, Laporte and Semet [15]. Larson and Odoni ([3], Chapter 7) discuss general considerations related to the simulation of problems similar in form to the ambulance-planning problem.

So what is new in this study?

First, our simulation directly reuses the data recorded in the CAD database. Real calls are fed through the simulation, rather than calls generated using the usual simulation techniques. Justification for our use of trace-driven simulation and discussion of some of the key issues can be found in Section 4.3. Such an approach resolves many difficulties, including accurate modeling of the complex dependence structure of the information related to calls including time of occurrence, location, need for transport and so forth, Of course, it also introduces other problems.

Second, we employ a sophisticated model, adapted from a model developed and used by the Auckland Regional Council [16] for regional planning purposes, to compute travel times. These travel times are used to determine which ambulance to dispatch to a call, the travel time for the ambulance to reach the call, and so forth. The effort we devote to this topic is justified by the great sensitivity of results to travel time assumptions, as noted both by the authors in a preliminary queueing analysis, and by a large proportion of the papers dealing with ambulance planning. For example, Carson and Batta [17] describe how the 30% savings predicted by their model turned into a 6% savings in actual tests, primarily due to the model not effectively capturing a certain travel time/distance relationship. The use of a simpler model based on the "square root law" [18, 19] or other approximations leads to rather large errors due to the highly irregular geography of Auckland; it is basically an isthmus between two oceans, containing many dormant volcano vents. The complex waterways and vents provide significant barriers to travel, leading to a somewhat convoluted road network. A further complication is that travel times are heavily time-dependent. The simulation makes extensive use of the travel model, and we employ several heuristics to reduce the computational effort involved. Many of the techniques used here could be used in other applications requiring travel time calculations where the travel time is time-dependent.

Third, we employ a geographic information system (GIS) to display simulation results and to examine historical performance calculated from real data. To our surprise, none of the ambulance service providers that we have talked with have used such tools in the past, and all have been tremendously excited by their potential. This has occurred in spite of the growing number of sites where a GIS is being used to draw insights from recorded data; see Peters and Hall [20]. Of course, GISs have been used many times to obtain input for simulation models (see, e.g., [21]), but GISs are not often used for displaying discrete-event simulation output. The graphical displays produced by GIS programs allow decision makers to digest copious amounts of information that were previously given in large tables. GIS output displays are currently under-utilized in discrete-event simulation studies, perhaps because of the form of the models involved. But as the ability to link discrete-event simulation software, databases, and standard GIS packages together increases, the use of GIS output display should become more prevalent.

We have been contacted many times by individuals interested in applying BARTSIM methodology to planning problems in the other emergency services, namely fire and police departments. There are many potential applications to these areas from the work presented here, and we believe that such extensions could be tremendously helpful from the practical standpoint. However, it is important to recognize some of the vital differences in these problems from the ambulance-planning problem. These differences mean that substantial effort would be required to tailor the planning methods used here. For example, the utilization rates of fire appliances are typically on the order of a few percent, while it is not uncommon to have ambulance utilization rates, at least in metropolitan areas in New Zealand, as high as 60%. In terms of police patrol planning, an important function of police patrols is to maintain police visibility, so the problems one faces can be quite different.

The remainder of this chapter is organized as follows. In Section 4.2 we discuss some of the particulars of the St. Johns problem, and outline the process that is followed when St. Johns receives an emergency call. Section 4.3 provides an overview of the simulation model underlying BARTSIM and describes some of the data-reuse issues alluded to above. Section 4.4 describes the travel model and the heuristics used to reduce computational overhead. In Section 4.5 we introduce BARTSIM itself, outline some of its GIS-based analysis capabilities, and describe how these analysis capabilities were used to provide useful insights into several decisions faced by St. Johns. Conclusions and suggestions for future research are offered in Section 4.6.

Further details on BARTSIM can be found on the BARTSIM web site (www.esc.auckland.ac.nz/stjohn).

4.2 THE PROBLEM FACED BY ST. JOHNS

St. Johns contracts to Crown Health Enterprises to supply emergency medical transport. The contracts stipulate that St. Johns supplies a minimum level of service as specified by certain performance targets. These targets relate to response time, which is defined as the time interval between receiving a call to the time that an ambulance first arrives at the scene. The performance targets are broken down by the location of the call (whether the call is in metropolitan Auckland, or in a rural area) and the priority of the call. St. Johns classifies its emergency calls, as opposed to patient transfers and other non-emergency calls, into two levels. Priority 1 calls are those for which an ambulance should respond at all possible speed, including the use of lights and sirens. Priority 2 calls are calls for which an ambulance may respond at standard traffic speeds. The performance targets that St. Johns faces are shown in Table 4.1.

Table 4.1 Contractual service targets

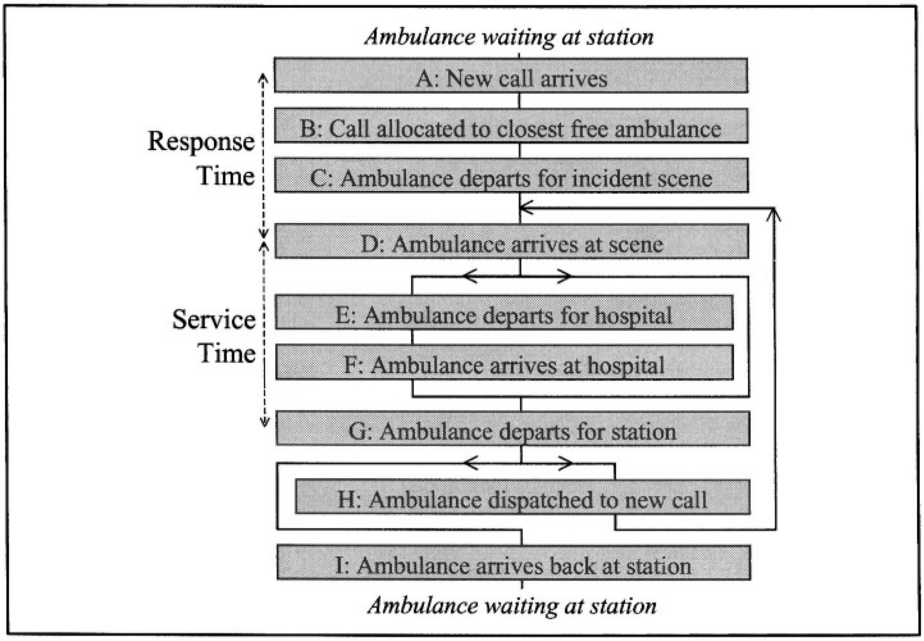
	Priority 1 Calls	**Priority 2 Calls**
Metropolitan	80% in 10 minutes 95% in 20 minutes	80% in 30 minutes
Rural	80% in 16 minutes 95% in 30 minutes	80% in 45 minutes

It is interesting to note that no guidance is given in the contract as to how these figures need to be interpreted. Interpreting the targets as applying, for example, to the entire Auckland area over the entire year in aggregate will lead to far lower resource requirements than assuming, for example, that the targets must be met in each suburb during each hour of each day. One of the goals of this project has been to develop tools to assist management in exploring performance under a range of possible interpretations of the contract.

St. Johns uses a computer-aided dispatch (CAD) system that logs, in a database, information on every call that is received. The database then enables St. Johns to prepare monthly reports that describe how well they meet their performance targets. When St. Johns first contacted us, these reports indicated that the organization was finding it more and more difficult to meet its service targets. It was (and continues to be) believed that this is primarily due to increasing congestion on Auckland roads.

To fully understand these service targets it is necessary to understand the ambulance dispatch and service delivery process. Figure 4.1 shows this process, and identifies the contractual response time discussed earlier. This flowchart also helps to explain the key steps that are captured within the simulation model. When a call arrives at St. Johns, staff in the control room identify an available ambulance (i.e., an ambulance either idle at its base station or returning from a previous job) and dispatch this vehicle to the scene. After initial treatment at the scene, the ambulance typically transports the patient to a hospital, performs a 'handover' to hospital staff, and then returns to its base station. If transport is not required, the ambulance returns directly to its base from the scene. In either case, the vehicle is considered available to receive calls as soon as it begins returning to base.

Figure 4.1 The ambulance dispatch and service delivery process



Ambulance waiting at station
A: New call arrives
B: Call allocated to closest free ambulance
C: Ambulance departs for incident scene
D: Ambulance arrives at scene
E: Ambulance departs for hospital
F: Ambulance arrives at hospital
G: Ambulance departs for station
H: Ambulance dispatched to new call
I: Ambulance arrives back at station
Ambulance waiting at station
Response Time
Service Time

4.3 THE SIMULATION MODEL

The simulation model is written using a high-level programming language without using specialist simulation software. The simulation is trace-driven, and ambulances are routed using a time-dependent travel model. Each of these aspects of the simulation is now discussed in more detail.

We decided not to use an “off-the-shelf package for simulating St. Johns’ operations for several reasons. First, the logical complexity of the decisions that must be made within the model would be difficult to code in a standard package. For example, the dispatcher may redirect an ambulance that is responding to a Priority 2 call to a Priority 1 call. Such a decision requires detailed knowledge of travel times, ambulance locations and so forth. This decision is far easier to code using custom software in a high-level language (C) than standard simulation packages. The second reason was speed. The simulation must be very fast to facilitate the large number of what-if analyses that need to be performed. Consequently, we decided to code the simulation in C, and then embed the simulation program within a custom-developed Microsoft Visual C++ application to provide a user-friendly interface. Third, this approach has allowed us to tightly couple the simulation with specialized data visualization (GIS) tools, providing integration benefits that would have been hard to achieve using any of the off-the-shelf systems that were available at the time. (Since the software development was completed, simulation software has made great strides in allowing integration with database software and code segments written in other languages.)

We were very lucky in that several years’ worth of historical data was made available to us. We used this data by running trace-driven simulations: the calls that we simulate are real calls that are read in from a stored file. See p. 133 of Bratley, Fox and Schrage [22] for a discussion of issues relating to the direct reuse of historical data from the general perspective of discrete-event simulation. We confine our remarks to specifics related to the ambulance-planning problem.

The data used from each call are call arrival time, call priority, call location, time spent by an ambulance at the scene, destination to which the patient was transported (if any) and time spent at the destination. The use of this historical data obviates the need to develop a statistical model for generating calls. This is a decisive advantage, as the correlation structure of calls, both temporally and spatially, is rather complex; see, for example, Lubicz and Mielczarek [10]. For example, the location of a call is somewhat correlated with the time of day at which it is received.

Of course, if we were to use BARTSIM for long-range planning (say more than two years into the future), we might be more wary about using historical data, because the existing data may not be representative of conditions in the future. In such a case, one might want to use an approach similar to that used in the development of the United Network for Organ Sharing Liver Allocation Model [23]. That model uses non-homogeneous Poisson processes to generate "arrival times"; other information about the "arrival" is obtained through a bootstrapping procedure.

An area of concern that arises in using historical data in this fashion is data validity. Indeed, many of the logged calls contain entries that are difficult to believe. For example, it is not uncommon to see durations of 1 second for the time spent at the scene of an incident. Discussions with ambulance personnel revealed that this can occur when personnel forget to notify the CAD system (through a button situated on the dashboard of an ambulance) that they have arrived at the scene. When they realize their error, they "catch up" by pushing the button multiple times. This sort of error not only corrupts the recorded time spent at the scene, but also any surrounding times, such as travel times, that are used elsewhere. Identifying such errors and devising methods for dealing with them are important research areas that we have not explored. Instead, we adopted an ad-hoc procedure where the data for a particular call is "cleaned" if it is "close" to being reasonable, or the call is deleted if the logged data is beyond repair. Of course, if too many calls require cleaning or deletion then we should be concerned, and this is the reason why more research is required in this area. Fortunately, in the St. Johns application such calls appear to occupy a very small percentage of the total calls processed, so they cannot greatly sway the overall results.

The use of trace-driven simulation allows one to deal effectively with many other issues, such as that of multiple-response calls. Multiple response calls occur, for example, because the personnel who initially respond are not legally qualified to administer needed drugs, or because the number of injured parties is large. Each response to a multiple response call is logged in the St. Johns CAD database and linked to previous entries. Within our simulation we simply replay these calls. This very simple approach could lead to potential errors when the personnel that initially respond in the simulation are qualified to assist the patient, so that further ambulance responses are not necessary. A more sophisticated simulation approach might avoid such errors by carefully analyzing the data record, but we did not do this. In any case the number of such multiple response calls is quite small.

Ambulance availability is specified in terms of when and where an ambulance is to be brought into operation, and when it is to be removed from circulation. This allows shifts to be effectively captured, along with (for example) meal breaks that must be held at the ambulance's base and have a certain minimum duration.

A vital component of the simulation is a travel time model that computes travel times between any pair of locations in Auckland at any time. An important step in this project has been to establish collaborative links between St. Johns and the Auckland Regional Council, a local government body actively involved in developing strategic policy for the city of Auckland. The Auckland Regional Council made available a road network model that details both road layout information and travel times along roads (arcs) at various times of the day, including the morning and evening rush periods. The use of this data in BARTSIM is discussed in more detail in the next section.

It is possible to run the simulation and see ambulance operations unfolding on the screen. In particular, one sees ambulances traveling along the road network to and from calls. As calls arrive, they are plotted on the screen in a color indicating their priority. As calls are assigned to an ambulance, the calls change color, indicating that they are being served. This animation is extremely useful for verification and validation purposes, and for visualizing St. Johns' operations. It is also tremendously helpful in getting St. Johns personnel to accept the simulation model as a reasonable reflection of reality, and has proven invaluable in communicating our work to staff and management throughout the organization. This aspect of the simulation may seem somewhat trivial from a theoretical point of view, but has been absolutely critical in obtaining "buy in" from the decision makers. We view this selling point as a key advantage of simulation over other operations research methodologies for the ambulance-planning problem. The BARTSIM approach is intuitive and easy to understand for people with non-technical backgrounds.

When one wishes to collect performance measures, the animation is an unnecessary computational overhead. In this case, animation is turned off, and the simulation proceeds without graphical feedback. We do not report confidence intervals for our performance measures. This is mostly due to the fact that the theory of error estimation from trace-driven simulations is not well understood, so that it is not clear how to develop confidence intervals. This is an area where more research could certainly help.

A simulation model on the scale of BARTSIM requires a great deal of effort in verification and validation to ensure that the model that has been implemented is indeed what was desired, and that the model appropriately represents reality. Instead of entering into a full discussion of our efforts in this regard, which are mostly direct applications of the usual methods as outlined in Law and Kelton [24], we content ourselves with a few examples.

The animation facilities of BARTSIM proved invaluable in verifying the model. By watching simulated ambulance operations over extended periods, many errors in the database of real calls were identified. As well as replaying existing calls, BARTSIM also has a facility for interactively generating calls. This was used to place calls at strategic locations for checking that the ambulance responses were as expected. Shortest paths were generated and displayed over the road network to verify the quality of the chosen routes.

The validation of a model involves ensuring that the model appropriately represents reality. In this regard, we worked very closely with a number of individuals at St. Johns. These people were closely involved in the development phase, and also assisted in performing test runs. Furthermore, we demonstrated the software and described the simulation model to groups of ambulance drivers, who provided feedback on the quality of the model. These steps also helped in the accreditation of the model, where the model is accepted and trusted by decision makers. The decision makers were so closely involved in the development and testing of the model that they felt some form of "ownership" over the system.

4.4 THE TRAVEL TIME MODEL

Auckland is built around two large harbors between two coastlines, and is dotted with dormant volcano vents. Consequently it has a highly irregular topology. Any plausible simulation of road travel cannot rely on 'as the crow flies' routes, or simple modifications of these to take into account a moderate number of obstacles, but must incorporate knowledge of the road network including the effects of motorways and major highways. Furthermore, the model must also incorporate the often dramatic changes in travel times that arise from varying congestion levels across the day and the week.

We obtained road data from the Auckland Regional Council detailing a network with about 2,200 nodes and 5,000 directed arcs. This Auckland Regional Transport Model (ART) is a relatively detailed transport model developed for medium term (15-25 years) project and policy planning and evaluation of regional transport strategy [16]. Traffic volumes are determined in ART using equilibrium solutions driven by origin-destination trip demands. Because the trip demands are determined using an underlying demographic model, travel times can be predicted over any planning horizon for which population forecasts are available. This ability to perform long-term planning is most useful when evaluating strategic decisions such as the location of ambulance bases.

We denote the ART road network by $G = (V, A)$, where V is the set of nodes, and A is the set of directed arcs (i, j) from node $i \in V$ to $j \in V$. By entering trip demands for different times of the day, a range of equilibrium solutions can be found, each with different travel times for the arcs. The ARC data includes the 8 a.m. morning peak travel time t_{ij}^{8}, 12 p.m. midday travel time t_{ij}^{12}, and 5 p.m. evening peak travel time t_{ij}^{17} for each arc (i, j). Weighted combinations of these times are used to estimate the travel time t_{ij}^{h} during any other hour h of the day. The weights are chosen using regression models based on actual travel times available in the St. Johns database.

We could use this model to compute dynamic shortest paths for ambulances based on time-dependent travel times whenever the simulation requires such paths. However, this would be a time-consuming computation that would greatly slow down the simulation. As a reasonable approximation, we instead pre-compute and store a range of shortest paths as follows. Of the 2,200 nodes in the network, 1,435 are used to spatially locate bends in the roads, while 765 are 'decision nodes' that define points at which a driver has a choice of direction (ignoring U-turn options). More formally, a node j belongs to the set D of decision nodes, $j \in D$, if there exists both an arc $(i, j) \in A$ and two distinct arcs from j, $(j, k_1) \in \mathrm{A}$, $(j, k_2) \in A$ with $k_1 \neq j$ and $k_2 \neq j$.

For each pair of decision nodes $i \in D$ and $j \in D$, we pre-compute three shortest paths, P_{ij}^{8}, P_{ij}^{12} and P_{ij}^{17} using the morning peak, midday and evening peak travel times respectively. This decision-node path information is stored in memory.

During the simulation we need to find the shortest path $S \rightarrow F$ between any arbitrary start point S and arbitrary finish point F. The shortest path process we use is heuristic, but nevertheless appears to provide a good level of accuracy.

We note that S and F need not correspond to nodes in the network. The first step in our process is to determine the spatially closest non-motorway nodes, $s \in V$ and $f \in V$, to S and F, respectively. We next determine the sets of decision nodes, $D(s) \subseteq D$ and $D(f) \subseteq D$, that are 'immediately connected' to s and f. The set of decision nodes $D(s)$ is given by $D(s) = T_s \cap D$, where $T_s \subseteq G$ is a tree with root s and with branches each constructed by adding 'outward pointing' arcs until the first decision node is reached. More formally, T_s is initialized with root $T_s=\{s\}$, and then T_s is grown by iteratively adding each arc/node pair $\{(i,j),j\} : i\in T_s\backslash D, (i,j)\in A$. Similarly $D(f)$ is determined from $D(f)=T_f\cap D$, where T_f is a tree built at f by adding all 'inward pointing arcs', i.e., adding each arc/node pair

$$\{(i,j),j\}: j\in T_f\backslash D, (i,j)\in A.$$

We then consider all the paths given by

$$P=\{S\to s\to d_s \overset{h}{\to} d_f\to f\to F\text{: } d_s\in D(s),\ d_f\in D(f),\ h\in\{8, 12, 17\}\},$$

where $S\to s$ (and $f\to F$) denotes 'as the crow flies' travel from S to s (and f to F), $s\to d_s$ denotes the unique path from s to d_s in T_s,

$$d_s \overset{h}{\to} d_f$$

denotes the pre-computed shortest path from decision node d_s to decision node d_f at hour h, $h\in\{8, 12, 17\}$, and $d_f\to f$ denotes the unique path from d_f to f in T_f. Each of these paths is then evaluated using the interpolated travel times for the hour in which the journey begins. The $S\to s$ and $f\to F$ travel is at some assumed off-network speed. The fastest of these paths is deemed the shortest path.

The decision node concept provides two primary benefits. First, without the use of this concept, we would need to solve an 'all shortest paths' problem on 2,200 nodes for each of the three sets of travel times. An 'all shortest paths' problem on n nodes can be solved using the Floyd-Warshall algorithm in $O(n^3)$ time (Papadimitriou and Steiglitz [25], p. 133). With the decision node concept, we solve an 'all shortest paths' problem on approximately one third (765) of the nodes, and therefore reduce the computational effort by a factor of $3^3 = 27$. We also reduce the memory required to store the shortest path solutions by a factor of $3^2=9$. Second, we consider several paths involving different combinations of decision nodes when deciding which route to take between any origin and destination. This means that the chosen route is a compromise between a pre-solved single fixed route, and the true shortest path as would be determined by solving a dynamic shortest path problem while the simulation is running.

When an ambulance responds to a Priority 1 call, it travels at 'lights and sirens' speed. We have captured this effect within the simulation using a multiplicative factor to decrease travel times from more standard travel speeds. This factor was fitted to data available in the database. We are currently exploring other improvements to the modeling of travel speeds.

4.5 BARTSIM

BARTSIM consists of the simulation program, the travel model, and various analysis tools. The simulation and travel models have been outlined in previous sections. This section describes the analysis capabilities of BARTSIM. These capabilities may be applied to historical data as recorded by the St. Johns organization, as well as simulated data generated by the simulation component of BARTSIM. *Informed* comparisons can then be generated between alternative strategies for operating the ambulance service. These analysis capabilities have proven very useful in St. Johns' decision making, several instances of which are mentioned below.

To protect St. Johns' confidentiality, all figures presented in this section are based on simulated data, rather than actual historical data. Road travel times have been perturbed, and all performance figures subjected to random perturbation. The number of ambulances operating out of each base has also been modified, with the result that we see a lower level of performance and greater variability over the Auckland region in terms of response time than is actually the case with historical data.

We record the response time performance on every call, so that a call can be classified according to which performance targets have been met. These "micro-statistics" may be aggregated into response time performance within every suburb of Auckland, within every half hour of the week. When a run consists of multiple weeks of real data (the runs usually consist of several months of real data), then results in the same time period in different weeks are accumulated together. Statistics are also collected on ambulance utilization.

By recording the response time performance on every call, we can generate plots such as that given in Figure 4.2. In Figure 4.2 a black dot indicates that a call was answered within the 80% time requirement, a gray dot means that the call was answered within the 95% time requirement, and a white dot indicates that neither of these response time bounds was met. (These colors have been modified from those used in the software to improve reproduction.) One can visually identify localized areas of poor performance. This is a very powerful capability that St. Johns have found extremely useful in allowing management to visually interpret data that was previously only available in aggregated database report tables. In particular, using these plots we were able to verify a belief held by some at the St. Johns organization that Silverdale (a suburb of Auckland) needed more resources, perhaps because of the strong recent growth in the region. A long-dormant station in Silverdale has since been reopened.

Figure 4.2 Response time performance in the Auckland region (data is illustrative only)

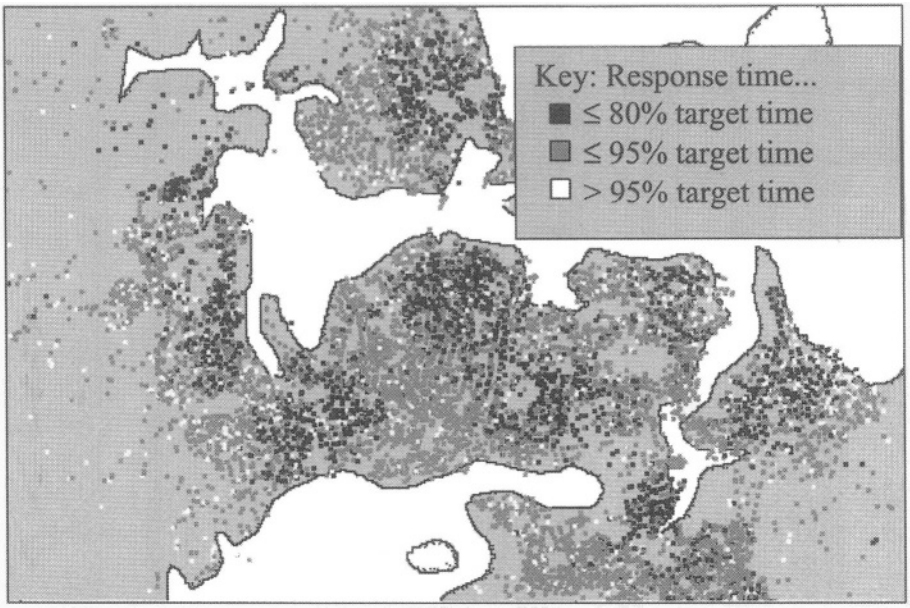
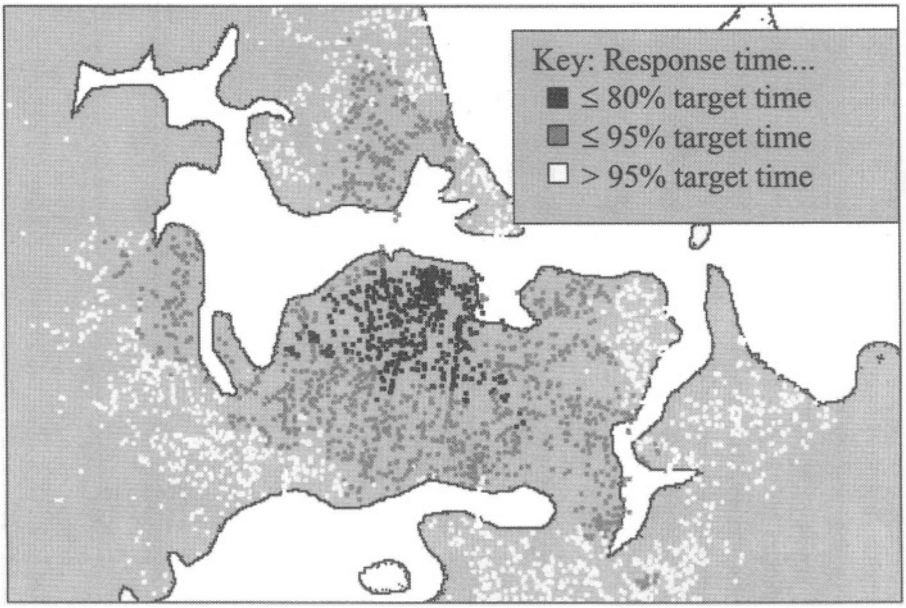


Figure 4.3 Plot of the “reach” of Pitt St. Station during the late morning/early afternoon period on weekdays (data is illustrative only)



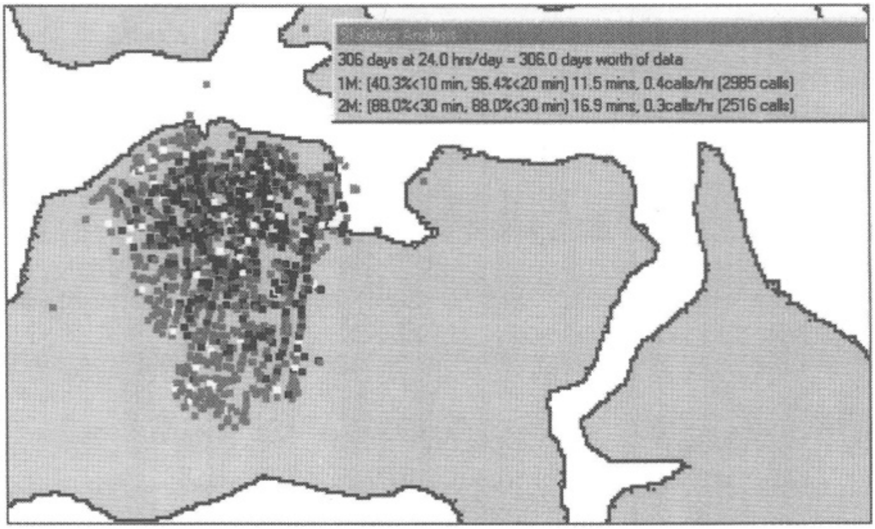
BARTSIM has proved to be a useful decision support tool for assisting with the allocation of ambulances to stations. During periods of low call demand, performance targets can be met by using just a few stations to cover the entire Auckland region. We can identify the "reach" of a station by producing plots like that of Figure 4.3.

In this plot, we computed the travel time from a single station to all calls. By coloring the call locations as above, we obtain a vivid picture of the area that can be covered by positioning an ambulance at a given station. Since travel time varies dramatically with the time of day, we can obtain a clearer picture of the station's reach at a given time by filtering the calls, so that we only display those arriving during a subset of the week. Figure 4.3 contains only those Priority 1 calls received in the late morning/early afternoon on weekdays. By repeating such plots for several stations, we can identify a suitable subset of stations that may be used to cover Auckland during various times.

As mentioned above, we can filter the calls so that one can "zoom in" on a particular time, or a particular area of Auckland, or both. The performance measures for the time and area of interest are then calculated, allowing one to identify response time performance for centrally located calls, for example. A sample screenshot of such an analysis is given in Figure 4.4. The small window in the upper screen area contains detailed information on contractual target performance for a case where ambulance allocation is too light, so that the targets are not met.

The plots described above are very useful for providing an overview of performance. In addition, plots such as those in Figure 4.4 allow one to provide precise numerical information on performance in a localised region. It is also desirable to be able to summarise on-time performance (relative to the contractual targets) over the entire Auckland region at once; Figure 4.5 is an example of such a plot. In this figure, the Auckland region has been broken down into rectangular regions. Within each region, we compute the percentage of Priority 1 calls reached within the required time limit (10 minutes for urban calls, 16 minutes for rural calls). To allow one to focus on regions containing significant numbers of calls, regions containing a small number of calls are suppressed in the output. Furthermore, the size (area) of the rectangles reflects the number of calls received within the region. We can also substitute other performance measures, such as the number of calls received, or the percentage of Priority 2 calls reached within the required time limit, in place of the performance measure used in this example.

Figure 4.4 Filter applied to results to identify performance in the city centre (data is illustrative only)



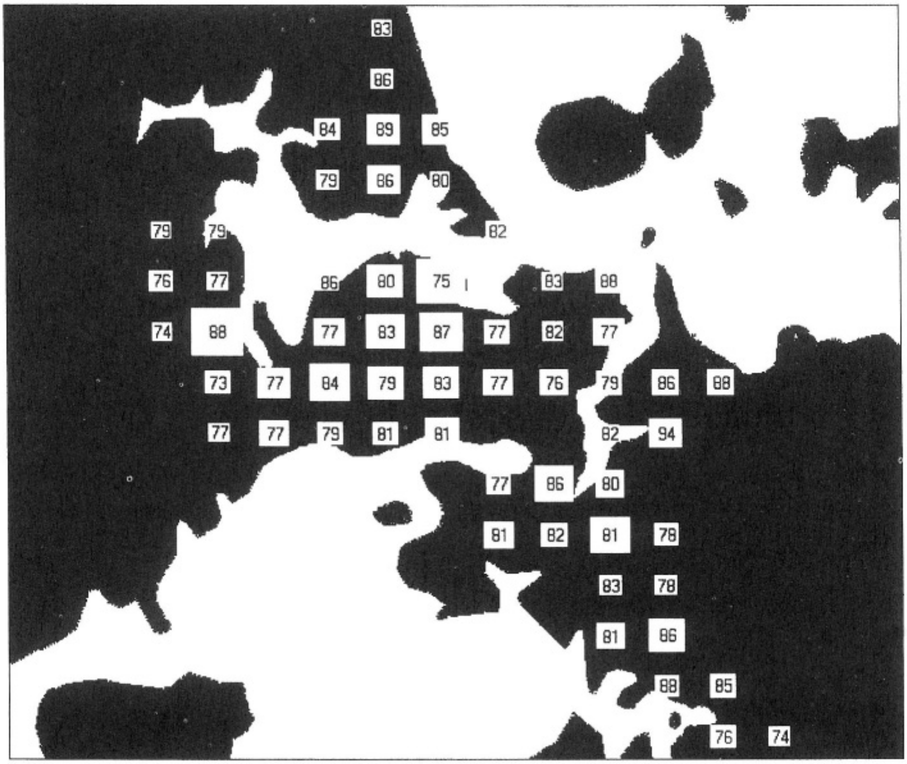
```
Statistics Analysis
306 days at 24.0 hrs/day = 306.0 days worth of data
1M: [40.3%<10 min, 96.4%<20 min] 11.5 mins, 0.4calls/hr [2985 calls]
2M: [88.0%<30 min, 88.0%<30 min] 16.9 mins, 0.3calls/hr [2516 calls]
```

Figure 4.5 is perhaps the most useful of all the plots described thus far in terms of determining required ambulance allocations. We vary the ambulance allocations between bases (usually heuristically, but one could also use optimisation methods), run the simulation, and then observe the performance in terms of these plots. Using these plots, we can locate areas with both a poor overall on-time performance and a large number of calls. These areas are good candidates for extra ambulance resources. Furthermore, by filtering the calls by time and producing the same plots, we can identify times when extra ambulances are most likely to have a large impact on the performance measures.

These plots revealed something unexpected when applied to historical data for the St. Johns organisation. In one small suburban area (not shown), a disproportionate (relative to neighbouring areas) number of calls were appearing. Upon investigation it was discovered that there are several accident and emergency clinics in this area, and such clinics generate many calls for St. Johns. The St. Johns organisation was apparently unaware of this situation, and is considering our recommendation that they ensure that an ambulance be relocated close to this vicinity.

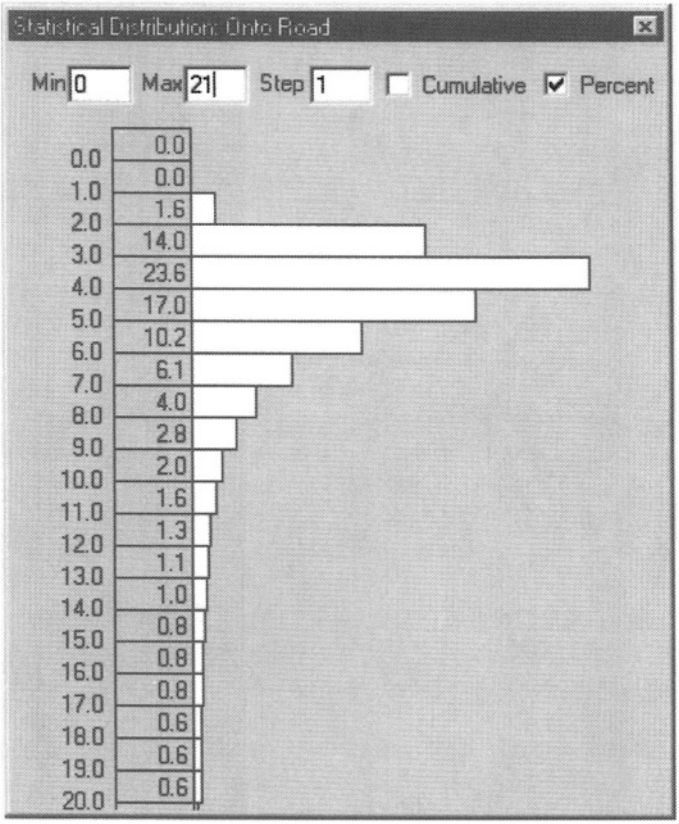
BARTSIM can also produce simple histograms of various characteristics of calls, such as response time, time spent by an ambulance at the scene, and so forth. One such histogram is given in Figure 4.6, showing the time between

Figure 4.5 Plot of average service quality (indicated by the numerical values) and the number of calls (indicated by the size of the white squares) for grid areas in Auckland (data is illustrative only)



a call being received and an ambulance being dispatched for a set of simulated metropolitan Priority 1 calls. The histogram shows very clearly that for many of the calls, a large amount of time is spent before an ambulance is dispatched to a call. Time spent in the dispatch process reduces the amount of time that an ambulance has to reach the scene of a callout if it is to meet the contractual performance targets. A plot similar to this for the historical data recorded by St. Johns was one of our most important findings for the organisation. Small decreases in these dispatch times can have (as simulations quantified) a large impact on contractual performance, so that it is worth devoting considerable effort to determining ways in which the dispatch time can be reduced. Apparent inefficiencies in the dispatch process can, when considered in view of the overall goals of the organization, actually be viewed as efficiencies, especially when the alternative expense of additional ambulance units is considered.

Figure 4.6 Distribution of the interval (in minutes) between a call being received and an ambulance responding (by radio) that it is en route (distribution is illustrative only)

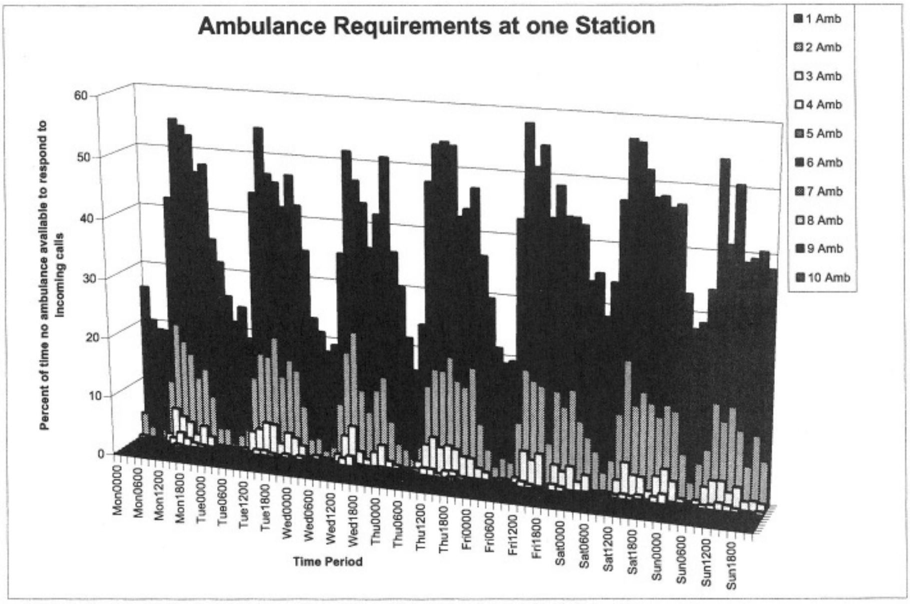


Statistical Distribution: Onto Road

Min 0 Max 21 Step 1 □ Cumulative ☑ Percent

BARTSIM also produces statistics on ambulance utilisation. These statistics may be imported into a spreadsheet (we use Microsoft Excel), and analysed from there. An example of the type of graphs that can be produced is given in Figure 4.7. This graph depicts the underlying demand near one of the stations operated by St. Johns. Each row of bars reflects the performance that can be expected over the week when a given number of ambulances are stationed at the base. In particular, each individual bar reflects, for a given number of ambulances and time of the week, the percentage of time that no ambulance is available to respond to incoming calls. This information is extremely useful for getting a first approximation to the number of ambulances required at each individual base at different times of the week. Of course, one would cover some proportion of these calls from other stations, but the plot gives an impression of the underlying demand.

Figure 4.7 Ambulance utilisation/requirements at one station (data is illustrative only)



As a final example of the nontraditional uses of BARTSIM, we mention that at a certain stage St. Johns was considering the use of a dispatching strategy that was expected to have a number of effects. First, it would better match the skills of the staff with the patient's requirements at the scene, thus resulting in better care. Second, it would result in fewer Priority 1 dispatches being made because the improved data collection would allow more cases to be classified as Priority 2. Priority 2 cases have a longer target response time so the performance targets for these cases would appear to be easier to meet. However, vehicles on Priority 2 dispatches do not use lights and sirens, so the time a vehicle spends on a case increases if it is changed from Priority 1 to Priority 2. The improved case classification would come at the cost of increased dispatch times. These changes were built into the simulation using approximations for the extent of the effects, and then comparisons between the current and proposed system were drawn based on the plots discussed in this section. The analysis played a large role in determining whether the proposed system would be adopted.

4.6 CONCLUSIONS

BARTSIM has been used to evaluate several decisions considered by St. Johns, including the use of a dedicated non-emergency patient transfer service, the possible introduction of a new dispatching method, and changes to where and when ambulances should be allocated. The results of these studies have been used to shape policy at St. Johns, and we continue to work with St. Johns on these and other issues, including rostering requirements for their staff. This experience has convinced us that simulation is a powerful tool in emergency service planning that is currently underutilized. Good simulation visualization tools have proven invaluable as a communication tool for describing our work to management and staff of St. Johns. The spatial data visualization capabilities have provided management with a significantly improved understanding of their current performance and, in conjunction with the simulation model, allowed results from what-if analyses to be readily communicated and understood.

It is important in vehicle simulation models to accurately capture travel time information. We have developed heuristics that allow both accurate modeling of travel times and rapid simulation run times. In addition, we introduced the notion of a decision node, which dramatically decreases the time required to compute shortest paths in the networks. This concept may be of interest in other applications where shortest paths must be calculated in large networks.

The travel times predicted by our model are deterministic: the same time is always predicted for travel from one point to another at a given time on a given day. However, travel times can vary tremendously depending on unpredictable events such as traffic congestion, weather, and traffic accidents. It is our belief, based on some initial analysis with very simple models, that randomness in travel times can have a material effect on the predictions of a model, and this is an area that we are beginning to investigate. Some care is needed, as it is not immediately clear how to generate random travel times. In general, there will be "macro" effects, such as those described above, which affect many ambulance trips in the same way, whereas other "micro" effects, such as traffic light phasing, might be confined to a single ambulance trip.

The combined simulation and data visualization tools introduced here have been of tremendous help to St. Johns, and several other ambulance companies have expressed interest in using the system within their organization. In our experience, the combination of CAD databases, CIS visualization methods and simulation leads to more informed decision making, and better utilization of resources, than the previous state of the art has supplied.

Since preparing this chapter, BARTSIM has been selected in a competitive tendering process for use in Melbourne, one of the larger cities in Australia. As part of this work, BARTSIM has evolved into a more powerful system known as SIREN (Simulation for Improving Response times in Emergency Networks) (see http://www.optimal-decision.com). Enhancements include call generation using non-homogeneous Poisson processes, introduction of stochastic travel times, more detailed case classifications, and more sophisticated simulation logic to handle the increased operational complexity of this new problem. For example, SIREN can dispatch several vehicles to a call, one of which is left at the scene while the ambulance officers travel in the other vehicle to the hospital. Upon leaving the hospital, this vehicle then travels back to the scene where the officers return to their original vehicles. The transport model has also been enhanced to reduce the memory requirements of the pre-computed shortest paths, allowing a network with 6,000 nodes and 14,000 arcs to be handled. This network also allows shortest distance (in addition to fastest time) routes to be calculated, and includes arc-specific times for lights and sirens travel. It is pleasing to see the value that SIREN can add being recognized by another ambulance organization.

Acknowledgments

The authors would like to thank the management and ambulance officers at St. Johns for their support of this project. Tony Francis's technical assistance has been most useful, while the contributions of Shaun Camp, Marion Etches, Pauline Mellor and Scott Osmond have all proven invaluable. We are also very grateful to Russell Jones, Don Houghton, Anatole Sergejew, and the rest of the ART team at the ARC for making their transport model available to us. We wish to acknowledge the outstanding contribution of Jenny Smith in her work on an initial queuing analysis of the St. Johns problem. We also acknowledge the huge contributions made by Optimal Decision Technologies, Ltd. (www.optimal-decision.com) and Alex Currell and his team at the Metropolitan Ambulance Service in Melbourne, Australia, in the development of SIREN, BARTSIM's successor. Finally, we are most grateful to Michael Clist for programming assistance in the early stages of BARTSIM's development. This work was partially supported by New Zealand Public Good Science Fund grant UOA803, and by National Science Foundation Grant DMI-0230528.

References

[1] Swersey, A.J. (1994). The deployment of police, fire, and emergency medical units. In Pollock, S.M., M.H. Rothkopf, and A. Barnett, eds., *Operations Research and the Public Sector.* North Holland, Amsterdam.

[2] Swoveland, C., D. Uyeno, I. Vertinsky, and R. Vickson (1973). Ambulance location: A probabilistic enumeration approach. *Management Science,* 20, 686- 698.

[3] Larson, R.C. and A.R. Odoni (1981). *Urban Operations Research.* Prentice-Hall, Englewood Cliffs, NJ. Also available at http://web.mit.edu/urban_or_book/www/book/

[4] Brandeau, M.L. and R.C, Larson (1986). Extending and applying the hypercube queueing model to deploy ambulances in Boston. In A. Swersey and E. Ignall, eds. *Delivery of Urban Services,* TIMS Studies in Management Sciences 22, Elsevier. 121-153.

[5] Savas, E.S. (1969). Simulation and cost-effectiveness analysis of New York's emergency ambulance service. *Management Science,* 15, B608-B627.

[6] Fitzsimmons, J.A. (1971). An emergency medical system simulation model. *Proceedings of the 1971 Winter Simulation Conference.* New York. 18-25.

[7] Fitzsimmons, J.A. (1973). A methodology for emergency ambulance deployment. *Management Science,* 19, 627-636.

[8] Fujiwara, O., T. Makjamroen, and K.K. Gupta (1987). Ambulance deployment analysis: A case study of Bangkok. *European Journal of Operational Research,* 31, 9-18.

[9] Daskin, M.S. (1983). A maximum expected coverage location model: Formulation, properties and heuristic solution. *Transportation Science,* 17, 48-70.

[10] Lubicz, M. and Z. Mielczarek (1987). Simulation modeling of emergency medical services. *European Journal of Operational Research,* 29, 178-185.

[11] Ingolfsson, A., E. Erkut, and S. Budge (2003). Simulation of single start station for Edmonton EMS. *Journal of the Operational Research Society,* 54, 736-746.

[12] Erkut, E., R. Fenske, S. Kabanuk, Q. Gardiner, and J. Davis (2001). Improving the emergency service delivery in St. Albert. *INFOR,* 39, 416-433.

[13] Harewood, S.I. (2002). Emergency ambulance deployment in Barbados: A multi-objective approach. *Journal of the Operational Research Society,* 53, 185-192.

[14] Ingolfsson, A., S. Budge, and E. Erkut (2003). Optimal ambulance location with random delays and travel times. Preprint. University of Alberta School of Business, Edmonton, Alberta, Canada.

[15] Brotcorne, L., G. Laporte, and F. Semet (2003). Ambulance location and relocation models. *European Journal of Operational Research,* 147, 451-463.

[16] Auckland Regional Transport (1994). Auckland Transport Models Project: Technical Working Paper 1 'Network Development And Inventory,' Environment Division, Auckland Regional Council, Auckland, New Zealand.

[17] Carson, Y.M. and R. Batta (1990). Locating an ambulance on the Amherst Campus of the State University of New York at Buffalo. *Interfaces,* 20, 43-49.

[18] Kolesar, P. (1975). A model for predicting average fire engine travel times. *Operations Research,* 23, 603-614.

[19] Kolesar, P., W. Walker and H. Hausner (1975). Determining the relation between fire engine travel times and travel distances in New York City. *Operations Research,* 23, 614-627.

[20] Peters, J. and G.B. Hall (1999). Assessment of ambulance response performance using a geographic information system. *Social Science and Medicine*, 49, 1551-1566.

[21] Pidd, M., F.N. de Silva, and R.W. Eglese (1996). A simulation model for emergency evacuation. *European Journal of Operational Research,* 90, 413-419.

[22] Bratley, P., B.L. Fox, and L.E. Schrage (1987). *A Guide to Simulation.* Springer, New York.

[23] Pritsker, A. (1998). Life and death decisions. *OR/MS Today,* August.

[24] Law, A.M. and W.D. Kelton. (2000). *Simulation Modeling and Analysis,* 3rd ed. McGraw-Hill, Boston, MA.

[25] Papadimitriou, C.H. and K. Steiglitz (1982). *Combinatorial Optimization: Algorithms and Complexity.* Prentice Hall, Englewood Cliffs, NJ.

5 SUPPLY CHAIN MANAGEMENT OF BLOOD BANKS

William P. Pierskalla

Anderson Graduate School of Management
University of California at Los Angeles
Los Angeles, CA 90095

SUMMARY

The chapter starts with a strategic overview of the blood banking supply chain. We then proceed to ask and answer questions concerning (i) the blood banking functions that should be performed and at what locations, (ii) which donor areas and transfusion services should be assigned to which community blood centers, (iii) how many community blood centers should be in a region, (iv) where they should be located and (v) how supply and demand should be coordinated. Then the many tactical operational issues involved in collecting blood, producing multiple products, setting and controlling inventory levels, allocating blood to hospitals, delivery to multiple sites, and making optimal decisions about issuing, crossmatching, and crossmatch releasing blood and blood products are presented. The chapter concludes with areas for future research.

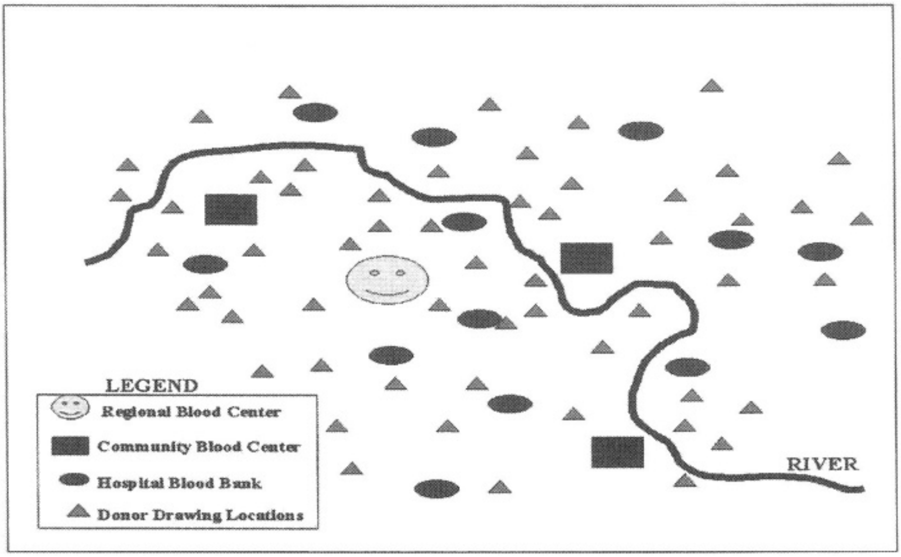
KEY WORDS

Supply chains, Blood inventories, Blood bank models

5.1 INTRODUCTION

As a supply chain, the flow of blood and blood products from the donor to the patient would seem to be one of the simplest inventory and distribution problems in the supply chain literature. Perhaps it is. One merely collects whole blood from donors, processes it into its components at a regional blood center or a community blood center and delivers the components to hospitals where they are transfused into patients. Geographically, the situation is shown in Figure 5.1.

Figure 5.1 A geographic region for blood supply and demand



In a geographic region, a regional blood center (RBC) with satellite community blood centers (CBCs) or, in smaller regions just the regional center without satellites, will be responsible for providing a supply of blood products (components) to hospitals for patients. To do this, a schedule of donor drawing locations is made some months in advance. Donors are solicited to give blood at the locations as the drawing time nears. Mobile phlebotomy vans with medical and service personnel and equipment are sent to the sites on the scheduled days. Decisions are made to prepare various components from the whole blood so the appropriate bags are used when drawing the blood. The drawn whole blood is returned to a processing location where it is recorded, tested for viruses and diseases, and the components are prepared. The resulting components are then inventoried and appropriate shipments are made to the hospitals based on their inventory needs. The hospital staffs then make decisions on how and when to use the blood components. If a particular blood component exceeds its allowable age it cannot be used for transfusion to a patient and must either be discarded or, for a few products, some modest salvage is possible. Some components, such as platelets, can be obtained directly from a donor by a process called pheresis. In this process a donor is connected to a machine that continuously circulates the donor's blood through the machine. The desired component is extracted from his/her blood and the remaining blood is returned to the donor. The process is costlier than the extraction of platelets from donated whole blood.

What makes this problem interesting and/or difficult from a research perspective? First, blood is a perishable commodity and whole blood has many components, each of which has a different shelf life before it perishes. The preparation of different components involves significant costs. Second, the supply of whole blood at a donor drawing location is a random variable that often has a large variance and, for planning purposes, the donor drawing locations and drawing dates are themselves sometimes random variables (Figure 5.1). The supply is also impacted by the need to screen out a growing list of viruses and diseases before the blood and its components may be used for transfusions; more variability and more risks are introduced. Third, the demands for blood components at a hospital in both their amounts and frequency are random variables (Figure 5.2). Fourth, many interacting decisions must be made at the strategic design, strategic policies, and operational and tactical levels. All are affected by the need to control costs, to minimize outdating and waste and, above all, to control potential shortages. Fifth, the entire blood supply chain can be examined as an essentially whole system and not just a subsystem of some larger system as occurs in most other supply chains. And finally, from a research perspective, much technically interesting, generalizable theoretical research can be extracted from the real problem regarding perishable inventories and regarding disease testing. In the future research section, other interesting unresolved theory questions will be raised.

Figure 5.2 shows the daily number of whole blood units drawn by the community blood centers in the Chicago area for one year. The drawing amounts range from zero to over 1,100 units and the variation is very large. It can be seen that in the January and November-December periods and in the summer the numbers of units drawn are below average and indeed there are often critically low inventories for patient needs.

O+ blood is one of the most common blood types. Figure 5.3 shows the range of daily demands for patient needs from a low of less than 10 units to a high of over 140 units and with significant daily variation throughout the year. This variability is typical for all blood types at large and small general hospitals that treat both emergent and elective admission patients.

Figure 5.2 Daily phlebotomy drawings by the CBCs in the Chicago area for one year

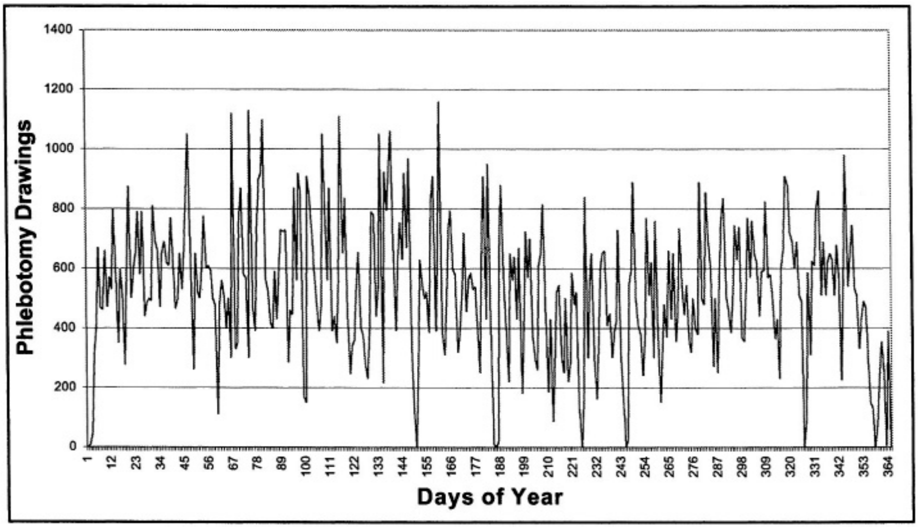
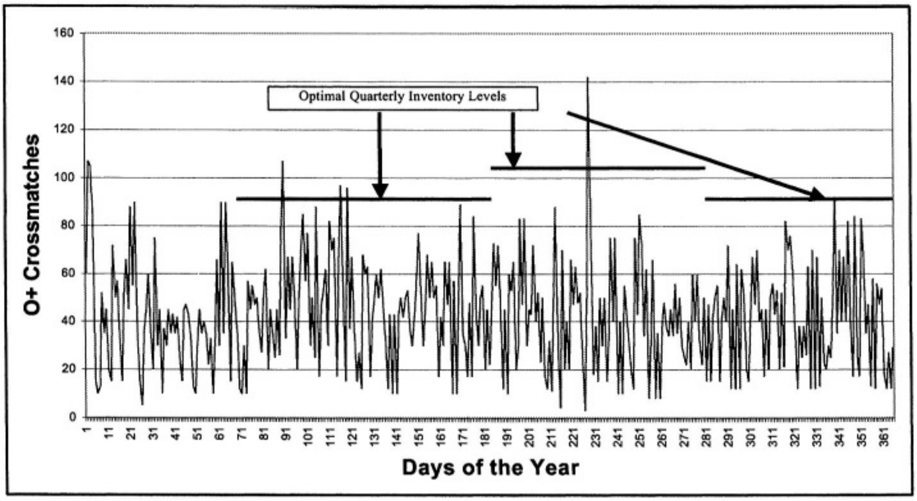


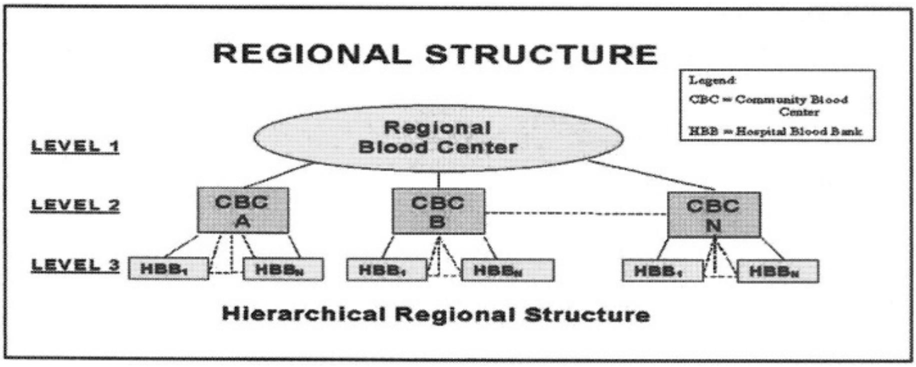
Figure 5.3 Daily O+ crossmatches for a large Chicago hospital for one year



The basic supply chain for whole blood and its components is given in Figures 5.4 – 5.7.

The organizational structure and the geographic region for a CBC or a RBC has usually evolved as the region's system of hospitals has grown and changed (Figure 5.4). In the early years as blood and components came into therapeutic use, hospitals began to draw blood and make components them-

Figure 5.4 Hierarchical regional structure



selves. Today many very large metropolitan hospitals still do so for a significant amount of their supply needs. But, in general, as the demands grew, hospitals found it to be more cost effective to seek a dependable central source for blood and components and also for the latest knowledge and research. As this growth occurred, CBCs evolved to meet the needs of groups of hospitals. In some regions only one CBC became dominant and met the needs of the region, whereas in other regions several CBCs successfully met the region's needs. In all cases, the intent of the hospitals was to obtain a dependable supply at minimal cost. For dependability this supply also had to be of the highest quality (free of blood borne diseases and meeting the best standards for therapeutic use) and always available when needed (no shortages). Because a significant part of the cost is recruiting donors and drawing, processing, storing, documenting and transporting blood, in order to minimize costs it was also necessary to minimize outdating and waste.

Depending upon the various levels of demand and the geographic location of a hospital, the CBC will make regular shipments of whole blood and components on a twice daily, daily, biweekly or weekly basis to the hospital. In a metropolitan area, most regular shipments would be on a daily basis. For outlying rural areas, the shipments may only be weekly.

In this process of inventory and distribution management the CBC must decide:

1. its own optimal inventory levels to maintain,

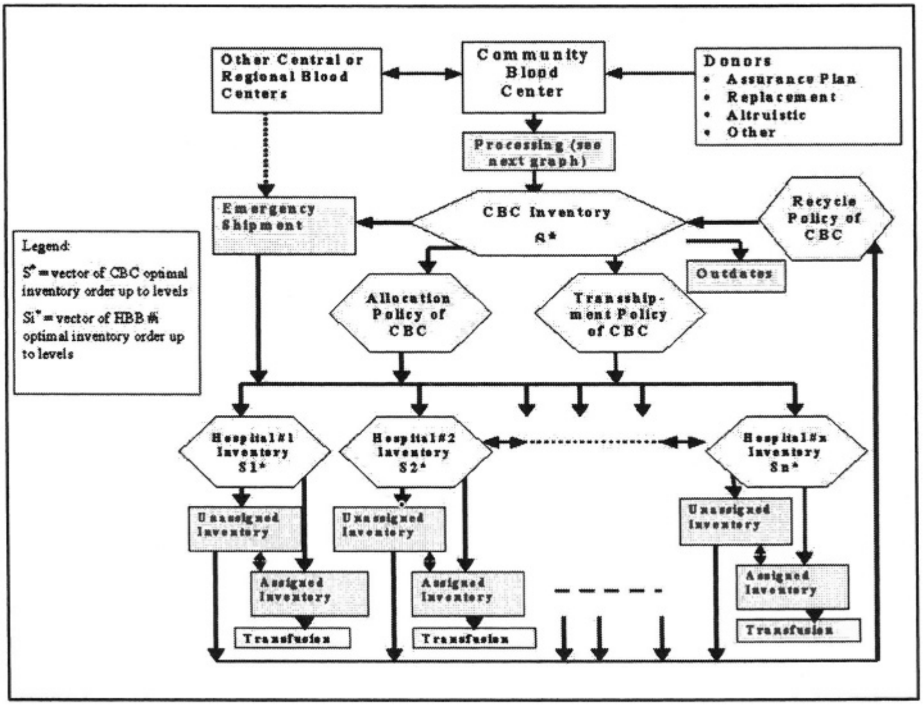
2. its inventory allocation policy in the event demands from the Hospital Blood Banks (HBBs) exceed the CBC inventories,

3. a trans-shipment policy from some HBBs to others in the event that there is an overall system shortage but some HBBs have a greater risk of shortages than others, and

4. a recycle policy of bringing old but still useful blood at an HBB back to the CBC for use at other HBBs with higher levels of demands and higher probability of using that blood before its expiration (Figure 5.5).

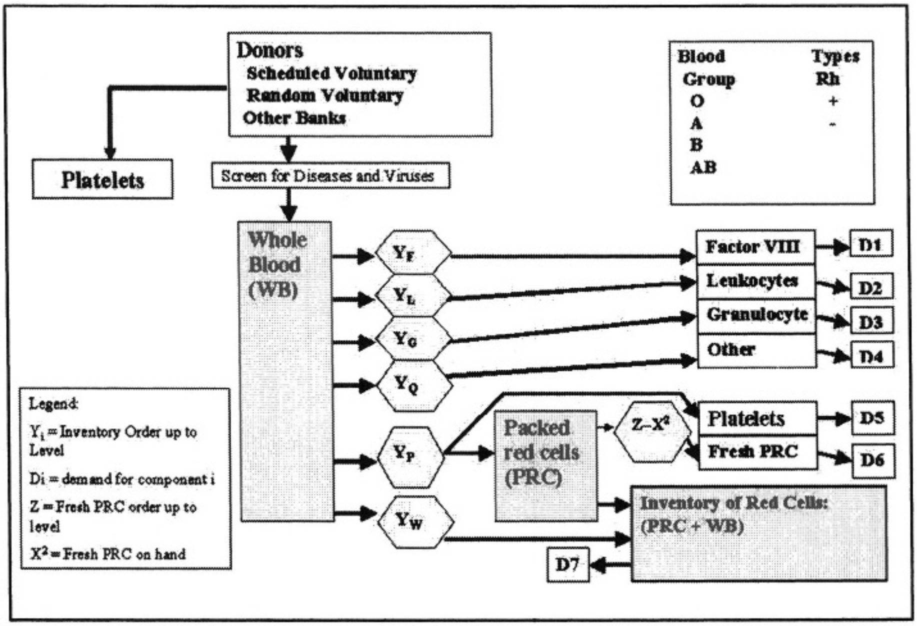
The HBB, itself, must decide its own optimal inventory levels to maintain. Depending on the corporate or contractual relationship between the HBB and the CBC, these levels may be made independently of or in conjunction with the CBC. More will be said about these optimal inventory levels later in the chapter.

Figure 5.5 The regional supply chain



In most cases, the demands for red cells and for the various blood components are independent random variables. However, since the red cells and components come from the same source – donors – there is a high level of dependence created on the supply side (Figure 5.6). Furthermore, the process of collecting the whole blood in appropriate types of bags and then making, storing and distributing the components can be costly, depending

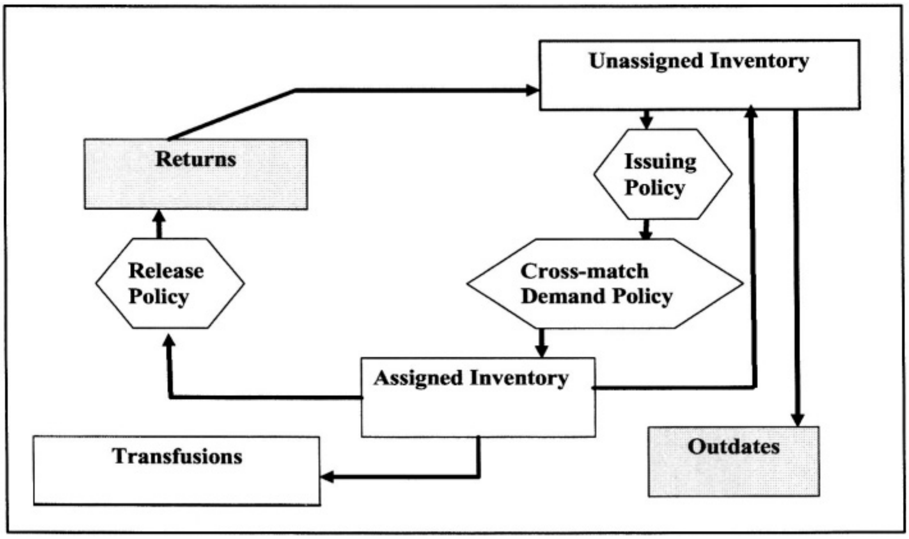
Figure 5.6 Processing whole blood into components



on the numbers and types of components made. Finally the components have different shelf lives, further complicating the supply chain processes.

In addition to its optimal inventory levels, the HBB must decide its issuing policy (usually last-in-first-out (LIFO) or first-in-first-out (FIFO)), its cross-match demand and its cross-match release policy (Figure 5.7). Cross-matching is the process of testing for incompatibilities between the patient's blood and the donated blood that the patient could potentially receive. The cross-match demand policy is the number of units of blood or a component that should be cross-matched to a patient's blood and assigned to that patient prior to its use. For whole blood and packed red cells (PRCs), the number of units will often be about one to two standard deviations above the average needed for the procedure. The cross-match release policy is the number of days after the patient's procedure that the unused blood or components will stay assigned to the patient in the event of emergency needs due to complications. For whole blood and PRCs this is often one or two days. Obviously the units continue to lose shelf life while on cross-match to that patient. Once released from this assignment, the units return to inventory (older) and can be cross-matched for use by another patient when needed. The reason for the assigned inventory is that its takes time to do the cross-matches and in an emergent situation the patient will not be able to wait.

Figure 5.7 Hospital i's inventory process and decisions



5.2 LITERATURE REVIEW

Research on the regional and the local management aspects of the blood supply essentially started in the 1960s, peaked in the late 1970s and early 1980s and then dropped off significantly to the present time. Excellent reviews of the work to the mid-1980s can be found in Prastacos [1], especially with regard to blood bank management policies and decisions, and in Nahmias [2] regarding theories of perishable inventories. In the years since these two reviews were published, almost every OR/MS researcher has left this area of research to pursue other interests. To some extent this exodus was caused by the collapse of federal funding for studies in the area (which reduces support for MS and PHD students), in the increasing difficulty of the remaining problems in the area, and in the shift of emphasis to do research on blood supply safety.

Since the mid-1980s, the published management-oriented work in blood banking has mostly been in the development of information systems to support donor screening, inventory management, blood ordering, blood usage review and compatibility testing [3]. Indeed, much prior work on information systems (IS) has been reported in earlier decades, but the advent of the personal computer (PC) and PC networks has driven the development of new structures and uses for blood information. In addition to improved IS, more new technologies are being introduced to improve the logistics and safety of the blood supply and delivery [4]. The National Blood Service, which is the central blood service for the United Kingdom (in a sense a super

RBC), is considering introducing electronic tags for every blood bag with sensors that tell all donor-specific details, blood type, and the exact location and temperature in real time for that bag. Clearly such a technology will greatly increase the safety and quality of the blood supply as well as provide logistics data for optimal supply chain management.

Few studies of note in the application of OR/MS have appeared in the past two decades. A platelet inventory management model was developed to determine outdate and shortage rates as a function of base stock levels and mean daily demand [5]. Using simulation, the model provided the base stock levels for different mean daily demand such that the platelet outdates and shortages in a region were significantly reduced. In another study, the task of scheduling donors at a bloodmobile site was undertaken [6, 7]. This modeling involved issues of donor motivation and psychology, layout of the collection facility and managing serial and parallel queues. Using a simulation model, the authors were able to improve the registration, screening and phlebotomy processes, which in turn improved donor satisfaction and reduced donor balking and reneging in future blood drives. The employers at the sites that the bloodmobile visited were also better satisfied because the new layouts and scheduling reduced employee waiting times to donate.

5.3 THE REGIONAL BLOOD BANKING SYSTEM

5.3.1 Regional structures and economies of scale

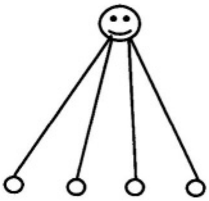
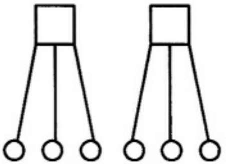
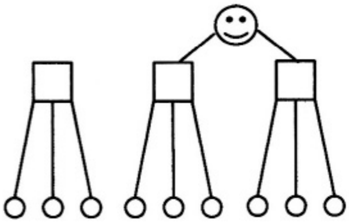
A strategic question regarding the regional supply chain for blood and components is: what are the economic and organizational consequences of different forms of regionalization of blood banking services?

If regionalization is to be effective, it must make a positive contribution to the achievement of one or more of the following objectives: reducing costs, reducing shortages and outdates, reducing extra-regional dependencies, improving the quality of the products, and reducing the confusion of overlapping jurisdictions. A search for economies of scale was thought to be the most logical starting point to analyze these factors. It is already known that by well planned operations, regionalization can reduce shortages and outdates by smoothing the region-wide supply and demand fluctuations (law of large numbers); however, issues of improved cost only will occur if there are economies of scale in regional operations.

The regional structures of interest are embedded in Figure 5.4 and illustrated in Figure 5.8. Level 1 is the Regional Blood Center, Level 2 is the Community Blood Centers and Level 3 is the Hospital Blood Banks (HBBs) and other Transfusion Service (TS) locations such as clinics and surgi-centers. (For ease of writing we will consider TSs as little HBBs and not use the TS notation.) In a region, if Level 1 does not exist, it means all HBBs are served by two or more CBCs only. In a region, if Level 2 does not exist, then a single RBC serves all HBBs (effectively operating as the sole CBC).

Figure 5.8 illustrates the different regional structures of interest. These are the single community blood center for the entire region, a collection of independent CBCs for the region or a collection of CBCs controlled or coordinated by an RBC. As a general rule, as the size of a region changes due to an increase in demand or geographic reach, the single community blood center may not adequately fill the needs of the region, and one of the other two structures will tend to replace it over time [8, 9].

Figure 5.8 Different organizational structures for a region

Single Community Blood Center for the Entire Region (Option 6)	**A Collection of Independent CBCs for a Region (Option 1)**	**A Collection of CBCs Controlled or Coordinated by an RBC (Options 2-5)**
		

In order to identify the economic and organizational consequences of different forms of regionalization of blood banking services, data were gathered from seven Chicago community blood centers and 66 Chicago area hospitals, as well as five other regional blood centers from around the nation.

Because wage rates, depreciation, purchasing costs of goods and supplies, rent, utilities and other costs vary greatly from one region to another, a proxy for costs was used. Instead of dollar costs for the geographic regions, man-hours per unit were used to derive the production function for each functional area of blood banking and for combinations of functional areas. The functional areas of main interest are:

(i) donor services (recruitment of donors and donor organizations),

(ii) phlebotomy on mobile units (collection and transport of the whole blood from donor locations),

(iii) phlebotomy at the community center,

(iv) processing (testing, typing and component preparation),

(v) inventory and distribution (storing and transport to hospitals), and

(vi) administration.

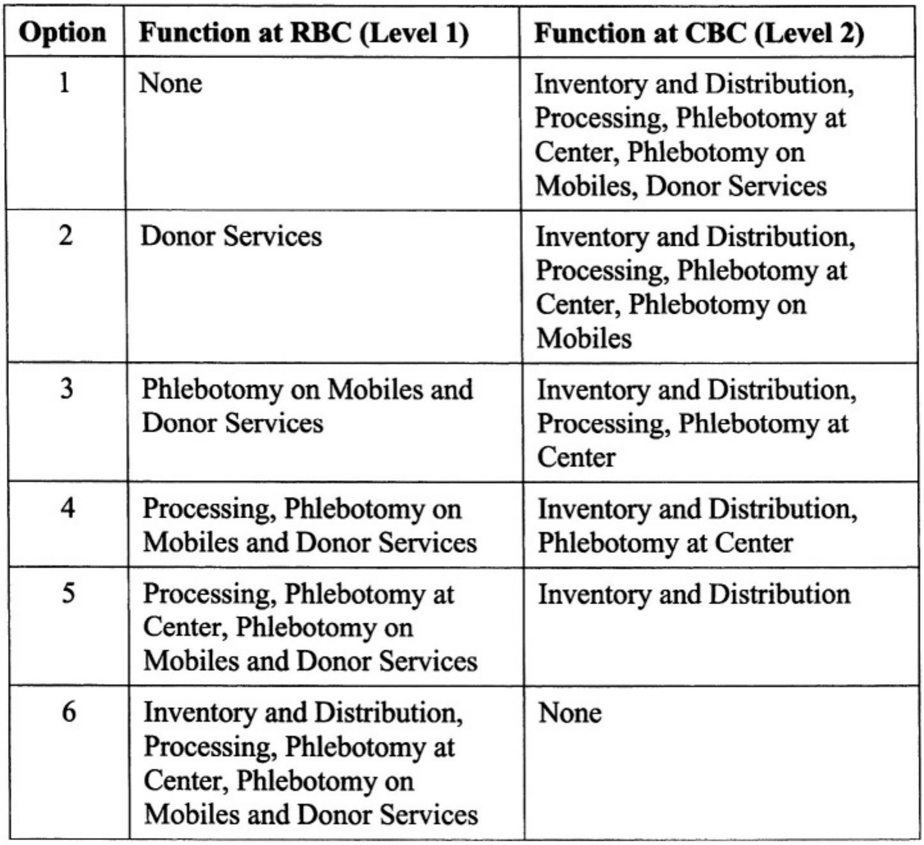
The overall total costs were also analyzed for scale economies.

The choice of man-hours removes the need to adjust dollar costs for the different wage rates experienced throughout the country. Great variation in man-hours per unit occurs across centers. Some of this variation is due to economies of scale or may result from different geographic distances covered, different proportional amounts of components produced, saturation of the donor market, style of management and expansion dislocations.

To reduce some of the data variations in the workload at the different centers for collecting, processing and inventory and distribution activities due to different proportions of components, a study of the times required to make the different components was undertaken. Using the results of these time studies, time-weighted volumes of activity were defined for each blood bank function and each center. In this manner, it was possible to compare the workload activities at all the centers for each function. In mobile phlebotomy, the number of units used to measure the workload were the amounts of whole blood drawn on the mobiles. In donor recruiting, the units were the whole blood drawn at the blood center, satellites and mobiles. In processing, the units were the whole blood drawn, plus weighted handling and processing times for the other components based on a normalized weight of 1.0 for whole blood. In the inventory and distribution area, the units were the total units shipped including whole blood and components. In the administrative and total manpower areas, the units were the appropriate units for each of the functional areas weighted by the percentage of the staff in each of the areas.

Because of the nature of the supply chain processes, some or all of the functions can be performed at Level 1 (the RBC) or at Level 2 (the CBCs). However, the process flow determines the order in which the functions are performed. Consequently there are only six possibilities for deciding which

Table 5.1 Options for regional operations where the blood banking functions are performed in different combinations at the RBC and CBC levels



Option	Function at RBC (Level 1)	Function at CBC (Level 2)
1	None	Inventory and Distribution, Processing, Phlebotomy at Center, Phlebotomy on Mobiles, Donor Services
2	Donor Services	Inventory and Distribution, Processing, Phlebotomy at Center, Phlebotomy on Mobiles
3	Phlebotomy on Mobiles and Donor Services	Inventory and Distribution, Processing, Phlebotomy at Center
4	Processing, Phlebotomy on Mobiles and Donor Services	Inventory and Distribution, Phlebotomy at Center
5	Processing, Phlebotomy at Center, Phlebotomy on Mobiles and Donor Services	Inventory and Distribution
6	Inventory and Distribution, Processing, Phlebotomy at Center, Phlebotomy on Mobiles and Donor Services	None

functional combinations can be performed at which levels as we analyze what is the best organizational structure for regional blood banking.

The combinations of functional areas are the options designated 1 to 6 in Table 5.1. For example, Option 5 reflects all tasks except inventory and distribution to be performed at Level 1 (the RBC); Option 2 reflects all tasks except donor services to be performed at Level 2, and so on. Using this set of six options, it is possible to analyze the structures given in Figure 5.8.

It was hypothesized that economies of scale exist in all options, with the possible exception of donor services. In donor services, as the geographic area expands and the donor market reaches a saturation level, it was hypothesized that increasingly more donor recruiter hours are needed to obtain the additional units of blood.

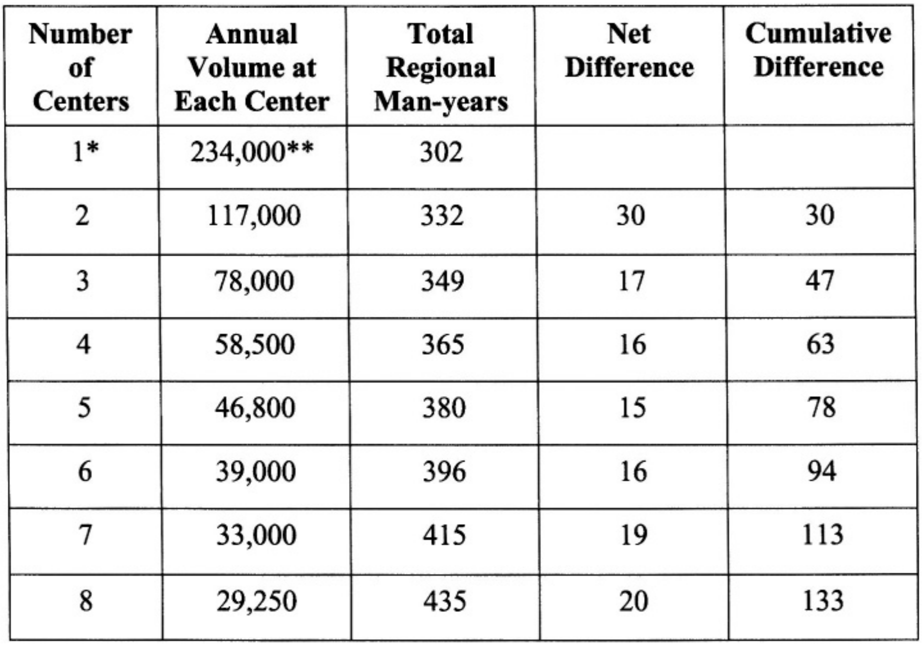
For the individual functions it was found that (i) the economies of scale hypothesis was significant for inventory/distribution and (ii) economies of scale occur initially, later followed by constant returns to scale in phlebotomy at the center, mobile phlebotomy, administration and processing. Donor services seemed to exhibit diseconomies of scale. When all functions are performed in a single center, economies of scale exist initially and are significant [10, 11].

For the options that correspond to specific regional organizational structures ranging from totally centralized activities to totally decentralized activities (Options 1 and 3-6) there are economies of scale. In particular, at the lower volumes (10,000 red cell units annually), the economies of scale are very significant. From 50,000-75,000 units, economies of scale are not as dramatic. Above 75,000 units the curves tend to flatten out but still show some small economies of scale. Caution should be exercised in using the curves past 200,000 weighted units since they were derived with only four data points. Option 2 exhibited economies of scale at the CBCs but diseconomies of scale at the RBC because in this option the RBC provides only donor services.

This analysis leads to two related conclusions. First, a regional system with community blood centers that are operating below 50,000 weighted units can realize significant economies of scale by increasing volume. These economies come from a more efficient utilization of space, equipment and vehicles, specialized skills and learning curve effects. Second, a regional system with one community blood center is more economical than a regional system with two CBCs, two are more economical than three, and so on. The example in Table 5.2 shows that none of these community blood centers should operate at less than 50,000 red cell units annually. Thus a region with slightly over 200,000 units annually is operated most economically with one center. The costs of two or three community blood centers (even if all are over 50,000 units) rise rapidly. This analysis leads to the conclusion that a region should have only one CBC (which would also be the RBC by definition). If a region needs more than one CBC due to geography and very large blood volumes, then the number of CBCs should be kept to a minimum.

Using the economies of scale results, we can gain an understanding of the cost implications of various regional structures as a basis for planned change in a region when such change is warranted. We can determine [12]:

Table 5.2 Example of the man-years needed in a regional system drawing 234,000 red cell units annually using Option 1



Number of Centers	Annual Volume at Each Center	Total Regional Man-years	Net Difference	Cumulative Difference
1*	234,000**	302		
2	117,000	332	30	30
3	78,000	349	17	47
4	58,500	365	16	63
5	46,800	380	15	78
6	39,000	396	16	94
7	33,000	415	19	113
8	29,250	435	20	133

* Option 1 when it has only one center is the same as option 6.

** In the year of the study, the seven Chicago community blood centers handled 234,000 units and used 416.5 man-years. However not all seven centers were of equal size as the example above has assumed.

(i) the blood banking functions that should be performed and at what locations,

(ii) which donor areas and transfusion services should be assigned to which community blood centers,

(iii) how many community blood centers should be in a region,

(iv) where they should be located, and

(v) how supply and demand should be coordinated.

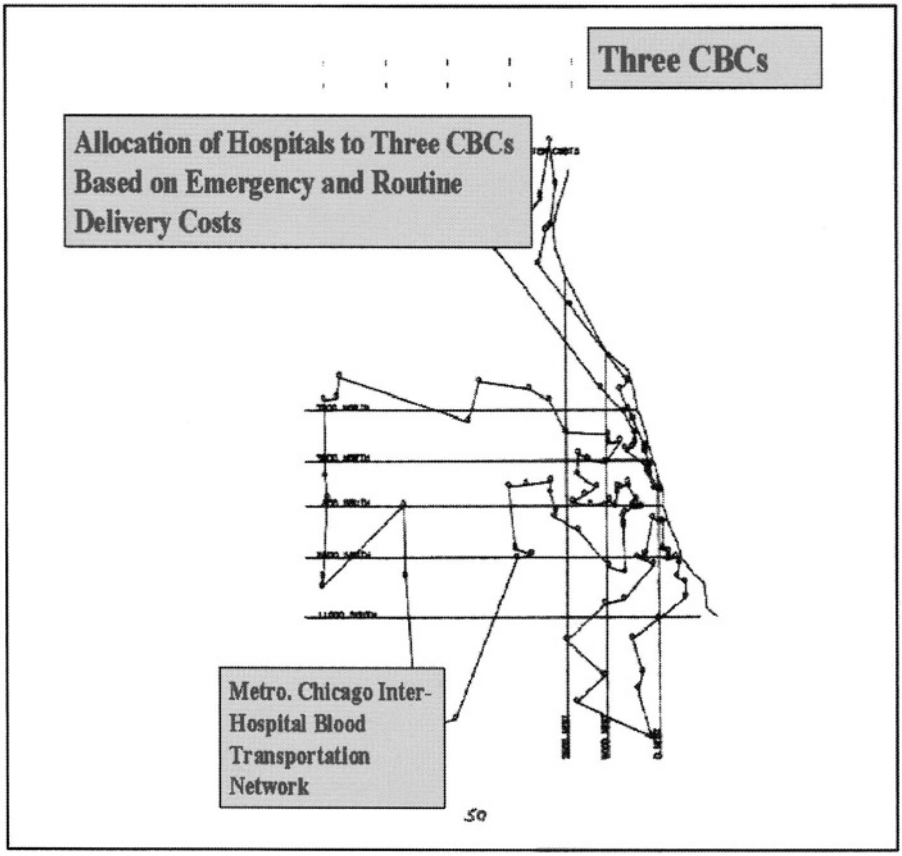
The next step in the analysis of the regional supply chain is to use the cost analysis to construct a model and decision support system to find: the number and location of community blood centers, the allocation of hospital blood banks to each CBC, and the routing of delivery vehicles from the CBCs to their HBBs to minimize (regular shipping costs + emergency shipping costs + operating costs) subject to constraints on: capital availability, personnel and facilities, budget, quality assurance, system reliability and demands for blood components. We call this model the Blood Transportation-Allocation Problem (BTAP) [13]. BTAP is a large constrained integer nonlinear program. BTAP is solved by decomposing the model into two sub-models: a demand model and a supply model. The demand model finds the best locations of the CBCs, allocation of the HBBs to the CBCs and the routing of the delivery vehicles for the distribution of the blood components to minimize the total costs of routine and emergency deliveries plus the system costs of operations subject to constraints. The supply model finds the best allocation of the donor supply locations to the CBCs to minimize the supply-side transportation and recruiting costs subject to constraints. This supply model is a constrained transportation-type problem.

The demand model takes as inputs the locations of HBBs in a region, the distances separating them, and their whole blood and component needs. Distances can be in any metric but for the analysis, the driving time between locations was used. The user then specifies the number of community blood centers to be evaluated (from 1 to 10) and their desired locations. In addition, the user specifies the desired option from Table 5.1 to be evaluated. Locations and options are then varied to achieve the most practical optimal locations and allocations. The results of one run of the model are shown in Figure 5.9. In this figure, the loops indicate which HBBs are assigned to which community blood center to meet the demands for blood at minimal cost.

The supply model was developed to allocate the supplies of blood to each CBC. This model takes as inputs the allocation of transfusion services with their demands given in the previous model and the available supplies of blood in the metropolitan area by zip code areas and then assigns the supplies to the CBCs in such a way as to meet the demands in each center and minimize costs of collection. Figure 5.10 illustrates the results of the supply assignment model corresponding to Figure 5.9.

Delivery Vehicle Routing. As part of a regional blood bank design model, the problem of vehicle routing for blood product deliveries was considered. The basic problem involves selecting vehicle routes for each central blood bank subsystem that minimize overall transportation costs between the CBC and its member HBBs. For each configuration of blood banks, a "sweep" algorithm was used. The algorithm is a heuristic method that is incorporated into the overall regional blood bank location and central bank allocation model (BTAP). Figure 5.11 indicates a typical regional design solution for the metropolitan Chicago area that also contains optimal vehicle routes.

Figure 5.9 Allocation of hospitals to three CBCs based on emergency and routine delivery costs

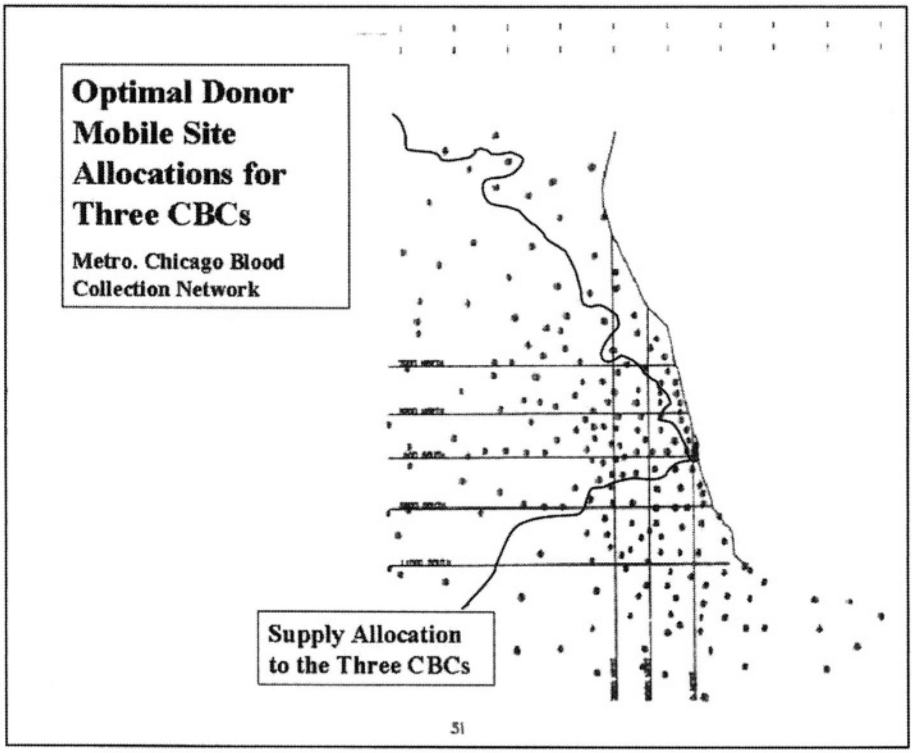

Three CBCs
Allocation of Hospitals to Three CBCs Based on Emergency and Routine Delivery Costs
Metro. Chicago Inter-Hospital Blood Transportation Network
50

In the preceding paragraphs it was concluded that some benefits of a regionally controlled structure would be:

- smoothing of the supply of blood from donors and a reduction in competition for donors,
- smoothing of the demands faced by a community blood center for blood and components by averaging the demands from many hospital blood banks, and
- economies of scale by operating community blood centers at levels above 50,000 units annually.

These benefits would lead to reduced shortages, outdating and costs.

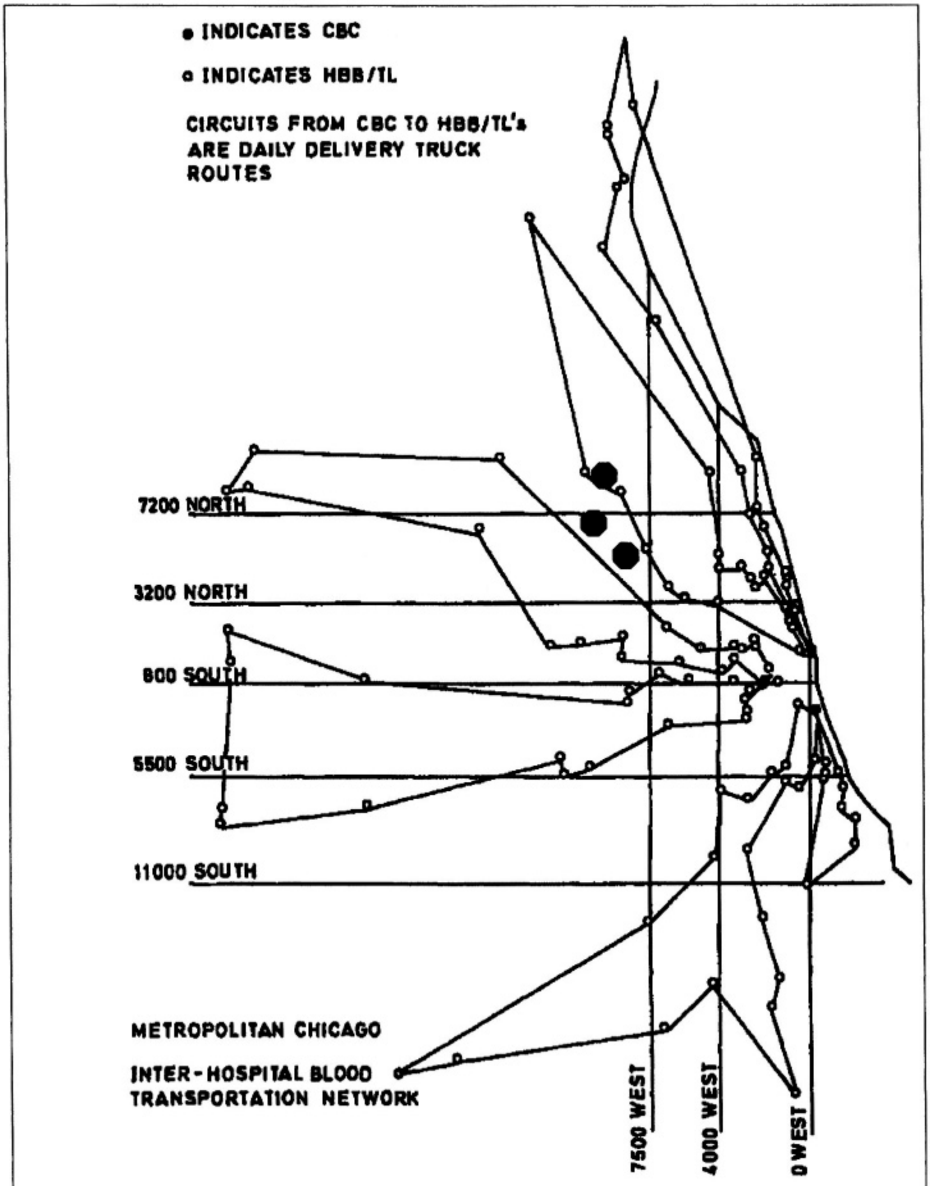
Figure 5.10 Optimal donor mobile site allocations for three CBCs


Optimal Donor Mobile Site Allocations for Three CBCs
Metro. Chicago Blood Collection Network
Supply Allocation to the Three CBCs
51

5.3.2 Tactical and operating decisions in a regional blood banking system

Before proceeding to detailed discussions of the tactical and operating decisions for a CBC and for HBBs, it should be noted that all of the optimal decision rules concerning inventory amounts, cross-match release policies, issuing policies, trans-shipment policies, vehicle routes and other factors interact with one another. That is, if one changes the policy in one area, it could affect the policies being followed in the other areas. These interactions will become more apparent as we proceed through this section and more will be said about them in subsequent discussions. Since it is not possible to present all of these policies simultaneously, each will be presented separately and the reader should keep in mind that they all interact. Usually for smaller changes, the interactions are not significant, i.e. the decision rules are robust.

Figure 5.11 Daily delivery truck routes



5.3.3 Forecasting

As in any business, the driving force for all decisions is the amount and types of demands that the business must meet to be successful. This fact is no different in blood bank management. Demands drive decisions. Any organization with random demand that does not do rational forecasting is condemned to work frequently in crisis and higher cost mode. Consequently it is necessary to have a good understanding of the demands, past, present and future. From our prior discussion, we know demand for the variety of blood products carried in inventory is a major source of uncertainty in the management of blood banks. Accurate forecasts of the quantity and timing of future demands become key inputs to inventory control and donor recruiting decision making. In particular, decisions relating to the quantities of blood products to be carried in stock, the scheduling of drawings from donor lists or mobile drawings, and ordering from other blood banks must all be made with such forecasts in mind.

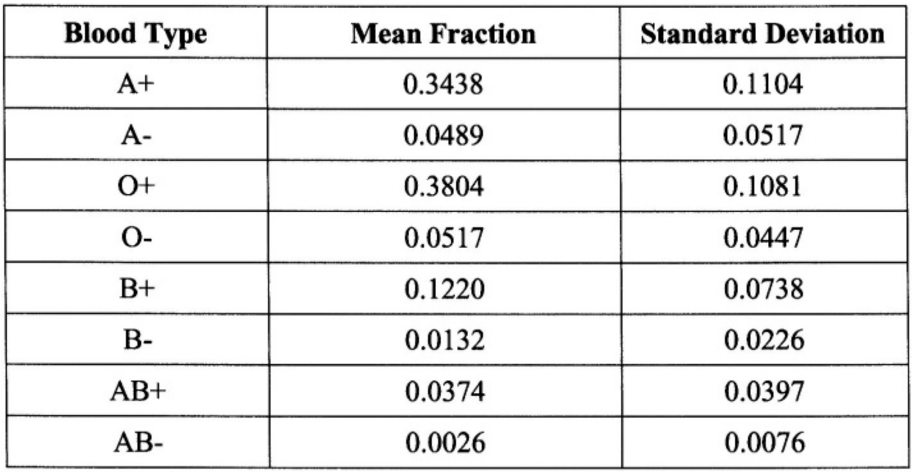
Demand for blood products can be computed by observing the number of those patients in a hospital who may require transfusions on any given day (cross-match requests) and the number of units requested for cross-match for each patient. Mean or average demand (cross-match quantity) then is simply the product of the mean number of requests times the mean number of units per request.

In order to specify the probability distribution for the number of units of a specific category or type, Yen [14], building on the work of Elston and Pickrel [15, 16], demonstrated that it is sufficient to estimate two parameters, the mean number of patients per day requiring transfusion (d_N) and the mean number of units requested for each patient (d_R). Moreover, the Neyman A distribution characterized by these two parameter values gave an adequate representation of the demand distribution obtained from data collected from a particular hospital. Subsequent analysis with regard to target blood bank inventory decisions [17-19], indicated that it was not necessary to keep track of these two components separately since effective system performance can be obtained by basing blood inventory decisions on mean demand alone (i.e., the product $d_N \times d_R$). Thus, in order to control the blood inventory effectively, forecasts of mean daily demand must be generated.

However, most blood inventory decisions are not reevaluated on a daily basis. In particular, target inventory levels probably would be updated on a monthly or quarterly basis taking seasonality into consideration. Figure 5.3 illustrates such a target levels, computed on a quarterly basis.

In order to forecast monthly demand and to identify seasonal cycles, it is necessary to collect data for several years. Often in HBBs only aggregate data (summed over all blood types) are available for such an extended period. For many planning purposes such aggregate forecasts are sufficient. In those cases where forecasts specific by blood type are needed, a reasonable approach would be to forecast demand levels on the basis of the aggregate data and then use estimates of the distribution of demand over blood types as a means of disaggregating these estimates into blood type specific forecasts. We tested the validity of this approach by examining the standard deviation of blood type fractions over one year of observations (Table 5.2). These standard deviations were observed to be relatively small when compared to the mean for the more common blood types. Moreover, the demand fraction for these blood types also was symmetrically distributed about its mean value. The rare blood type fractions exhibited significant variation relative to their mean and their distribution tended to be skewed. Since the rare types do not influence the aggregate blood demand significantly, we may conclude that forecasting of aggregate demand and subsequent disaggregation is a reasonable approach to generating longer term blood type specific forecasts for the common blood types. Further analysis, however, is needed for the rare blood types. (Cohen et al. [20] show that equation (1) can be used with reasonable accuracy for all blood types.)

Table 5.2 Transfusion requests as fraction of total demand (by blood type)



Blood Type	Mean Fraction	Standard Deviation
A+	0.3438	0.1104
A-	0.0489	0.0517
O+	0.3804	0.1081
O-	0.0517	0.0447
B+	0.1220	0.0738
B-	0.0132	0.0226
AB+	0.0374	0.0397
AB-	0.0026	0.0076

The best fit for the monthly demand series (12 years of monthly data from an HBB) using Box-Jenkins methodology is as follows:

$$D(t) = Z(t-1) - 0.812\, E(t-1) + 0.259\, E(t-12) - 0.211\, E(t-13) \tag{1}$$

where

$D(t)$ = the forecast for month t

$Z(t)$ = the month t actual transfusion request level

$E(t)$ = the forecast error $(Z(t)-D(t))$.

This forecast equation indicates that the moving average (error) term has cyclical components with periods of 1, 12 and 13 months. However, even for aggregate monthly figures there is significant variance and it is difficult to forecast monthly aggregate cross-match request quantities accurately at the single hospital blood bank level. We will later see that the optimal inventory order-up-to level is not very sensitive to the errors in prediction over a broad range around this optimal inventory level.

5.3.4 Target inventory levels for an HBB

As noted previously, the major responsibility of a hospital blood bank is to ensure that all blood-related demands are met in a manner that minimizes wastage through outdates and spoilage, maintains high quality standards and reduces shortages that require either emergency shipments from other blood banks, emergency demands on donors, appeals to the hospital staff for donations or the delay of nonemergency and elective medical procedures. In order to achieve these goals, it is important for the hospitals (HBBs) to set inventory levels that trade off shortage versus outdate rates and minimize total operating costs.

This section establishes a simple decision rule for an HBB which yields the optimal inventory level for each blood type for whole blood and red cells as a function of factors in the blood bank environment (the demand for blood by group and Rh, i.e. the blood type, and the ages of the blood units received from a CBC) and on the management decisions in the hospital itself (the inventory levels, the transfusion to crossmatch ratio, the crossmatch release period and the blood issuing policy – usually FIFO or LIFO). In using this simple decision rule, it is not necessary for the hospital blood bank administrator to choose a shortage rate for system operation since the inventory level recommended by the rule reflects the optimal tradeoff between shortages and outdates [17-19].

Demand data from hospitals in the metropolitan Chicago area were used. In order to understand the complex interactions among the environmental, managerial and random variables affecting the hospital blood bank, a series of models were constructed. The analysis began with a simulation model and a full factorial statistical design of the key variables to develop the response surface for various inventory levels' effects on shortages, outdates and costs. Then from the response surface a Cobb-Douglas model was used with log-linear regression to determine the optimal inventory order-up-to policy (optimal target inventory levels) for any whole blood/red cells Rh-blood type. This optimal inventory policy would apply to any hospital blood bank whose environmental and managerial data fell within the ranges of those variables used in the factorial design. These ranges were chosen to include most or all of the hospital blood banks in the United States. Using the optimal inventory policy, Cobb-Douglas models with log-linear regression were again developed to predict the resulting shortage and outdate levels under varying environmental and managerial decisions. These models used as input factors the system environment and hospital managerial decision variables. The factors considered include: parameters to specify the daily demand distribution, the age of units supplied from donors and/or the CBC, target inventory levels, the transfusion-to-cross-match ratio, cross-match release time, issuing policy, shortage cost, and outdate cost.

Model outputs include detailed records of all inventory transactions and the age distributions of both assigned (cross-matched) and unassigned inventories. These outputs are used to estimate the "optimal decision rule" i.e., the relationship between the cost-minimizing target inventory level, S*, and the various factors. In a similar way, the outdate rate, O_r, and shortage rate, S_r, were determined by relating them to the various factors as well as the target inventory rule.

The decision rule for the target inventory level is summarized in equation (2).

$$S^* = \frac{4.755(d_m)^{0.6964}(p)^{0.1146}(L)^{0,1332}}{R^{0.0453}} \tag{2}$$

where d_m is the mean daily demand for a blood type, p is the average transfusion to cross-match ratio, L is the maximum shelf life for red cells (either 35 or 42 days) and R is the cross-match release time in days. All coefficients are significant at the 0.01 level or less and $r^2 = 0.99$.

For each blood type, the blood bank manager computes the appropriate optimal inventory level, S*, and on a daily basis orders enough blood units to bring the available inventory on hand up to S*. If the blood bank receives deliveries only on a triweekly, biweekly or weekly basis then the d_m value that should be used in the calculation should be the mean demand over the number of days between deliveries.

A range of 2 to 50 units demanded per day (which corresponds to an annual volume of between 300 to 10,000 transfusions) was considered in the experimental design. Almost all blood banks have type-specific mean demand volumes that fall into these ranges, and hence equation (2) has wide applicability.

The small values of 0.1146 for the power of p, 0.1332 for the power of L and 0.0453 for the power of R in equation (2) indicate that their influence on S* is not nearly as large as that of d_m, with its power of 0.6964. Taken singly over the respective ranges of each variable, with the others held constant, the effect of p, L or R on S* is, at most, 6 percent to 8 percent.

For fixed values of p, L and R, a positive exponent of 0.6964 for mean daily demand in the optimal decision rule indicates that as the mean daily demand increases, there is less than a proportional increase in the optimal order quantity. Alternatively, a blood bank that doubles its activity (in terms of mean daily demand) should increase its optimal inventory level by no more than 62 percent (provided that p and R remain the same).

In a similar manner we can develop equations for the effects of the environmental factors and managerial decisions on the outdate rate and the shortage rate for a specific blood-Rh type at a hospital blood bank.

The outdate rate is the ratio of the mean number of units outdated to the mean number of units transfused plus units outdated, and the shortage rate is the fraction of days on which a shortage occurs. In establishing the relationship between the outdate rate and its causal variables, it was evident that two additional explanatory causal variables should be the deviation of the hospital's actual mean inventory level, S', from the optimal inventory level, S*, and the mean age of delivered units, A, from the CBC. If S' > S*, then outdates should increase because more blood is on hand than needed; if S' < S*, then outdates should decrease. In each case the reverse holds for the effect on shortages when S' differs from S*.

We also can hypothesize the effect of the other causal variables such as the crossmatch release time, R, and the mean age of units, A. As either increases, outdates should increase. The reverse should hold for the variables d_m, p and L. That is, the larger the mean demand, transfusion-to-cross-match ratio or the shelf life, the lower should be the outdates. The regression for O_r is given by

$$O_r = \frac{4.11052(R)^{0.66033}(A)^{1.57255}(e)^{0.00799(S'-S^*)}}{(d_m)^{0.8856}(p)^{2.54564}(L)^{3.01945} \qquad (3)$$

where *e* is the base of the natural logarithm.

From the regression results in equation (3), we can see that these expectations are true. All of the coefficients are significant at the 0.01 level or less and their algebraic signs agree with the above hypotheses. Furthermore, these variables explain 71 percent of the variation in the dependent variable $(r^2 = 0.71)$.

This regression function captures the effects of these six variables on the outdate rate. These same causal variables were used to explain the variation in the shortage rate, except that instead of the deviation (S' – S*), the reverse deviation (S* - S') was used. Consequently, if S' < S*, the shortage rate should increase because the actual inventory level S' is below the optimal level; and if S' > S*, the shortage rate should decrease. The other variables are expected to have the same effect on the shortage rate as they did on the outdate rate. As d_m, P, or L increase, the shortage rate should decrease. As R or A increase, the shortage rate should increase.

As shown in equation (4), these expectations have been realized. All of the coefficients are significant at the 0.01 level or less and are of the correct sign, The log/exponential linear regression explains 59 percent of the variation in the dependent variable. $(r^2 = 0.59)$. The regression equation is

$$S_r = \frac{0.09629\,(e)^{0.17356\,(S'-S^*)}\,(A)^{0.57441}\,(R)^{0.05359}}{(d_m)^{0.34867}\,(p)^{0.43568}\,(L)^{1.09577}} \tag{4}$$

where *e* is the base of the natural logarithms.

The variations in p and R represent examples of internal management policies since p and R are affected by the working relationships between the blood bank and the ordering physicians. Variations in A and $|S' - S^*|$ represent external management since the age and amount of arriving blood at the hospital often depend upon the policies of a regional blood center. Variations in L are set by government regulations (either 35 or 42 days for red cells) and are outside the scope of managerial decision.

The amounts of shortages, outdates and costs are determined by a complex interaction among these environmental and managerial factors. To capture the full effects of the benefits from following the optimal inventory policy, the other variables must not be allowed to deteriorate, i.e., p should not drop, R and A should not increase, and the actual inventory level S' should be held close to the target inventory level S* given by equation (2). Some of these variables are under the control of the blood bank administrator and others are possibly under the control of external administrators such as community blood center directors. Significant reductions in shortages and outdates, however, can be made by a combination of "good" overall internal and external management.

5.3.5 Optimal blood issuing for crossmatching (FIFO vs. LIFO)

FIFO (first in first out) and LIFO (last in first out) issuing policies were considered in conjunction with varying values of the cross-match release period, R, from 0 to seven days. R = 0 corresponds to an inventory system where all crossmatched units are transfused or immediately released after the procedure and R = 7 corresponds to a system where non-transfused crossmatched units remain in the assigned inventory for a period of one week. In all, 16 issuing-crossmatch policy combinations were considered (eight for FIFO and eight for LIFO). Simulation was again used to determine the optimal issuing policy.

Table 5.3 gives results averaged over a number of runs and Figure 5.12 is a graph of cumulated outdates for increasing values of R for both FIFO and LIFO issuing policies. The following observations can be made.

1. When R = 0, as predicted by the theory of Pierskalla and Roach [21], FIFO is optimal in terms of minimizing the outdates and shortages and costs (since costs increase as the number of shortages and/or outdates increase).

2. As ? increases under FIFO issuing, outdates increase and when R = 7 shortages appear.

3. As ? increases under LIFO issuing, outdates are very large but decreasing and shortages increase.

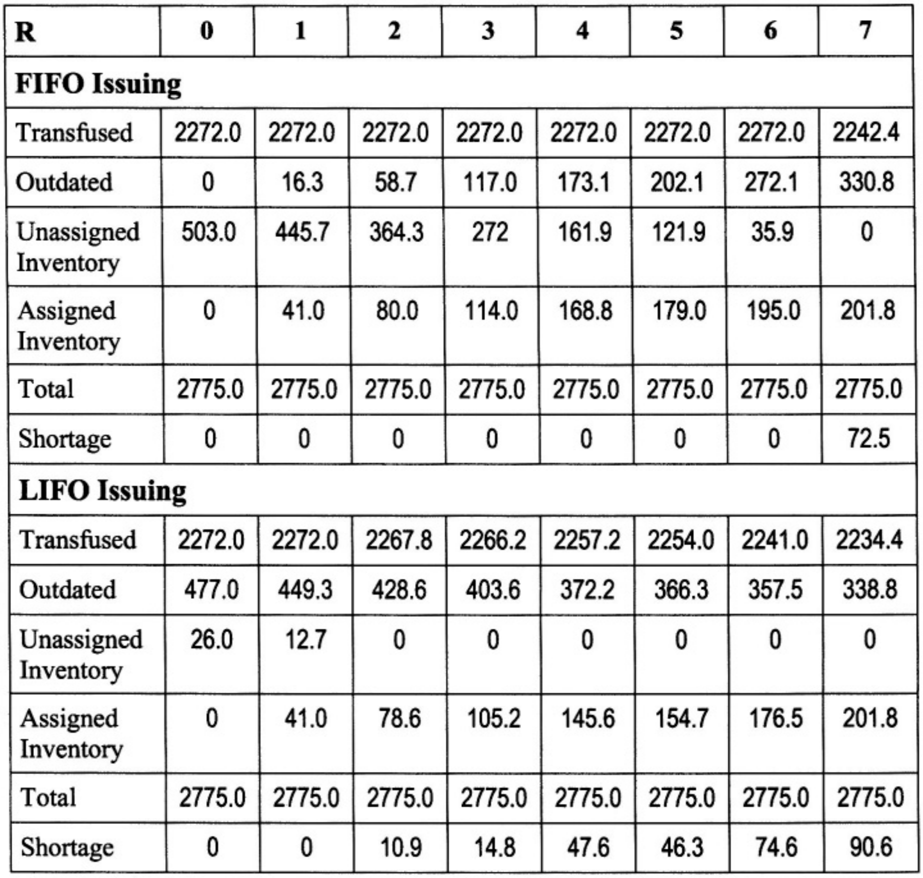
4. For reasonable R in the range 0-2 days (common in most hospital blood banks), FIFO dominates LIFO.

5. Although not shown, it is possible to generate examples where outdates under FIFO exceed those under LIFO for R sufficiently large (greater than seven days).

6. There is great sensitivity to changes in R under both issuing policies. The smaller the value of R, the higher is the system performance.

Choosing between the two policies, FIFO vs. LIFO, for any reasonable values of the crossmatch release period, FIFO is optimal and should be used by the hospital blood bank manager. Furthermore the hospital blood bank

Table 5.3 FIFO vs. LIFO issuing policy results on outdates and shortages for crossmatch release periods (R) ranging from zero to seven days



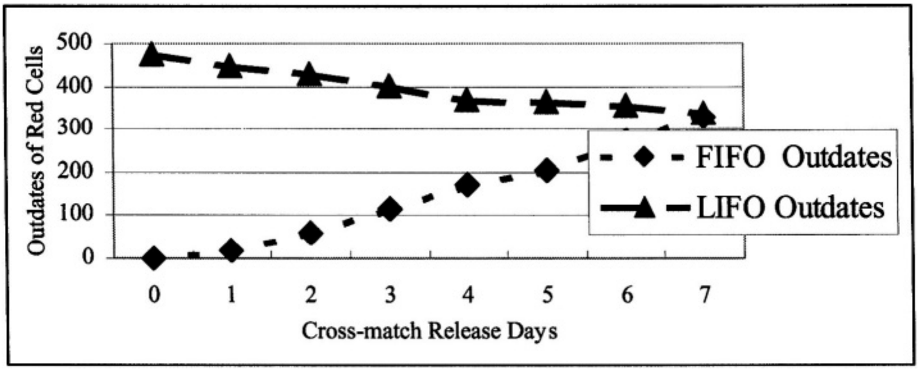
R	0	1	2	3	4	5	6	7
FIFO Issuing								
Transfused	2272.0	2272.0	2272.0	2272.0	2272.0	2272.0	2272.0	2242.4
Outdated	0	16.3	58.7	117.0	173.1	202.1	272.1	330.8
Unassigned Inventory	503.0	445.7	364.3	272	161.9	121.9	35.9	0
Assigned Inventory	0	41.0	80.0	114.0	168.8	179.0	195.0	201.8
Total	2775.0	2775.0	2775.0	2775.0	2775.0	2775.0	2775.0	2775.0
Shortage	0	0	0	0	0	0	0	72.5
LIFO Issuing								
Transfused	2272.0	2272.0	2267.8	2266.2	2257.2	2254.0	2241.0	2234.4
Outdated	477.0	449.3	428.6	403.6	372.2	366.3	357.5	338.8
Unassigned Inventory	26.0	12.7	0	0	0	0	0	0
Assigned Inventory	0	41.0	78.6	105.2	145.6	154.7	176.5	201.8
Total	2775.0	2775.0	2775.0	2775.0	2775.0	2775.0	2775.0	2775.0
Shortage	0	0	10.9	14.8	47.6	46.3	74.6	90.6

manager must endeavor to keep R as small as possible in order to minimize its effect of increasing outdates and shortages. This effect was also seen in the outdate and shortage functions, equations (3) and (4) above.

5.3.6 Target inventory levels for a community blood center system or a centralized regional blood banking system

In a community blood center or regional blood center, the management of inventories of whole blood and components also involves a complex and interrelated set of decisions concerning collection, processing, record keeping, storage, issuing and transportation of units. In this section, some management decision problems are analyzed to determine easily implement-

Figure 5.12 Outdates for varying crossmatch release periods for FIFO and LIFO issuing policies



ted rules that yield the "best," or at least "very good" operating results at the CBC and its satellite HBBs [10].

It has been recognized that benefits can be obtained by pooling resources using a community blood center. The most apparent benefit to the hospital is that the blood bank staff is relieved of the responsibility of donor recruitment, blood procurement and blood processing. This permits the hospital blood bank to channel its energies and efforts toward the resolution of patient-related transfusion problems. Another advantage to the hospital is the opportunity to pool widely fluctuating, largely unpredictable demands with those of other hospitals in the system. Within the system the variations often cancel each other and produce a smoother, more predictable aggregate demand. This will enable member blood banks to maintain lower inventories without degrading their outdate and shortage performance.

Since the demand to which the community blood center must respond is generated outside its control, its decision making processes must focus primarily on inventory management. While management decisions regarding donor recruitment, phlebotomy and processing are essential, they can be handled effectively only after efficient optimal inventory control policies have been implemented. This control at the community blood center requires setting inventory levels to maintain the optimal tradeoff between system-wide excess inventory, with consequent outdating, and system-wide excess amounts of shortages.

Inventory levels can be developed for the CBC by using an outdating/shortage cost-minimizing procedure similar to that described in the previous section for the single hospital blood bank. The optimal inventory level at the CBC for each blood type is a function of the number of HBBs served by the CBC, their mean daily transfusions, their transfusion to crossmatch ratios and their crossmatch release periods. It is assumed that all issuing is done using FIFO. Associated with the optimal inventory function for the CBC is an optimal inventory level for each member HBB that is a function of its demand and transfusion to crossmatch ratio for that location. For those hospital blood banks that belong to a centralized system it is to be expected that their optimal inventory levels will differ from the values indicated by the decision rule of the previous section which is appropriate for independent hospitals in a decentralized system.

In order to study some of the benefits and shortcomings from a centralized blood banking system, a simulation model was constructed. The simulation model is described in detail in Yen [14]. Among the issues to be discussed in this section are the optimal inventory levels at each HBB (denoted $\mathbf{S_i}$ for each hospital i), the impact on total system cost of high cross-match to transfusion ratios, the allocation of units from the CBC to the HBBs, the trans-shipment policy among HBBs and the effect of a limited and somewhat random supply to the community blood center. In addition, the sensitivity of system cost to changes in the number and size of HBBs in the system was considered.

As one might expect, the results indicate that the total amount of optimal inventory levels in the hospital blood banks increase at a decreasing rate with incorporation of more HBBs into the centralized system. Also, after a certain system scale is reached the marginal benefits received from lower shortages and lower outdates can be expected to approach zero as more HBBs are added to the system. Finally, as more HBBs are included in a centralized blood banking system, the total average distances between the CBC and the HBBs, as well as their information needs, increase and thus the corresponding transportation and information costs increase. So, as HBBs are added, a saturation number of hospital blood banks in the system is reached, further inclusion of local banks is not likely to reduce the system cost per unit and indeed as has been shown previously may lead to diseconomies of scale above 200,000 units annually.

5.3.7 Optimal daily inventories at the CBC

The daily amount of whole blood and components to be maintained centrally at the CBC depends upon the amounts maintained at each HBB in the system. If the total inventories at the HBBs are larger than would be optimal, then the amount at the CBC should be small and vice versa. However, large inventories at the HBBs could result in more outdates and/or outdate-anticipating trans-shipments. Similarly, small inventories might incur more emergency shipments and/or shortage-anticipating trans-shipments. Because-of these possibilities, there must be a balance between the inventory at the CBC and the inventories at the HBBs in the supply chain.

Equations for determining optimal inventories of whole blood and packed cells at the HBBs were given in the previous section. Using a similar simulation-optimization-regression approach and making the same reasonable assumptions concerning the system costs of shortages and outdates, the optimal inventory level at the CBC was established.

The outdate cost consists primarily of the average costs per unit of recruiting, processing, storing and transporting one unit. When a unit outdates, these costs are basically lost. Actually a more appropriate cost to charge for outdates would be the marginal per unit costs of these blood bank activities rather than average per unit costs. However, it is not easy to obtain actual marginal costs since the cost figures available are not sufficiently precise to define the appropriate marginal relationship. Furthermore, since the average cost includes many variable items such as bag costs, record keeping and hours of work, it is reasonably representative of the marginal cost.

The shortage cost at the CBC was based on the cost for processing and handling a unit on an emergency basis and for recruiting and/or transportation from another source on an emergency basis. Again, marginal costs per unit would be better but they were not available. The shortage cost at the HBBs was based on the average per unit cost of maintaining a buffer stock of frozen blood units either at the HBB or the CBC or shipping a unit by emergency shipment from another regional center. Finally, it should be recalled that what is important about these costs is not their absolute levels, but rather their relative magnitudes. Hence, if inflation should cause them to rise in the same relative proportions, the results still hold. Furthermore, the results hold even when the relative magnitudes are varied over reasonable ranges.

Many variables were considered in this inventory supply chain analysis to find the optimal target inventory levels at the CBC and its independent satellite HBBs. A complete list of these variables is shown below. However, and somewhat surprisingly, only three of these many variables are needed to make optimal decisions in this centrally controlled supply chain. The optimal target inventory level at the CBC needs only to know d_0 and N and the optimal target inventory levels at the satellite $HBB_j s$ need only d_j. This contrasts for the optimal target inventory level at the independent HBB that needs three variables as shown in equation (2).

Variables initially considered were the following:

d_0 = mean demand for type specific whole blood and packed red cells at the community blood center.

d_j = mean demand for whole blood (WB) and packed red cells (PRCs) at the hospital blood bank.

It is assumed that demand at the bank is a Neyman type A distributed random variable characterized by the mean number of patients per day and the mean units requested per patient. Yen [14] demonstrated that the Neyman A fits the data well. Other variables initially considered were:

R = the crossmatch release period, the time lapse before a unit is returned to the unassigned inventory if not transfused (in days)

S_O = inventory level at the CBC (in units of WB and PRCs)

S_j = inventory level at location j (in units of WB and PRCs)

N = number of HBBs in the system

p_j = probability of a cross-matched unit of WB or PRC being transfused at location j

O_0 = shortage at the CBC (in units of WB and PRC)

v_j = shortage at location j (in units of WB and PRC)

O_j = outdate at location j (in units of WB and PRC)

n = number of times a unit of WB or PRC is cross-matched in its lifetime

a = age of a unit of WB or PRC when it is cross-matched for the first time.

The variables d_0, d_j, S_0, S_j, v_0, v_j and O_j are computed for each group and Rh factor; rather than have two subscripts, one for location and the other for ABO and Rh, the second subscript has been suppressed for ease of writing the results. However, for low volume rare blood groups when the target levels and the demands are small, say, one or two units, it is better to maintain more stock at the CBC rather than incur excessive trans-shipping of units.

The optimal level target inventory level at the CBC is:

$$S_0^* = 3.14\,(d_0)^{.72}\,(N)^{.93} \tag{5}$$

We find $R^2 = 0.993$ and F = 3529. All coefficients are significant at level 0.001.

The corresponding optimal target inventory level for the HBBs which belong to and are fully coordinated by the CBC is:

$$S_j^* = 7.99\,(d_j)^{.78} \tag{6}$$

We find $R^2 = 0.995$ and F = 8675. The term d_j is significant at level 0.001.

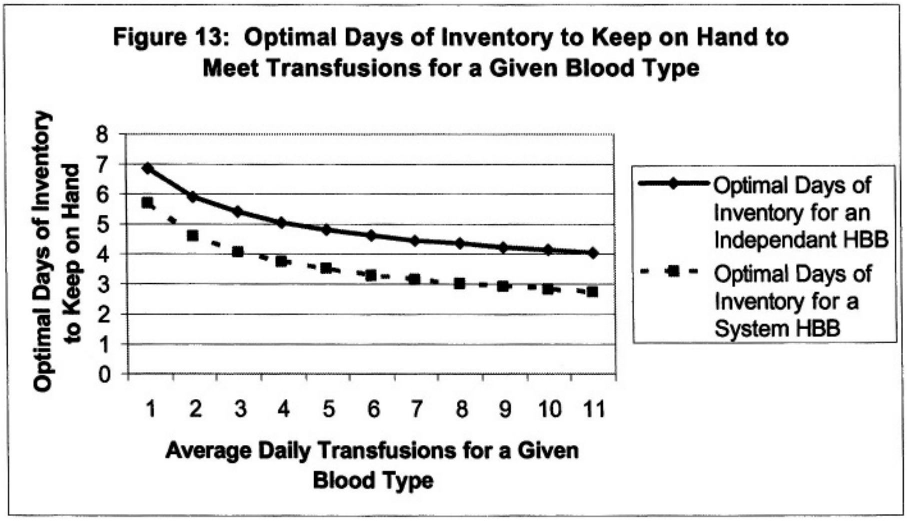
As noted above, all of the other variables that were used in the original Cobb-Douglas function were not significant and did not contribute to the analysis of variance so they were removed from the regressions and only the variables shown in equations (5) and (6) were used in the final analysis.

The relationship between the level of demand and the optimal inventory level in terms of days of blood usage for both an independent bank and a member of a central system is illustrated in Figure 5.13. This figure was computed from the equations above and from equation (1) for target inventory at an independent bank for the case where the transfusion fraction at each bank is p = 0.5 and the cross-match release time is R = 2 days. The optimal inventory level at a hospital blood bank can be reduced by 20 to 50 percent for an HBB that has its inventory level managed by a community blood center.

5.3.8 Centralized blood bank issuing and allocation policies to HBBs

After the CBC receives all the requests from the HBBs, the orders are filled by drawing from the inventory in the CBC using an oldest to youngest age of units issuing policy. For purposes of simplification as well as good medical practice, each group and Rh factor is considered independent of the other groups and Rh factors. When the sum of all type-specific HBB demands exceeds the total inventory in the CBC, the CBC may backlog the excess demand or may fill all demands by calling in donors, by contacting other CBCs, by using frozen packed red cells or by requesting an emergency shipment from still higher echelon (regional) blood banks. In this analysis the CBC uses different approaches to handle the excess demand depending upon whether the orders are routine or emergency. Routine orders are placed by the HBBs at the beginning of each day to build up their inventory to a specific level. Emergency orders are placed during the day when the inven-

Figure 5.13 Optimal days of inventory to keep on hand to meet transfusions for a given blood type



tory of the HBBs cannot meet their respective users' demands. For routine orders, the CBC will fill the orders as long as its inventory lasts and disregard the excess demands, if any. Consequently, the HBBs may not receive the full amount they ordered. For emergency orders, the CBC still fills the orders as long as its inventory lasts. However, if there are excess emergency demands, the CBC will attempt to fill them from the inventory of the HBBs within the system. Furthermore, if there is insufficient stock in the whole system to fill the excess emergency demands, then the CBC will fill them by contacting exogenous sources. The rationale of the different treatments for the three types of excess demands, i.e., the three types of "shortages" between routine and emergency orders, is that the routine orders are used to build up the buffer inventory in the HBBs. These routine orders may not represent actual transfusion demands that day. Therefore if the excess of the routine orders over the available inventory at the CBC is not filled, a true shortage will not necessarily occur. On the other hand, the emergency orders, if not filled, will most likely create a shortage, since the buffer inventory in the HBB has to be essentially depleted before the HBB will place an emergency order.

Since each HBB may not receive all that it has ordered, a systematic process is needed to allocate the available stock in the CBC to HBBs. This allocation process is called the allocation policy. Essentially there are three distinct practical alternatives:

1. The CBC picks an HBB and fills its order by the First Come First Served (FCFS) issuing policy, and then goes on to fill the next HBB order until all the stock runs out or all orders from HBBs are filled. This type of allocation process resembles the practice that exists in some blood banking systems.

2. The CBC ships an amount to each HBB such that the ratio of the amount received to the amount ordered is the same for each HBB. Furthermore, all shipments have the same ratio of the amount of different ages received to the amount ordered. This type of allocation process resembles proportional rationing of scarce resources and is intended to be fair to all users with regard to their stated target needs by treating each user equitably. (See Cohen, Pierskalla and Yen [22] for a theoretical treatment of this problem.)

3. The CBC ships each unit to the hospital where the shortage probability is the highest in the system. In other words, the delivery of each unit is intended to adjust the system stock configuration such that total system shortage probabilities may be improved. If the target level needs in policy 2 above are based on shortage probabilities, then this alternative policy coincides with policy 2. However, if the target level needs are based on some tradeoff between shortages and outdates, then policies 2 and 3 may differ slightly.

After all HBBs receive their orders it may be desirable to trans-ship units among them. Basically there are three reasons for, or types of, such trans-shipments: an emergency need at an HBB that cannot be met by the CBC; the shortage anticipating trans-shipment; and the outdate anticipating trans-shipment [14, 23, 24]. If one location anticipates a shortage while another location does not, then a trans-shipment from the latter to the former may be beneficial to the system in reducing the system shortage cost. Similarly, if one location has an excessive amount of old units while another location does not, an outdate anticipating trans-shipment can be initiated for the benefit of the system. Before a trans-shipment is made, the exact stock configurations of the locations, as well as the demand distributions of the locations, must be known in order to evaluate the benefit of the trans-shipment. When such information is available, the CBC is in the best position to direct the trans-shipments in the system. Obviously, for these types of actions a sophisticated information processing system is needed.

In the case where such information is not available, the benefits of trans-shipping are uncertain and no trans-shipment should be made directly from one HBB to another. However, since each HBB knows its own stock/age configuration, it can choose to return excessively old, but still usable, units to the CBC. In this way old units are recycled to other hospitals in the system. This particular type of outdate anticipating trans-shipment will be called the recycle policy.

Of the three conditions for a trans-shipment, the most important condition is when an HBB has an emergency demand and the CBC does not have sufficient stock on hand to meet it. In this case a check of the other HBBs should be conducted and a trans-shipment made provided the HBB which furnished the units will not be placed in a precarious shortage situation, that is, provided the probability of shortage at the sending HBB does not become too large after depletion of its stock.

Less important trans-shipments occur due to shortage or outdate anticipating trans-shipments. For shortage anticipating trans-shipments, a unit is trans-shipped from location A to B if the shortage probability in A is greater than that in B and if the difference of the two probabilities is greater than a certain number. The number should be large enough so that the trans-shipment will be beneficial to the system. It is calculated according to the following formula:

[shortage probability at A - shortage probability at B]
> transportation cost/shortage cost (7)

If the transportation cost is estimated to be about 5 percent of the shortage cost, then the number used in the determination of whether or not to trans-ship a unit is 0.05 (i.e., initiate a trans-shipment if the differential shortage probability is reduced by 0.05). Note that the shortage cost is assumed to be the same for all HBBs and the transportation cost is independent of the facilities where the trans-shipment occurred. This simplification is justified because the majority of the transportation costs are often not the direct costs, e.g., gas and time consumed in the shipment, rather the indirect costs related to the handling, labeling, accounting and information exchanged between the two facilities. All these indirect costs, however, depend upon the size of the system. Therefore, the number 0.05 can at best be described as an educated guess.

For outdate anticipating trans-shipments, a unit is trans-shipped from A to B if the outdate probability in A is greater than that in B and if the difference of the two probabilities is greater than the transportation cost divided by the outdate cost:

[outdate probability at A - outdate probability at B]
> transportation cost/outdate cost (8)

Again, 0.05 was used for this ratio (the number 0.05 is based on similar calculations and assumptions as those used above). It should be noted that in both cases the number 0.05 is somewhat arbitrary since actual costs are not known precisely. However, in the range between 0.03 and 0.20 there appears to be no significant difference in the number of units trans-shipped. Indeed, for this range, virtually no shortage or outdate anticipating trans-shipments will occur [14].

One reason why there are few shortage-anticipating trans-shipments stems from the allocation policy in the CBC. Recall that units are available for trans-shipment only after each HBB has received its delivery. But under allocation policies 2 or 3, the units in the CBC are issued one by one to the location with the highest shortage probability or proportionally to their target needs. So at the end of the allocation process each HBB will have an essentially identical shortage probability except when there is insufficient inventory in the CBC to make them equal or when there is a tie in shortage probabilities before the issuance of the last few units. In both of these cases, some discrepancies among shortage probabilities will occur, but they are rather negligible under relatively wide ranges of target inventory levels at all locations. Consequently, the conditions to initiate shortage trans-shipment would rarely occur, hence hardly any units are shortage trans-shipped. For this reason the shortage trans-shipment policy has virtually no significant effect on the shortages in the system.

The insensitivity of the outdated units to the outdate trans-shipment policy can be explained as well. By observing that a unit will be outdated only after several passages through the cross-matching process, the quantity of expected daily outdates is fairly small simply because the probability of outdate given by $(1-p_j)^n$ is usually a very small number where n is the number of times the unit is cross-matched prior to outdating. Hence, there are very few units which outdate, when optimal inventory, issuing, p_j and R policies are followed, regardless of whether an outdate trans-shipment policy is in effect or not. Consequently, the outdate trans-shipment policy can be expected to have virtually no significant effect on the outdates in the system.

It should be mentioned here that the simulation model also indicated that while there are some units trans-shipped, the actual quantities were insignificant even when the inventory levels at different locations were varied over wide ranges. However, if the actual inventory levels used are far larger than the optimal target inventory levels at the HBBs, then as one would expect, outdate trans-shipments would become significant if the allocation policy is changed to the FCFS allocation policy. Both of these decisions are extreme and should not be followed. That is, the CBC should use optimal target inventory levels and should not use allocation policy 1 (FCFS).

We now summarize the best trans-shipment and allocation policies:

1. Use allocation policy 2. Allocation policy 3 is also good but requires more computation and time for implementation.

2. Trans-ship units from one HBB to another.

   a) If there is an emergency need at an HBB and if the CBC is out of stock and if the sending HBB does not incur an excessive probability of shortage (say over 10 percent).

   b) If the probability of shortage at HBB_i minus the probability of shortage at HBB_j is greater than or equal to the ratio of unit transportation cost to shortage cost.

   c) If the probability of outdate at HBB_i minus the probability of outdate at HBB_j is greater than or equal to the ratio of unit transportation cost to outdate cost.

5.3.9 Optimal cross-matched release and issuing policies from the CBC

Cohen and Pierskalla [17] show that if a unit is cross-matched at an HBB and not reported transfused within a short time (R = 1, 2 or 3 days), further information should be obtained on the status of the demand for which the unit was issued. If the demand had disappeared, the unit should be made available for possible reassignment either at the same bank or another hospital blood bank. In this manner, the cross-match release time, R, should be kept as low as possible. As long as R can be maintained below 4 days, the FIFO issuing policy should be followed at the CBC for those HBBs which receive daily or at least tri-weekly deliveries from the CBC. If R exceeds 7 days, last-in first-out (LIFO) will be somewhat better than FIFO but both policies will then have excessive outdates and shortages.

In another study of issuing policies in an HBB [25], it was shown that for a department which has low usage and low values of p_j, a LIFO issuing policy for that department should be followed. The underlying reason why LIFO should be followed rather than FIFO is to increase the probability of transfusion of the cross-matched unit. This same reasoning applies to some HBBs in a CBC system, namely, those HBBs which require infrequent deliveries (weekly) and have low transfusion probabilities. This case often occurs at small distant rural HBBs. For these HBBs, the CBC should issue by LIFO and then at the next delivery pick up any non-transfused units, replacing them with younger units. The slightly older units which are then picked up may be made available to HBBs with higher volume needs which have higher transfusion probabilities.

Optimal policies include:

1. The cross-match release period, R, should be 1 or 2 days (the smaller that R is, the lower are the shortages, outdates, and costs).

2. For HBBs which receive daily, triweekly or biweekly deliveries, the units which are shipped to them should be issued on a FIFO basis (unless fresh units are needed for special purposes such as cardiac surgery).

3. For HBBs with infrequent deliveries (once a week), the units which are shipped to them should be issued on a LIFO basis and unused units from the prior shipment should be picked up and replaced with younger units.

5.4 CONCLUSIONS AND FURTHER RESEARCH

This chapter has considered a number of contributions to the development of operational procedures for blood bank management. In regionalization, it was shown that economies of scale exist in most of the blood bank management functions. Consequently a centralized community blood center is more efficient than a decentralized system. In addition, algorithms were developed to provide optimal allocation of HBBs and donor sites to CBCs in the case in which a region has multiple CBCs. Optimal target inventory levels, allocation, trans-shipment and issuing policies were shown for CBCs with central and with coordinated controls. Time series methods were applied to daily type-specific cross-match and monthly total cross-match data. These methods led to models for forecasting mean daily demands that are required for inventory control. A simulation model and statistical analysis was used to develop a target inventory decision function for inventory levels at an independent hospital blood bank, at HBBs that are a part of a centralized system and at the CBC. The mean daily demand, the transfusion to cross-match ratio and the cross-match release period were shown to be significant variables. Many of the key decisions in blood system management were analyzed and developed. However, there are still many open research questions that should be addressed for a more complete understanding of this supply chain.

In 1984, Prastacos [1] noted some unresolved research issues in his survey paper. They are still unresolved today. He noted the need for research on:

Optimal component processing policies Because the demand for components has risen greatly due to new medical technologies and therapies, upwards of 95% of whole blood units drawn are processed into various components. Furthermore many components are being collected by pheresis. There are differing quality and cost aspects to these two methods that need analysis and modeling. In addition to the practical needs of the blood banks in this area of component processing, there is a major need for more research in inventory theory for developing and analyzing mathematical models in which a common input source is subdivided into value added components. Deuermeyer [26, 27] developed optimal inventory policies for a product model that also produced a valuable by-product. But very little theory has been developed since his work.

Distribution scheduling of multiple products from the Center to the hospitals With the increase in use of components and their differing shelf lives, the immediacy of delivery for some of them in order to maximize their useful lives combined with the less demanding delivery of relatively long shelf-life red cells poses new logistics problems for the CBC.

Organizational structures for regional systems Although much work has been done (as noted above), there are major problems of centralization/ decentralization involving contractual relations between the CBC and the HBBs. These problems involve, but are not restricted to, who owns the blood products and at what points in time, what are the agency relationships and how can they be priced to maximize the overall societal benefits vis-à-vis the individual parties' benefits and what are the game relationships among the parties and is there equilibria. Here again there is need for theory to illuminate the issues and practice to achieve the most desirable results for donors and patients.

Pricing of blood products and inter-regional cooperation To some extent there is a war out there. Many of the suppliers are in heavy, mostly negative competition among themselves and with many of the HBBs.

Donor scheduling algorithms Frequently it is the case in a region that mobile and in-house drawings and pheresis drawings are seasonally bunched or else have seasonal gaps. In either case, the supply is not smoothed to meet demand and there is either excess outdating or shortages. Because of the very stochastic nature of both the supply and the demand processes, adaptive stochastic modeling is needed to improve the system.

There are many more research areas that could be mentioned but the above areas give a flavor of the still large knowledge needs for optimal blood products supply chain management. Since some studies [28, 29] have estimated that blood products can be very costly due to their significant utilization in many procedures, and this use accounts for about 1% of total hospital costs in the United States, small improvements can yield significant national savings.

References

[1] Prastacos, G. (1984). Blood Inventory Management: An overview of theory and practice. *Management Science,* 30, 777-800.

[2] Nahmias, S. (1982). Perishable inventory theory: An overview. *Operations Research,* 30, 680-708.

[3] Kern, D.A. and S.T. Bennett (1996). Informatics applications in blood banking. *Clinics in Laboratory Medicine,* 16, 947-960.

[4] Roberts, S. (2003). When the supply chain becomes a matter of life and death. *Frontline Solutions,* 12, 14-16.

[5] Sirelson, V. and E. Brodheim (1991). A computer planning model for blood platelet production and distribution. *Computer Methods and Programs in Biomedicine,* 35, 279-291.

[6] Brennan, J.E., B.L. Golden, and J.K. Rappoport (1992). Go with the flow: Improving Red Cross bloodmobiles using simulation analysis. *Interfaces,* 22, 1-13.

[7] Michaels, J.D., J.E. Brennan, B.L. Golden, and M.C. Fu (1993). A simulation study of donor scheduling systems for the American Red Cross. *Computers and Operations Research,* 20, 199-213.

[8] Cohen, M.A. and W.P. Pierskalla, (1975). Management policies for a regional blood bank. *Transfusion,* 15, 58-67.

[9] Cohen, M.A., W.P. Pierskalla, R.J. Sassetti, and J. Consolo (1979). An overview of a hierarchy of planning models for regional blood bank management. *Transfusion,* 19, 526.

[10] Cohen, M.A., W.P. Pierskalla, and R.J. Sassetti (1981). Regionalization of blood banking services: Alternative models of regional blood service systems. *Proceedings of the 1980 Conference on the Management and Logistics of Blood Banking,* Sponsored by the National Heart, Lung and Blood Institute, 5, 201-238.

[11] Cohen M.A., W.P. Pierskalla, and R.J. Sassetti (1987). Economies of scale in blood banking. In *Competition in Blood Services,* Clark, G.M., Ed, American Association of Blood Banks, 25-39.

[12] Or, I. (1976). *Traveling Salesman-Type Combinatorial Problems and Their Relation to the Logistics of Regional Blood Banking.* PhD Dissertation, Department of Industrial Engineering and Management Science, Northwestern University, Evanston, IL.

[13] Or, I. and W.P. Pierskalla (1979). A transportation location-allocation model for regional blood banking. *AIIE Transactions,* 11, 86-95.

[14] Yen, H. (1975). *Inventory Management for a Perishable Product Multi-Echelon System.* PhD Dissertation, Department of Industrial Engineering and Management Science, Northwestern University, Evanston, IL

[15] Elston, R.C. and J.C. Pickrel (1963). A statistical approach to ordering and usage policies for a hospital blood bank. *Transfusion,* 3, 41.

[16] Elston, R.C. and J.C. Pickrel (1965). Guidelines to inventory levels for a hospital blood bank determined by electronic computer simulation. *Transfusion,* 5, 465.

[17] Cohen, M.A. and W.P. Pierskalla (1979). Target inventory levels for a hospital blood bank or a decentralized regional blood banking system. *Transfusion,* 19, 444-454.

[18] Cohen, M.A., W.P. Pierskalla, and R.J. Sassetti (1981). Regional blood inventory control and distribution. *Proceedings of the 1980 Conference on the Management and Logistics of Blood Banking,* Sponsored by the National Heart, Lung and Blood Institute, 5, 21-88.

[19] Cohen, M.A., W.P. Pierskalla, R.J. Sassetti, and K. Walkky (1980). Time series forecasts of daily crossmatch quantities. Working Paper, Department of Decision Sciences, University of Pennsylvania, Philadelphia, PA.

[20] Cohen M.A., W.P. Pierskalla, and R.J. Sassetti (1982). The impact of adenine and inventory utilization decisions on blood inventory management. *Transfusion,* 23, 54-58.

[21] Pierskalla, W.P. and C.D. Roach (1972). Optimal issuing policies for perishable inventory. *Management Science,* 18, 603-614.

[22] Cohen, M.A., W.P. Pierskalla, and H. Yen (1980). Multi-echelon, age differentiated inventory systems. In *Multi-Level Production/Inventory Systems: Theory and Practice,* Schwarz, L., Ed, TIMS Studies in the Management Science Series, Elsevier, The Netherlands.

[23] Jennings, J.B. (1968). An analysis of hospital blood bank whole blood inventory control policies. *Transfusion,* 8, 335.

[24] Jennings, J.B. (1973). Blood bank inventory control. *Management Science,* 19, 637.

[25] Deuermeyer, B.L., W.P. Pierskalla and R.J. Sassetti (1976). Methods for reducing outdating without altering physician ordering patterns. Working Paper, Department of Industrial Engineering, Northwestern University.

[26] Deuermeyer, B.L. and W. P. Pierskalla (1978), A by-product production system with an alternative. *Management Science,* 24, 1373-1383.

[27] Deuermeyer, B.L. (1976). *Inventory Control Policies for Multi-Type Production Systems with Applications to Blood Component Management.* PhD Dissertation, Department of Industrial Engineering and Management Science, Northwestern University, Evanston, IL.

[28] Wallace, E.L. (2001). Blood services costs and charges. *Transfusion,* 41, 437-439.

[29] Cantor, S.B., D.V. Hudson, Jr., B. Lichtiger, and E.B. Rubenstein (1998). Costs of blood transfusion: A process-flow analysis. *Journal of Clinical Oncology,* 16, 2364-2370.

6 EVALUATING THE EFFICIENCY OF HOSPITALS' PERIOPERATIVE SERVICES USING DEA

Liam O'Neill[1] and Franklin Dexter[2]

[1] Department of Policy Analysis and Management
Cornell University
Ithaca, NY 14853

[2] Department of Anesthesia
University of Iowa
Iowa City, IA 52242

SUMMARY

Elective surgery typically generates 40 percent or more of a hospital's total revenue, and individual surgeons almost always have a net positive contribution margin. Perioperative services include surgical operations, pre-operative care of patients, and post-operative care. This chapter presents a method to identify best practices among hospitals' perioperative services using Data Envelopment Analysis (DEA). This analysis included 44,033 procedures performed by 3,502 surgeons at 53 non-metropolitan Pennsylvania hospitals. Eight procedures, each performed by one surgical specialty, were selected. For each hospital, DEA 1) identifies untapped markets for surgery; 2) identifies relatively high and low procedure volumes among specialties; and 3) suggests a strategy for increasing surgical volume for inefficient hospitals. Findings may be used by managers of perioperative services to aid in resource allocation decisions, such as hiring and recruitment among different surgical specialties.

KEY WORDS

Data envelopment analysis, Perioperative services

6.1 INTRODUCTION

Elective surgery typically generates 40 percent or more of a hospital's revenue, and individual surgeons almost always have a positive contribution margin [1, 2]. Hospitals depend on their surgeons for a steady flow of patients and revenue. Yet the management of the surgical functions within a hospital is a very complex and demanding task. This chapter presents a method to identify best practices among hospitals' perioperative services using Data Envelopment Analysis (DEA).

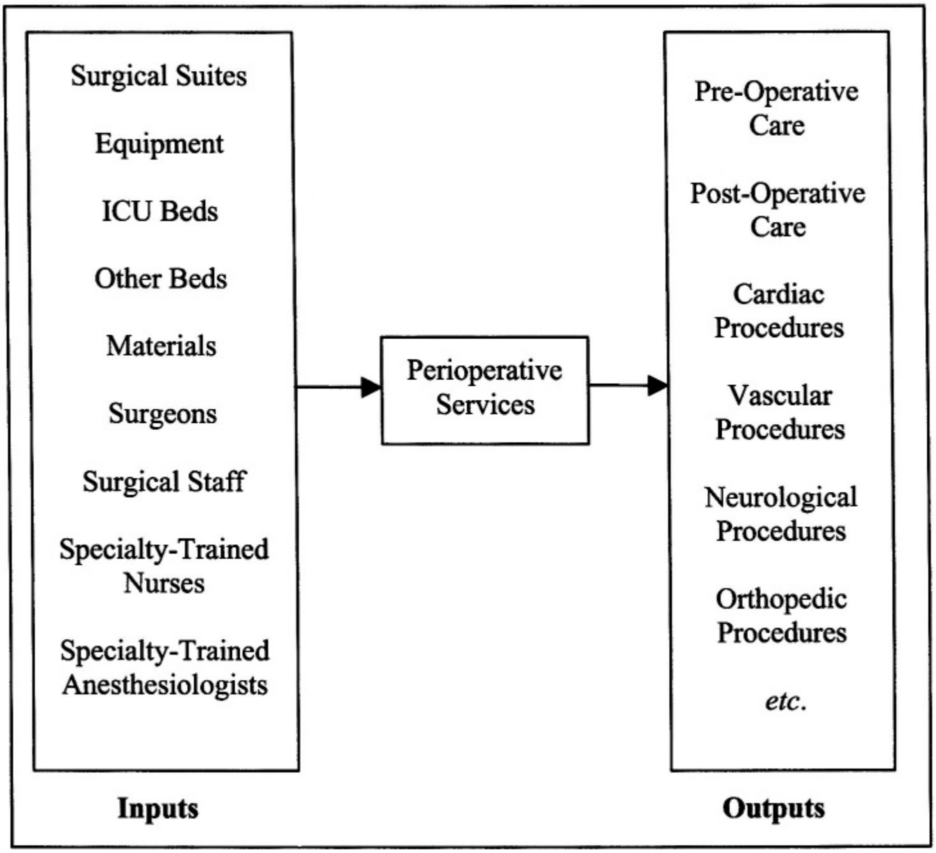
Elective surgery differs from non-elective surgery, such as trauma and transplant surgery, in that elective procedures are scheduled in advance. Perioperative care begins once the decision is made that a patient will undergo surgery at a hospital. We define *perioperative services* (POS) as the sub-system of a hospital that produces elective surgery, pre-operative care, and post-operative care. As such, POS is a complex system that uses multiple inputs, such as capital and personnel, to produce multiple products, such as procedures by specialty (Figure 6.1). POS can be thought of as a "hospital-within-a-hospital" that encompass all the functions associated with elective surgery.

Much of POS is isolated from the rest of the hospital, not just practically but physically. Personnel cannot enter operating rooms without wearing surgical scrubs and masks. Operating room (OR) nursing has little overlap with other types of nursing, and requires a year of additional training. Nurse anesthetists and anesthesiologists have little non-perioperative work. Surgical equipment and anesthesia machines are used under few other circumstances.

The Director of POS is typically a nursing or medical director, and if a medical director is usually an anesthesiologist [3]. In allocating scare resources, such as OR time, equipment, and staff, the director must weigh the demands of different surgical specialties. Operating room nurses in hospitals usually focus on three or fewer surgical specialties, as do many anesthesiologists. Deciding whether the anesthesiology department's new member has subspecialty training in cardiac surgery or regional anesthesia (i.e., for orthopedics) balances one surgical specialty against one another. Which specialties are favored can significantly impact surgeons' flexibility and access to POS.

The strategic factors that determine a hospital's potential workload for elective surgery have been well-established: Erickson and Finkler [4] showed that hospital market share is driven by its visibility in the community and the number of physicians with privileges at the hospital. They also

Figure 6.1 Model of production for perioperative services



showed that hospital visibility is driven in turn by the number of beds, number of services provided, and teaching status. Adams et al. [5] found that patients were willing to travel further for teaching hospitals with more acute care beds and more sophisticated services. All other things being equal, patients have strong preferences for local hospitals [6]. Hence the number of potential patients within a hospital's county and region are important predictors of the hospital's workload [5, 7].

The director of POS has little control over the strategic factors that determine a hospital's potential workload, but he or she does have significant control over operational factors that determine the proportion of the potential workload that is actually done at the hospital. This is particularly true in competitive markets where surgeons may have multiple hospital affiliations. The perioperative system typically has three bottlenecks: access to convenient operating room time, availability of specialized surgical equipment, and (for some procedures) open and staffed intensive care unit (ICU) beds. For operating room scheduling, the "jobs" are not patients but surgeons, since patients are more flexible in their availability. Surgical suites vary dramatically in their flexibility for booking cases. If the waiting time is too long, then the patient will likely receive care elsewhere, either with the same surgeon or a competing surgeon. On a long-term basis, surgeons who cannot get convenient OR time at one hospital tend to gravitate to other hospitals that can better meet their needs. Other operational factors that influence where surgeons choose to practice include availability of specialty-trained nurses and equipment, staff turnover times between consecutive cases, and availability of ICU beds.

6.1.1 Evaluating the performance of perioperative services

A number of practical difficulties arise in evaluating the performance of POS across institutions. Hospitals differ significantly in factors that influence the demand for elective surgery, such as the number of staffed beds, technological services offered, and size of the market. Therefore, a good evaluation method should compare a hospital's POS with peer entities that operate in a similar environment and use a similar combination of resources to produce a similar product mix. The method should accommodate system complexity in the form of multiple outputs and multiple inputs. The method should capture the tradeoffs faced by managers in allocating resources to different specialties, as well as the potential for substitution among the inputs [8]. Finally the measure should be clinically meaningful and relevant to physicians and OR managers.

Previous analyses of the performance of POS have mostly used ratio methods. Among the ratios that have been used are the following: delay in on-time start per case [9]; contribution margin per case [1]; labor cost per case [10]; patient waiting time per case [11]; anesthesia drug costs per case [12]; and anesthesia relative work units per case [13]. These ratios provide one-dimensional measures of how well POS is doing at one task or specialty. There is no clear way to collapse these multiple ratios into a single performance measure. Moreover, the ratios themselves are based on the workload performed at one hospital, rather than comparisons among hospitals. They do not measure or predict the facility's expected perioperative workload compared with the best practices at peer institutions.

DEA offers several advantages over previous ratio methods [8, 14]. First, DEA combines multiple ratios into a single ratio of productive efficiency. Second, DEA allows for resource substitution among the inputs as well as managerial tradeoffs among the outputs. Third, DEA compares each hospital to its peers and identifies benchmark facilities for inefficient hospitals.

This chapter extends the use of DEA in health care to perioperative services. The results of this model can be used by directors of POS to aid in resource allocation decisions, such as hiring and recruitment among different surgical specialties and capital equipment purchasing. For inefficient hospitals, the results can suggest how to increase surgical volumes. The remainder of this chapter is organized as follows: In the next section, we review DEA and the specific formulations that were chosen for this study. This is followed by a description of our data and methods (Section 6.3), results of our analysis (Section 6.4), model validation (Section 6.5), and conclusions (Section 6.6).

6.2 DATA ENVELOPMENT ANALYSIS

Data envelopment analysis (DEA) is a linear-programming-based technique to measure the technical efficiency of Decision-Making Units (DMUs). DEA works by estimating a piece-wise linear envelopment surface, known as the *best-practice frontier*. DEA is a deterministic, non-parametric technique, and thus makes no assumptions about the underlying form of the production function or the distribution of error terms. This technique accommodates multiple inputs and multiple outputs without prior knowledge of their relative prices.

DEA has been applied extensively in health care and has been shown to offer several advantages over other techniques, such as multivariate regression [15], ratio analysis [8], and other econometric approaches [16]. For a review of DEA health care studies, see Ozcan [17] and Hollingsworth et al. [16]. Areas of application include hospitals [15, 18-20], physicians [8, 14], nursing homes [21], and health maintenance organizations [22]. This chapter extends the use of DEA in health care to perioperative services.

To estimate the efficiency of surgical hospitals, the CCR (Charnes, Cooper, and Rhodes) input-oriented model was used [23, 24]. The CCR model can be formulated as follows: Suppose that there are *n* DMUs, each of which uses *m* inputs to produce *s* outputs. Let X_{ij} $(i = 1,..., m)$ be the amount of input *i* used by DMU*j*; let Y_{rj} $(r = 1,..., s)$ be the amount of output *r* produced by DMU *j* $(j = 1, ..., n)$. The technical efficiency of DMU 0 is then given by

$$max \;\; h_0 = \frac{\sum_{r=1}^{s} u_r Y_{r,0}}{\sum_{i=1}^{m} v_i X_{i,0}} \tag{1}$$

$$\frac{\sum_{r=1}^{s} u_r Y_{r,j}}{\sum_{i=1}^{m} v_i X_{i,j}} \leq 1 \quad j = 1,...,n, \tag{2}$$

$$u_r \geq 0, v_i \geq 0, \forall\, r, i.$$

Equation (1) represents the ratio of DMU 0's *virtual* output to its *virtual* input. Each DMU is free to choose the weights, u_r and v_i, that maximize its efficiency score, with only one set of constraints (equation 2). Efficient DMUs are those for which it is possible to find a set of weights for which the efficiency ratio is equal to one. Otherwise, the DMU's efficiency score will be less than one and it will be regarded as inefficient. The constant returns-to-scale CCR formulation was used because previous studies of physician efficiency have not found variable returns-to-scale [8, 25]. There is some evidence of increasing returns-to-scale for hospitals owing to horizontal integration [26].

In order to derive additional information about the hospitals we studied, we incorporated extensions to basic DEA including super-efficiency, known as the AP (Anderson and Peterson) model [27] and multifactor efficiency (MFE) [28]. The AP model is identical to the CCR model, except that the self-referential constraint in equation (2) is relaxed, allowing the efficiency score to exceed one [27]. The AP model has been used to identify potential data errors and to rank efficient DMUs [27, 29]. One drawback to the latter approach is that super-efficiency scores tend to be higher for *maverick* DMUs, i.e. those DMUs that place all their emphasis on one output and one input in equation (1) [28]. Multifactor efficiency overcomes this weakness by using the slack values from the AP model to rate each DMU with respect to all output-input combinations.

A robustness index, R_i, was calculated to measure the sensitivity of the AP scores with respect to changes in the input and output weights:

$$R_i = \frac{MFE_i}{AP_i}, \quad 0 < R_i \leq 1 \tag{3}$$

When R_i is close to 1, the AP score is relatively insensitive to changes in the input and output weights. A small value of R_i indicates a specialist orientation.

6.3 DATA AND METHODS

Patient data on inpatient admissions during 1998 from all non-Federal Pennsylvania hospitals were obtained from the Pennsylvania Health Care Cost Containment Council. Hospital variables were derived from the 1998 Annual Survey of the American Hospital Association. The study sample consisted of the 53 Pennsylvania hospitals that have at least 200 staffed beds and are located in non-metropolitan areas.. A non-metropolitan area was defined as a county with a population of less than one million people, based on the 1990 census.

6.3.1 Defining inputs and outputs

Eight surgical procedures were selected to measure surgical output (Table 6.1). These eight procedures were chosen to represent a wide spectrum of elective, scheduled, inpatient surgical procedures performed at a hospital. Each procedure serves as a proxy for the total surgical caseload within its respective specialty. Specifically, each procedure is performed by only one specialty. For example, we did not include carotid endarterectomy which is performed both by vascular surgeons and neurological surgeons, there is significant correlation with total inpatient workload for each specialty. Each of the eight procedures is correlated with the total inpatient workload for its respective specialty. Also, the procedures studied were those that are performed once per hospitalization. Thus, the number of hospitalizations is proportional to resource use. For example, hip replacement was included but not knee replacement, since some patients undergo one knee replacement during hospitalization (one such procedure) whereas others undergo bilateral knee replacement (two such procedures). Hospital discharges were selected based on the six ICD-9-CM procedure codes listed in the hospital discharge abstract.

We used the Diagnosis-Related Groups (DRG) Case-Mix index as a measure of the relative resource use of each procedure. The weights were determined by the modal DRG weight for hospital discharges including each of the procedures (Table 6.1). Coronary Artery Bypass Graft (CABG) had the highest DRG case-mix weight (5.65); hysterectomy had the lowest (0.77).

The eight procedures studied accounted for 7.5 percent of all inpatient discharges in the State of Pennsylvania. Hospitals also produce other outputs that were not included in this analysis, including outpatient care, medical and pediatric inpatient care, non-elective surgical care such as trauma and transplant surgery, post-graduate medical training, and research. However, this study focuses on the outputs of elective, scheduled perioperative care.

Hospital size and capacity were measured by the number of staffed beds (“BEDS”) and the use of technological services (“TECH”). Hospital technology was measured as the number of high-technology services offered, including the following: cardiac catheterization, cardiac surgery, shock-wave urological lithotripsy, megavoltage radiation therapy, magnetic resonance imaging, organ transplantation, neonatal intensive care, cardiac intensive care, and certified trauma care. A constant (c = 1) was added to the

Table 6.1 Description of input and output variables

Variable	Description	Specialty	DRG Weight
Outputs			
AAA	Abdominal Aortic Aneurysm resection	Vascular	4.08
CABG	Coronary Artery Bypass Graft	Cardiac	5.65
COLO	Colorectal Resection – excision of colon and/or rectum	General	3.13
CRAN	Craniotomy not for trauma – brain surgery (DRG 1)	Neurological	3.23
HIP	Hip Replacement	Orthopedic	2.37
HYS	Hysterectomy – removal of the uterus	Gynecology	0.77
LUNG	Lung Resection – excision of a piece or all of a lung	Thoracic	3.04
NEPH	Nephrectomy – removal of a kidney	Urology	2.65
Inputs			
BEDS	Number of staffed beds at the hospital		
TECH	Number of high-tech services offered at the hospital		
SURGEONS	Number of surgeons who did at least one of any of the above eight procedures		
COUNTY	Number of above eight procedures done on residents of hospital's county, weighted by case-mix		
CONTIG. COUNTY	Number of above eight procedures performed on residents of hospital's region, weighted by case-mix		
Explanatory			
AFFIL	Average number of hospital affiliations per surgeon		
HOSP-SURG	Hospital's market share among its surgeons		
RURAL	Hospital located in a rural county		

technology variable in order to prevent unbounded solutions in the AP model due to zeroes in the input data [28].

The input "SURGEONS" was defined as the number of surgeons who generated at least one hospital discharge for any of the above procedures. Most previous studies of hospital efficiency have excluded the number of physicians because they are independent contractors who may admit patients to multiple hospitals [15]. For our purposes, it is important to include surgeons as an input, since they largely determine both the volume and the type of procedures that the hospital can perform.

The demand for surgery depends on the number of surgeons, population size, and population demographics, such as age and gender. County demand ("COUNTY") was measured as the total number of the aforementioned procedures performed on residents of each hospital's county, weighted by DRG case-mix index. Demand from contiguous counties ("CONTIGUOUS COUNTY") was defined as the number of procedures performed on residents of all those counties sharing a common border with the hospital's county.

6.3.2 Explanatory variables

Surgeons typically have privileges at multiple hospitals. As the number of hospital affiliations per surgeon increases, the surgeon is available less often at each hospital [4]. Scheduling access to OR time becomes more challenging, resulting in idle capacity in the form of unused OR time and empty beds. Therefore, we would expect the efficiency of POS to decrease at facilities where the surgeons operate at many other hospitals. To test this hypothesis, two measures of hospital-surgeon relations were used: mean number of hospital affiliations per surgeon ("AFFIL"), and a hospitals' *market share* among its surgeons ("HOSP-SURG"). The hospital's market share among its surgeons was defined as the sum of the eight procedures performed at the hospital divided by the sum of the eight procedures performed by all surgeons with privileges at that hospital. For example, suppose 10 surgeons performed 100 procedures at Hospital A and 100 procedures at all other hospitals. Then Hospital A's market share among its surgeons would be 50 percent.

Another explanatory variable denoted whether the hospital was located in a rural county ("RURAL"), as defined by the Office of Management and Budget (www.nal.usda.gov). DEA assumes that hospitals are peer decision-making units. If our strategy was successful, our results should not be significantly different for rural hospitals.

In order to investigate the factors associated with technical efficiency, a series of parametric (t-Tests) and non-parametric (Mann-Whitney) tests were performed. The log transform of "SURGEONS" and "BEDS" was used for the t-Tests. The chi-squared test of independence was used for the dichotomous variable "RURAL." These tests were done as part of validation, in order to determine whether the efficiency scores were correlated with our input or control variables.

6.4 RESULTS

The characteristics of the input and output variables are presented in Table 6.2. Three of the eight procedures – colorectal resection, hip replacement, and hysterectomy – were performed by every hospital. The average number of surgeons per hospital was 66. The average number of hospital affiliations per surgeon was 1.64.

DEA identified 24 hospitals as efficient and 29 as inefficient (Table 6.3). The average efficiency score was 0.91, based on the CCR model. The AP model identified Hospital 38 as the most influential observation, with a superefficiency score of 7.67. The second highest AP score was 2.59. Hospital 38 is examined in more detail below.

The MFE and R_i measures indicate the robustness of the AP score with respect to all output-input combinations [28]. The mean MFE score was 0.63, compared with a mean AP score of 1.22. Only six surgical hospitals had $MFE_i \geq 1$. Hospital 48 had the lowest robustness index, $R_{48} = 0.23$, identifying the hospital as a maverick.

For inefficient hospitals, DEA provided information on the sources of inefficiency, as shown by the slack values in Table 6.4. Hospital 3 produced 130 fewer CABGs and 184 fewer hysterectomies, compared with the best-practice frontier. Overall, the surgical volumes for Hospital 3 could be increased by 1/0.99 = one percent, while holding all inputs constant.

Hospital 10 had a positive slack for the number of surgeons (13) as well as county market (1,037). This finding indicates relatively low productivity at the hospital among its surgeons. The surgeons may face barriers in getting surgery done at the hospital. Hospital 10 also had excess surgical demand in its county, indicating that this facility was losing market share to other hospitals. By contrast, the "county market" slack value is zero for surgical Hospital 50, indicating it has a large share of the local market but not the regional market, since its slack for CONTIGUOUS COUNTY is positive.

Table 6.2 Descriptive statistics for input and output variables

Variable	Mean	Standard deviation	Minimum	Maximum
Outputs				
AAA	25	20	0	85
CABG	229	290	0	1,113
COLO	138	81	11	381
CRAN	34	48	0	261
HIP	131	77	26	367
HYS	226	173	27	893
LUNG	28	24	0	110
NEPH	20	20	0	119
Inputs				
BEDS	312	112	203	659
TECH	6	2	1	10
SURGEONS	66	39	9	214
COUNTY	6,295	3,924	406	14,200
CONTIG. COUNTY	39,821	23,986	3,484	83,841
Explanatory				
AFFIL	1.64	0.46	1.03	3.20
HOSP-SURG	0.72	0.21	0.19	1.00
RURAL	0.19	0.39	0	1

6.5 MODEL VALIDATION AND INTERPRETATION

In order to test the validity of our model, we now focus on three hospitals in more detail and compare our results with other available evidence.

Hospital 38 is a tertiary facility in an integrated health system and health maintenance organization with more than two million enrollees. It is a regional referral center for central and northeast Pennsylvania. The hospital is located in a small, rural county with a population of 18,000. Hospital volumes were high for complex, resource-intensive procedures, such as craniotomy and CABG. The slack for county demand for this hospital was zero, indicating that it was drawing many of its surgery patients from outside

Table 6.3 Efficiency scores for 53 hospitals*

DMU	CCR	AP	MFE	R_i
1	1.000	1.204	0.610	0.507
2	1.000	1.599	0.978	0.612
3	0.992	0.992	0.771	0.777
4	0.912	0.912	0.628	0.688
5	1.000	1.339	1.113	0.832
6	1.000	1.215	0.948	0.781
7	1.000	1.012	0.557	0.550
8	0.660	0.660	0.458	0.695
9	0.976	0.976	0.587	0.602
10	0.858	0.858	0.652	0.760
11	1.000	1.478	1.052	0.712
12	1.000	1.137	0.724	0.637
13	0.706	0.706	0.335	0.475
14	1.000	2.586	1.005	0.389
15	1.000	1.291	0.703	0.545
16	1.000	1.230	0.413	0.336
17	0.919	0.919	0.554	0.603
18	1.000	1.252	0.946	0.756
19	0.728	0.728	0.384	0.528
20	0.974	0.974	0.385	0.395
21	0.929	0.929	0.302	0.325
22	1.000	1.232	0.768	0.623
23	0.818	0.818	0.411	0.502
24	0.710	0.710	0.235	0.330
25	0.752	0.752	0.453	0.603
26	0.899	0.899	0.672	0.747
27	0.963	0.963	0.378	0.393
28	0.993	0.993	0.477	0.481
29	0.933	0.933	0.456	0.489
30	0.934	0.934	0.480	0.514
31	1.000	2.129	1.115	0.524
32	0.538	0.538	0.313	0.582
33	1.000	1.220	0.710	0.582
34	0.836	0.836	0.715	0.855

Table 6.3 (cont.) Efficiency scores for 53 hospitals*

DMU	CCR	AP	MFE	R_i
35	1.000	1.122	0.664	0.592
36	0.842	0.842	0.353	0.419
37	1.000	1.067	0.386	0.362
38	1.000	7.669	2.255	0.294
39	1.000	1.214	0.639	0.526
40	1.000	1.325	0.802	0.605
41	0.561	0.561	0.336	0.599
42	0.863	0.863	0.270	0.313
43	0.736	0.736	0.495	0.672
44	1.000	1.031	0.407	0.394
45	1.000	1.674	1.179	0.704
46	0.609	0.609	0.306	0.503
47	0.987	0.987	0.451	0.457
48	1.000	1.830	0.423	0.231
49	0.652	0.652	0.482	0.739
50	0.833	0.833	0.371	0.445
51	1.000	1.138	0.712	0.626
52	0.885	0.885	0.656	0.742
53	1.000	2.386	1.138	0.477

* CCR = Charnes, Cooper, and Rhodes; AP = Andersen and Petersen; MFE = Multifactor Efficiency

its own county. The craniotomy volumes were in the 96th percentile and the CABG volumes were in the 88th percentile. The hospital's market share among its surgeons was 97 percent, the 5th highest in the sample. This was found to be the most influential observation, based on its superefficiency score.

Hospital 48 was found to be an efficient, "maverick" hospital (Table 6.3). This facility had some of the fewest surgeons (9), beds (204), and technological services (0) of all 53 hospitals. It is located in a small market, as measured by county (740) and regional (7,906) demand. Despite its difficult operating environment, this facility competed successfully by focusing on three procedures: colorectal resection, hip replacement, and hysterectomy. These procedures have a relatively low case-mix weight and require comparatively low investment in technology. The hospital was

Table 6.4 Slack analysis for inefficient hospitals

Table 6.4a Increased outputs

Hosp.	Effi-ciency Score	AAA	CABG	COLO	CRAN	HIP	HYS	LUNG	NEPH
3	99%	5	130	46	0	0	184	28	4
4	91%	0	0	0	16	6	81	17	11
8	66%	0	0	0	10	0	9	8	4
9	98%	1	0	21	49	0	0	5	16
10	86%	0	0	0	0	44	129	7	9
13	71%	13	94	0	16	0	0	11	8
17	92%	3	132	3	7	0	24	4	0
19	73%	1	0	0	9	7	73	19	12
20	97%	8	78	0	25	21	0	14	0
21	93%	4	8	0	0	8	0	4	1
23	82%	0	274	0	11	13	190	8	7
24	71%	3	4	0	1	0	0	0	2
25	75%	3	103	0	2	30	176	0	7
26	90%	0	94	0	44	19	58	1	0
27	96%	2	52	0	0	16	21	20	12
28	99%	10	249	0	29	34	0	0	5
29	93%	11	324	0	16	0	88	0	11
30	93%	0	62	0	12	33	95	1	9
32	54%	0	255	0	9	28	51	0	4
34	84%	0	54	0	0	64	103	0	4
36	84%	13	42	0	25	44	43	5	0
41	56%	9	156	0	5	16	66	0	9
42	86%	21	177	0	40	54	0	19	12
43	74%	5	158	0	5	0	5	15	2
46	61%	14	155	38	18	32	0	4	0
47	99%	14	142	0	33	11	0	12	0
49	65%	2	163	19	13	5	0	8	0
50	83%	8	159	0	29	12	54	19	0
52	88%	2	0	28	31	49	0	0	2

Table 6.4 Slack analysis for inefficient hospitals

Table 6.4b Decreased inputs

Hospital	Efficiency Score	Beds	Tech. Services	Surgeons	County Market	Regional Market
3	99%	0	0	1	3,053	0
4	91%	0	2	0	76	9,762
8	66%	0	1	0	1,097	15,957
9	98%	0	1	0	1,679	22,801
10	86%	0	0	13	1,037	0
13	71%	0	0	0	651	0
17	92%	72	0	0	677	0
19	73%	48	1	0	4,739	17,990
20	97%	82	1	0	250	34,818
21	93%	102	0	0	133	16,871
23	82%	0	0	0	1,935	0
24	71%	183	0	0	371	7,476
25	75%	0	0	0	4,102	33,470
26	90%	0	0	17	1,960	0
27	96%	27	2	0	10,481	44,521
28	99%	0	1	0	4,075	48,602
29	93%	0	0	0	6,988	44,337
30	93%	35	0	0	1,951	0
32	54%	0	0	0	3,536	16,294
34	84%	0	0	0	1,834	0
36	84%	147	0	0	4,755	11,780
41	56%	0	0	0	0	6,048
42	86%	15	0	0	498	0
43	74%	0	0	0	61	0
46	61%	0	2	0	2,528	10,847
47	99%	92	4	0	178	0
49	65%	0	1	0	958	0
50	83%	27	0	0	0	11,417
52	88%	0	0	15	1,911	17,523

efficient to a large extent because its market share among surgeons was 94 percent, which was in the 88th percentile. This means that 94 percent of the cases performed by these nine surgeons were performed at Hospital 48.

Hospital 10 is located in a competitive market with four other hospitals within its county. Table 6.4b shows that this facility has positive slack for both surgeons and county demand. This facility's surgeons have 2.8 hospital affiliations on average, the second highest in the sample. Its market share among its surgeons was 51 percent, which was in the 19th percentile. These findings corroborate the DEA results. If all the surgeons with privileges could be persuaded to admit all their patients to this hospital, then surgical volume would almost double.

The results of the post-hoc analysis for efficient and inefficient hospitals is shown in Table 6.5. As expected, efficient hospitals had fewer affiliations per surgeon than inefficient hospitals (1.45 vs. 1.80; P < 0.01). The distribution of affiliations per surgeon for efficient and inefficient hospitals is shown in Figure 6.2. Only two of the 24 efficient hospitals had an average of more than two affiliations per surgeon. By contrast, eight of the 29 inefficient hospitals averaged more than two affiliations per surgeon. Efficient hospitals had a higher market share among their surgeons (80 percent ±0.18 vs. 65 percent ± 0.22). This difference was statistically significant for the Mann-Whitney test (P = 0.009). Thus, a hospital's market share among its surgeons was positively associated with its overall POS efficiency.

There were no statistically significant differences between efficient and inefficient hospitals with respect to the other variables tested, including beds, surgeons, county, contiguous county, and rural location. Thus, there was little evidence of increasing returns-to-scale, as hospital POS efficiency was not associated with the size of the hospital or market.

6.6 DISCUSSION AND CONCLUSIONS

DEA has been applied extensively to other areas of health care, but this is the first study to apply DEA to hospitals' perioperative services. This study has demonstrated the usefulness of DEA in capturing the complexity and managerial tradeoffs that characterize perioperative services. In addition to measures of hospital capacity, our analysis included market factors that are significant predictors of surgical demand.

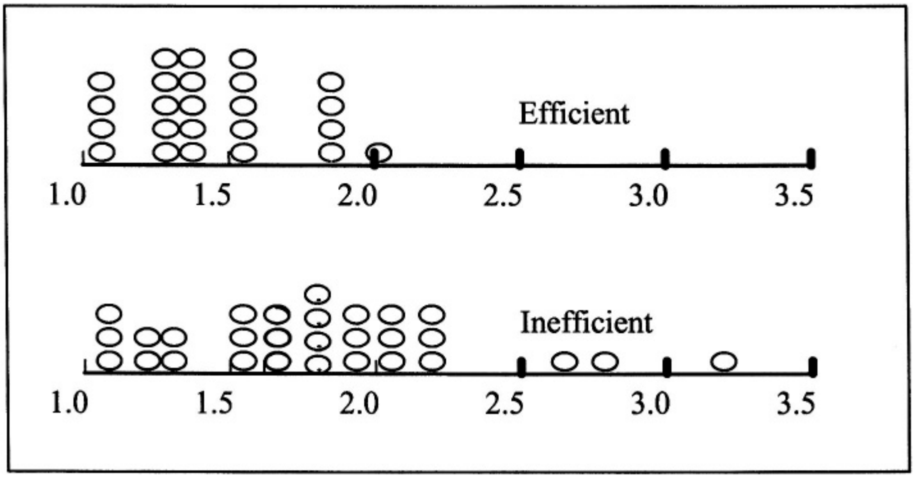
This study found that the strength of a hospital's relations with its surgeons is an important predictor of POS efficiency (P< 0.01). Hospitals having stronger relationships with their surgeons were more likely to be efficient.

Table 6.5 Differences between efficient and inefficient hospitals

Variable	Mean	t-Test	Mann-Whitney U	
Affiliations per Surgeon				
Efficient	1.45	2.96	195	Test statistic
Inefficient	1.80	0.005	0.006	p-value
Hospital's Market Share Among Surgeons				
Efficient	0.80	N/A	201	Test statistic
Inefficient	0.65		0.009	p-value
Beds				
Efficient	346	-1.90	-1.35	Test statistic
Inefficient	284	0.062	0.177	p-value
Surgeons				
Efficient	76	-1.28	-1.39	Test statistic
Inefficient	58	0.205	0.166	p-value
County				
Efficient	6,080	0.36	-0.58	Test statistic
Inefficient	6,473	> 0.2	> 0.2	p-value
Contiguous County				
Efficient	36,318	0.97	-1.03	Test statistic
Inefficient	42,719	> 0.2	> 0.2	p-value
Rural				
Efficient	0.208	N/A	0.111*	Test statistic
Inefficient	0.172		> 0.2	p-value

* Based on the chi-squared test of independence

Figure 6.2 Histogram of average affiliations per surgeon for efficient and inefficient hospitals



This finding is not surprising, since surgeons largely determine surgical volumes. In rural areas, larger hospitals may have a captive market for surgeons, since surgeons' choices may be limited by geography and other factors. In more competitive markets, hospitals may have to work harder to satisfy their surgeons and ensure a steady stream of patients.

Hospital managers and executives can use these results in several ways. For inefficient hospitals, DEA suggests ways to increase hospital volume. A positive slack for the number of surgeons indicates that the hospital has more surgeons than would be expected given its current surgical volume. This is an indication that the hospital needs to improve its market share among its surgeons. This could be accomplished by reducing scheduling delays in the OR, offering more amenities, or reducing turnover times between cases. Positive slack in some but not all procedures provides insight into which surgical specialties to focus on in capital equipment purchasing, recruiting sub-specialty trained anesthesiologists, and in training OR nurses. These are all operational factors that are under the control of the director of POS.

Future research should compare the DEA results with parametric, regression-based approaches in order to identify the comparative advantages of each method. Future research should also adapt this model to metropolitan areas where competition among hospitals for surgeons and patients is more intense.

References

[1] Dexter, F., J.T. Blake, D.H. Penning, and D.A. Lubarsky (2002). Calculating a potential increase in hospital margin for elective surgery by changing operating room time allocations or increasing nursing staffing to permit completion of more cases: a case study. *Anesthesia andAnalgesia.* 94, 138-142.

[2] Macario, A., F. Dexter, and R.D. Traub (2001). Hospital profitability per hour of operating room time can vary among surgeons. *Anesthesia andAnalgesia,* 93, 669-675.

[3] Mazzei, W.J. (1998). Should the director of perioperative services be a physician? *ASA Newsletter,* 62.

[4] Erickson, G. and S. Finkler (1985). Determinants of market share for a hospital's services. *Medical Care,* 23, 1003-1018.

[5] Adams, E.K., R. Houchens, G.E. Wright, and J. Robbins (1991). Predicting hospital choice for rural Medicare beneficiaries: the role of severity of illness. *Health Services Research,* 26, 583-612.

[6] Finlayson, S.R., et al. (1999). Patient Preferences for Location of Care: Implications for Regionalization. *Medical Care,* 37, 204-209.

[7] Cohen, M.A. and H.L. Lee (1985). The determinants of spatial distribution of hospital utilization in a region. *Medical Care,* 23, 27-38.

[8] Chilingerian, J. (1994). Exploring why some physicians' hospital practices are more efficient: Taking DEA inside the hospital, in *Data Envelopment Analysis: Theory, Methodology, and Applications,* A. Charnes, W. Cooper, A. Lewin, L. Seiford, Eds., Kluwer Academic Publishers, Boston, MA.

[9] Shelver, S.R. and L. Winston (2001). Improving surgical on-time starts through common goals. *AORNJournal,* 74, 506-513.

[10] Dexter, F., R.H. Epstein, and H.M. Marsh (2001). Statistical analysis of weekday operating room anesthesia group staffing at nine independently managed surgical suites. *Anesthesia and Analgesia,* 92,1493-1498.

[11] Dexter F, and R.D. Traub (2000). Statistical method for predicting when patients should be ready on the day of surgery. *Anesthesiology,* 93, 1107-1114.

[12] Dexter, F., D.A. Lubarsky, B.C. Gilbert, and C. Thompson (1998). A method to compare costs of drugs and supplies among anesthesia providers: A simple statistical method to reduce variations in cost due to variations in casemix. *Anesthesiology,* 88, 1350-1356.

[13] Abouleish, A.E., D.S. Prough, C.W. Whitten, M.H. Zornow, A. Lockhart, L.A. Conlay, and J.J. Abate (2002). Comparing clinical productivity of anesthesiology groups. *Anesthesiology,* 97, 608-615.

[14] Chilingerian, J. and D. Sherman (1990). Managing physician efficiency and effectiveness in providing hospital services. *Health Services Management Research,* 3, 3-15.

[15] Sherman, H.D. (1984). Hospital efficiency measurement and evaluation: Empirical test of a new technique. *Medical Care,* 22, 922-938.

[16] Hollingsworth, B., P. Dawson, and N. Maniadakis (1999). Efficiency measurement of health care: A review of non-parametric methods and applications. *Health Care Management Science,* 2, 161-172.

[17] Ozcan, Y.A. (1993). Sensitivity analysis of hospital efficiency under alternative output/input and peer groups: A Review. *International Journal of Knowledge and Public Policy,* 1, 1-31.

[18] Morey, R.C., Y.A. Ozcan, D.L. Retzlaff-Roberts, and D. Fine (1995). Estimating the hospital-wide cost differentials warranted for teaching hospitals: An alternative to regression approaches. *Medical Care,* 33, 531-552.

[19] Ozcan, Y.A., R. Luke, and C. Haksever (1992). Ownership and organizational performance. A comparison of technical efficiency across hospital types. *Medical Care,* 30, 781-794.

[20] Sexton, T., A. Leiken, A. Nolan, A. Hogan, and R. Silkman (1989). Evaluating managerial efficiency of veterans administration medical centers using data envelopment analysis. *Medical Care,* 17, 1175-1188.

[21] Chattopadhyah, S. and C.S. Ray (1996). Technical, scale, and size efficiency in nursing home care: A non-parametric analysis of Connecticut homes. *Health Economics,* 5, 363-373.

[22] Siddharthan, K., M. Ahern, and R. Rosenman (2000). Data envelopment analysis to determine the efficiencies of health maintenance organizations. *Health Care Management Science,* 3, 23-29.

[23] Charnes, A., W.W. Cooper, and E. Rhodes (1978). Measuring the efficiency of decision making units. *European Journal of Operational Research,* 2, 429-444.

[24] Charnes, A., W. Cooper, A. Lewin, and L. Seiford (1994). *Data Envelopment Analysis: Theory, Methodology, and Applications,* Kluwer Academic Publishers, Boston, MA.

[25] Pauly, M.V. (1980). *Doctors and Their Workshops: Economic Models of Physician Behavior.* University of Chicago Press, Chicago, IL.

[26] Harris, J., H. Ozgen, and Y. Ozcan (2000) Do mergers enhance the performance of hospital efficiency? *Journal of the Operational Research Society,* 51, 801-811.

[27] Andersen, P. and N.C. Petersen (1993). A procedure for ranking efficient units in Data Envelopment Analysis. *Management Science,* 39, 1261-1264.

[28] O'Neill, L. (1998). Multifactor efficiency in Data Envelopment Analysis with an application to urban hospitals. *Health Care Management Science,* 1, 19-27.

[29] Xue, M. and P. Harker (2002) Note: Ranking DMUs with infeasible super-efficiency DEA. *Management Science,* 48, 705-710.

7 BENCHMARKING USING DEA: THE CASE OF MENTAL HEALTH ORGANIZATIONS

Yasar A. Ozcan[1], Elizabeth Merwin[2], Kwangsoo Lee[3]
and Joseph P. Morrissey[4]

[1] Department of Health Administration
Virginia Commonwealth University
Richmond, VA 23298

[2] Southeastern Rural Mental Health Research Center
University of Virginia
Charlottesville, VA 22904

[3] Korean National Bureau of Health Insurance
Seoul, South Korea

[4] Cecil G. Sheps Center for Health Services Research
University of North Carolina
Chapel Hill, NC 27599

SUMMARY

This chapter uses windows and cone ratio analysis – a longitudinal and weight- restriction application of Data Envelopment Analysis (DEA) – to develop a methodology for analyzing organizational performance of community mental health centers (CMHCs); the chapter also develops measures of efficiency as a basis for improving productivity in behavioral health care. Specifically, non-hospital services provided by CMHCs were studied. Data limitations are noted in relation to use of the method and to the results. The model is shown to capture the impact of managed care on CMHC efficiency. The cone ratio version of the model, using weight restrictions, identified six super-efficient CMHCs, which had been consistently efficient since the implementation of managed behavioral care. The potential usefulness of this method for public and private mental health systems and for managed care companies is discussed.

KEY WORDS

Windows analysis, Efficiency, DEA, Weight restricted model, Mental health services

7.1 INTRODUCTION

The 1990s brought many challenges to behavioral health care organizations to improve the efficiency of their health care delivery. In response, behavioral health care organizations implemented different forms of managed care. Of particular importance has been the dramatic increase in managed care programs under Medicaid. Managed care contracts use pricing mechanisms to influence the use of services by controlling the amounts paid to health care providers and professionals. Effective cost control should of course be accompanied by a thorough understanding of the varying services provided by different mental health care providers, as well as by the use of good practice protocols for treating mental health conditions.

This chapter reports on a pilot investigation that uses data envelopment analysis (DEA) to develop methods for studying the technical efficiency of providers of community mental health care, in order to improve productivity. It focuses solely on the care of the seriously mentally ill (SMI) patients who receive services reimbursed by Medicaid. We examined 12 community mental health centers (CMHCs), all receiving traditional fee-for-service Medicaid reimbursement in years 1-2 (1994 and 1995). In years 3-5 (1996-1998), a mandatory, capitated Medicaid managed care program was implemented in the geographic areas served by eight of those CMHCs.

The measures of community mental health efficiency that we developed are tested by comparing the longitudinal patterns of provider efficiency over a five-year time frame, before and after implementation of the mandatory Medicaid managed care plan. We compare efficiency scores between the managed care site – the Tidewater area – and the control site – Richmond, Virginia. We also develop a structured method for identifying the effects of data limitations and the effects of ongoing modifications in managed care plans on the interpretation of findings.

The methods we developed offer a replicable, objective methodology that can be used to compare the operational efficiency of different types of providers who care for similar populations of clients. The methodology identifies consistent measures for comparison – numbers of patients treated – and provides a means of aggregating information on different numbers of patients to serve as a measure of organizational performance.

This methodology could be useful to public mental health systems as well as to private and public managed care companies, because it can identify the combinations of services that result in the most efficient care. That information can be used to change the mix of services that a managed care company will reimburse, and/or those that a provider chooses to use.

7.2 BACKGROUND

7.2.1 Relevance

The cost of health care in the United States was $943 billion in 1996, with over 10% of that money ($99 billion) spent on behavioral health care. Mental health disorders consumed 7% of the health care costs, with Alzheimer's disease/dementias and addiction disorders consuming 2% and 1% of total costs, respectively [1]. Eighteen percent of the expenditures on mental health went to multi-service mental health clinics, which include community mental health centers [1]. From 1986 to 1996, mental health costs rose 1% less than overall health costs did. One explanation for the lower rise is more use of cost-containment strategies by managed care companies, which resulted in increased efficiency and lower expenditures on mental health care [1]. Other possible reasons are Medicaid program design, reductions in inappropriate hospitalizations, use of non-mental health services, and the shift of mentally ill persons from inpatient care to the community [1].

Twelve percent of the United States population is covered under Medicaid for their health care. Medicaid's cost for behavioral health amount to 19% of its expenditures; per capita annual Medicaid mental health expenditure is approximately $481 [1]. These costs justify examination of efficiency in the provision of the mental health services by community mental health organizations.

It has been suggested that, to be effective, health insurance plans should provide the following services: 24 hour care/hospitalization, intensive community services, outpatient services, medication management, psychosocial rehabilitation, and outreach services (Frank et al. in [1]). It is less clear, however, which combinations of services are optimal for the treatment of specific population sub-groups.

Effective service delivery must result in desirable outcomes for patients. At the same time, budget constraints dictate that these outcomes must be achieved in a financially responsible manner: CMHCs must provide effective services with efficiency. The efficiency of mental health care providers is understudied and will be the focus of this chapter. We demonstrate the use of DEA as a methodology to answer the following questions: 1) How can different mental health services be used together for optimal efficiency? 2) How important are specific services in the overall efficiency and use of mental health services? 3) How do community mental health organizations compare in overall efficiency? These questions are explored within the context of the implementation of a Medicaid managed care program.

Efficiency has been evaluated in other states that have implemented Medicaid managed care programs. Results have been inconsistent. In Massachusetts the implementation of Medicaid managed care led to a 25% reduction in costly inpatient care, but an increase in rehospitalizations. In Utah there was no effect on use of inpatient care, but some differences in outpatient care. Utah enrollees with the worst mental health had the least improvement when Medicaid managed care was implemented. An evaluation of the Colorado program noted that utilization management did not focus as much on outpatient care as on inpatient care and that "utilization management strategies to provide outpatient services more efficiently are insufficient" [1-4].

7.2.2 Data envelopment analysis

DEA evaluates organizational performance by considering multiple inputs and outputs to identify the most efficient providers. DEA has been successfully applied in many industries [5], including the study of health care organizations and professionals. Sherman [6] and Nunamaker [7] were among the first to apply DEA measures to hospitals, examining hospitals in Massachusetts and Wisconsin, respectively. Huang and McLaughlin [8] applied DEA to programs for rural primary health care; Sexton and colleagues [9] applied DEA to the Veterans Administration Medical Centers. Applications of DEA in health proliferated in the 1990s, including studies of physicians [10-12], mental health programs [13, 14] nursing homes [15], aging agencies [16], and hospitals [17-19]. Collectively, these studies demonstrate that DEA is an effective research tool for evaluating the efficiency of health care providers, given varying input mixes and types and numbers of outputs.

DEA uses linear programming to search for optimal combinations of inputs and outputs, based on the actual performances of decision making units, in this case, CMHCs. In this chapter, we use DEA to evaluate the technical efficiency of each CMHC relative to "optimal" patterns of production. Patterns are computed using the performance of CMHCs whose inputs and outputs are not optimized by those of any other comparison or peer CMHC. DEA computes the relative efficiencies with which CMHCs combine major categories of inputs to generate general categories of outputs typically produced by providers. Controllable and uncontrollable inputs/outputs are taken into consideration, as is the size of each CMHC.

DEA also calculates inefficiency values for each CMHC. The inefficiencies are the degrees of deviance from the frontier. Input inefficiencies show the degree to which inputs must be reduced for the inefficient CMHC to lie on the efficient practice frontier. Output inefficiencies are the needed increase in outputs for the CMHC to become efficient. If a particular CMHC either reduces its inputs by the inefficiency values or increases its outputs by the amount of inefficiency, it could become efficient; that is, it could obtain an efficiency score of one.

Various types of DEA models can be used, depending upon the problem at hand. The DEA model we use can be distinguished by the scale and orientation of the model. If one cannot assume that economies of scale do not change as the size of the service facility increases, then a variable-returns-to-scale (VRS) type of DEA model, the one selected here, is an appropriate choice (as opposed to a constant-returns-to-scale, (CRS) model). Furthermore, if in order to achieve better efficiency, managers' priorities are to adjust their inputs (before outputs), then an input-oriented DEA model rather than an output-oriented model is appropriate.

The way in which the DEA program computes efficiency scores can be explained briefly using mathematical notation (adapted from [20]).

The VRS envelopment formulation is expressed as follows:

$$\mathrm{VRS}_p(Y_l, X_l, \mathbf{u}^l, \mathbf{v}^l): \ \min -(\mathbf{u}^l s + \mathbf{v}^l e)$$

$$\mathbf{Y}\lambda - s = Y_l$$

$$-\mathbf{X}\lambda - e = -X_l$$

$$1\lambda = 1$$

$$\lambda \geq 0,\ e \geq 0,\ s \geq 0$$

For decision making unit 1, $\mathbf{x}_{il} \geq \mathbf{0}$ denotes the $\mathbf{i}^{\text{th}}$ input value, and $\mathbf{y}_{rl} \geq \mathbf{0}$ denotes the $\mathbf{r}^{\text{th}}$ output value. $\mathbf{X}_l$ and $\mathbf{Y}_l$ denote, respectively, the vectors of input and output values. Units that lie on (determine) the surface are deemed *efficient* in DEA terminology. Units that do not lie on the surface are termed *inefficient*. Optimal values of variables for decision making unit 1 are denoted by the s-vector $\mathbf{s}^l$, the m-vector $\mathbf{e}^l$, and the n-vector $\boldsymbol{\lambda}^l$.

Although DEA is a powerful optimization technique that can assess the performance of each CMHC, it has certain limitations. When one has to deal with large numbers of inputs and outputs in service production, and a small number of organizations are under evaluation, the discriminatory power of the DEA is limited. However, analysts can overcome this limitation by including only those factors (input and output) that provide the essential components of service production, thus avoiding distortion of the DEA results. This is usually done by eliminating one of a pair of factors that are strongly positively correlated with each other.

In the majority of studies using DEA, the data are analyzed cross-sectionally, with each decision making unit (DMU) – in this case the CMHC – being observed only once. Nevertheless, data on DMUs are often available over multiple time periods. In such cases, it is possible to perform DEA over time, where each DMU in each time period is treated as if it were a distinct DMU. This DEA technique is called window analysis [21]. Using window analysis, one can examine changes in efficiency over time. A DMU's performance in an initial time period is compared to its performance in later time periods, and compared as well to the performance of the other DMUs. We employed window analysis to assess the changes in CMHC efficiency over time [22].

7.3 METHODS

7.3.1 Data and data sources

The primary source of data was the Department of Medical Assistance Services (DMAS) of Virginia. DMAS has extensive claims files that are made available for research purposes. This database records dates of services for each claim, and its diagnosis, procedure, and patient profile. Medicaid data come in three files: claims, recipients, and providers. The claims data set includes dates of services for each claim, its diagnosis, procedure, and patient profile. The recipient data set contains eligibility information on recipients of Medicaid. The provider data set contains the provider's location, practice type, and specialties. Data for five consecutive years of fee-for-service care (calendar years of 1994 through 1998), including two years of managed care encounters (1997 and 1998) were used.

Data were provided by the Virginia Medicaid agency with the cautions that managed care data have not been evaluated for reliability and validity and that there are known data quality concerns. All variables from the encounter data set are considered to have quality limitations, which we will point out and which should be considered in interpreting preliminary findings.

7.3.2 Sample selection and analysis plan

Only patients with Serious Mental Illness (SMI) were included in the study. Patients were identified as SMI patients using ICD-CM-9 diagnosis codes in the range of 295.00 – 298.99 (schizophrenia, major affective psychosis, paranoid states, and other non-organic psychoses). The claims of SMI recipients were merged and aggregated to the unit level of the community mental health center (CMHC), also known as a Community Services Board (CSB) in Virginia. To ensure experience and consistency in services for SMI patients, we examined 12 CMHCs (the providers) that had treated 100 or more claims of SMI patients in case management services in 1994. Those 12 CMHCs also were examined in the next four consecutive years (1995, 1996, 1997, and 1998). A total of 60 (12 × 5 years) CMHCs were thus included in the analysis and constituted the unit of analysis.

7.3.3 Variables

We included outcome and resource measures for community mental health centers derived from the DMAS database. These measures comprise two output and six input variables. Output variables are: the number of Medicaid SMI patients in the Medicaid eligibility category of supplemental security income (SSI), and the number of patients in all Medicaid eligibility categories except SSI. This categorization of outputs is a proxy for outputs "more severe" and "less severe", and hence serves as the case-mix difference for outputs. Inputs that we included (measured by number of claims in which these services appear) are as follows:

- use of non-emergency crisis support in CMHC;
- use of outpatient assessment;
- use of outpatient therapy;
- use of outpatient medication management;
- use of clubhouse;
- use of case management.

The above services are those most frequently provided by the CMHCs. Three of the services remained fee-for-service throughout the five-year time frame we considered, in both Richmond and Tidewater. Certain services – the use of crisis support, clubhouse, and case management – were part of a special program, the State Plan Option program, which was available in both settings with fee-for-service reimbursement. State Plan Option services support successful community care and residence. Non-emergency crisis support offered by CMHCs includes their crisis intervention services in the community, with the goal of stabilizing the client and allowing him or her to remain in the community. The clubhouse service is a psychosocial rehabilitation program that provides a supportive environment and promotes independent living in the community. The case management services include coordination and integration of care and services for the client.

In the Tidewater area, three services were covered by the capitated managed care plan in the last three years of our study – outpatient assessment, outpatient therapy, and medication management.

The remaining services, which changed from fee-for-service to managed care in the Tidewater area, include outpatient assessment, therapy, and medication management. These services are limited to those provided by CMHCs as the billing providers. By definition, these are outpatient providers. Assessment is defined as a psychiatric diagnostic interview (procedure code 90801). Therapy includes individual therapy (excluding un-timed billings), family therapy, and group therapy. Medication management includes prescriptions and evaluation of medication needs. It does not include administration of medications.

Only services provided by the CMHCs and paid for by Medicaid are included in the analysis. CMHCs may also provide services to clients that are not covered by Medicaid; these are not included.

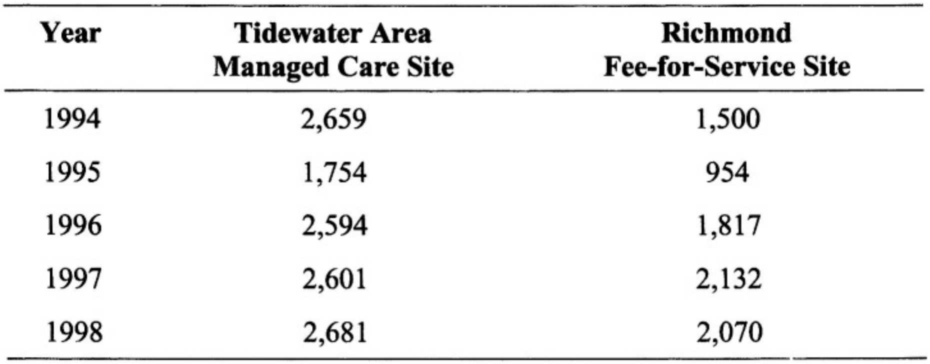
7.4 RESULTS

7.4.1 Trends for SMI patients

Table 7.1 shows the number of SMI claims in the study localities, by years. The percentage of the sample comprising SMI Medicaid recipients rose steadily from 9% in 1994 to 13% in 1998. Medicaid managed care for behavioral care was implemented in 1996 at the Tidewater site. Thus, it is prudent to examine the descriptive statistics for pre- and post-managed care in the experimental (Tidewater) and control (Richmond) groups of CMHCs.

Table 7.2 provides descriptive statistics for all output and input variables before and after managed care at both sites. For each variable, the table shows both its mean (first row for each variable) and standard deviation (second row). There is a notable difference in the volume of outputs from pre-managed care to post-managed care in Richmond and in the Tidewater areas. Furthermore, average output in the Tidewater area generally is higher than that in Richmond. On the input side, there are varying practice patterns between the two areas. However, these are not statistically significant, with the exception of outpatient assessment.

Table 7.1 Trends for Serious Mental Illness (SMI) by site: Number of claims by year*



Year	Tidewater Area Managed Care Site	Richmond Fee-for-Service Site
1994	2,659	1,500
1995	1,754	954
1996	2,594	1,817
1997	2,601	2,132
1998	2,681	2,070

* The 1995 data had missing recipient and claims data, so the reported figures are an undercount. Data for 1994, 1995, and 1996 include only fee-for-service claims. Data for 1997 and 1998 include fee-for-service and managed care claims.

7.4.2 Windows analysis – Efficiency results

Evaluations were performed using the data from 1994 through 1998. The windows analysis method of DEA developed by Charnes and colleagues [21] was employed.

Efficiency results are presented in Table 7.3. Windows of five years for the 12 CMHCs show that 31 occurrences out of 60 are classified as efficient. Since Richmond had four CMHCs, and Tidewater had eight CMHCs in this study, during the five-year window only seven efficiency results (out of 20) were observed from Richmond CMHCs. The average efficiency score for Richmond CMHCs was 0.753.

On the other hand, Tidewater CMHCs displayed much higher efficiencies, with an average score of 0.895, with 24 occurrences (out of 40) observed during the same five-year window.

The five-year trend of efficiency scores (shown later in Table 7.6) displays a generally increasing trend for the Tidewater CMHCs, but stagnation, even retrenchment in the Richmond CMHCs. Table 7.4 compares efficiency scores before and after managed care implementation in both localities, and shows that efficiency at the Tidewater CMHCs is significantly higher after managed care than before.

Table 7.2 Descriptive statistics before and after managed care*

	Before Managed Care (1994 and 1995)		After Managed Care (1997 and 1998)	
Variables	**Richmond N=8**	**Tidewater N=16**	**Richmond N=8**	**Tidewater N=16**
Outputs				
Numbers of Medicaid SMI patients treated in the Medicaid SSI eligibility category**	185.38 (171.23)	266.56 (175.73)	281.88 (276.61)	428.75 (301.22)
Numbers of patients treated in all Medicaid eligibility categories except SSI**	30.50 (15.84)	50.88*** (39.95)	84.00 (74.29)	98.25 (86.49)
Inputs				
Non-emergency crisis support	111.95 (137.46)	172.75 (161.47)	224.88 (312.92)	199.38 (166.83)
Outpatient assessment	16.55 (11.95)	42.94*** (51.46)	282.13 (526.20)	67.19 (55.74)
Outpatient therapy	241.75 (106.70)	349.13 (450.74)	439.13 (391.30)	306.88 (233.66)
Outpatient medication management	724.13 (775.70)	689.88 (695.96)	715.13 (622.89)	820.50 (524.11)
Clubhouse	1008.25 (1187.40)	1108.18 (1450.79)	2388.25 (3413.87)	1020.88 (1062.21)
Case management	1309.00 (1424.11)	1351.25 (1206.53)	1930.25 (1616.21)	1710.75 (1423.81)

* N=48; Mean numbers are shown in the first row for each entry; the entry in the second row (shown in parentheses) is the standard deviation.

** SSI = supplemental security income

*** $P < 0.01$

Table 7.3 Efficiency results*

	Richmond	Tidewater
Number of results		
Efficient	7	24
Inefficient	13	16
Total	20	40
Efficiency Score		
Efficients included	0.753 (0.21)	0.895 (0.16)
Efficients excluded	0.619 (0.13)	0.737 (0.15)

*N=60; numbers not in parentheses represent mean values; numbers shown in parentheses are standard deviations.

Table 7.4 Technical efficiency score differences between Richmond and Tidewater areas*

	Richmond (N=8)	Tidewater (N=16)	t-Value for Mean
Before Managed Care (1994 and 1995)	0.768 (0.20)	0.865 (0.17)	1.26**
After Managed Care (1997 and 1998)	0.748 (0.21)	0.938 (0.14)	2.30***

*N=60; numbers not in parentheses represent mean values; numbers shown in parentheses are standard deviations.
** P < 0.01
*** P < 0.05

7.4.3 Inefficiency score differences between Richmond and Tidewater areas

The sources of inefficiency are investigated and depicted in Table 7.5 for pre- and post-managed care in both localities. Tidewater CMHCs generally reduced inefficiencies after the implementation of managed care, as reflected by their efficiency scores. To do so, CMHCs must increase their outputs while reducing their inputs, or reduce their inputs while keeping the outputs steady. The majority of Tidewater CMHCs achieved this goal, but not completely. There is a room for further input reduction for the inefficient Tidewater and Richmond CMHCs. For example, after the implementation

Table 7.5 Inefficiency score differences before and after managed care*

	Before Managed Care (1994 and 1995)		After Managed Care (1997 and 1998)	
Input Variables	**Richmond N=5**	**Tidewater N=8**	**Richmond N=5**	**Tidewater N=5**
Non-emergency crisis support	40.10 (65.54)	36.19 (56.88)	105.54 (136.89)	19.32 (53.00)
Outpatient assessment**	4.85 (6.05)	16.97 (36.45)	230.90 (505.27)	4.03 (7.54)
Outpatient therapy**	142.27 (146.91)	49.17 (65.48)	181.59 (319.09)	60.27 (112.54)
Outpatient medication management**	325.14 (562.12)	221.56 (377.27)	184.59 (245.62)	100.27 (272.72)
Clubhouse	359.09 (796.00)	342.60 (1041.54)	1745.43 (3156.75)	333.10 (876.91)
Case management	467.42 (808.65)	204.42 (273.70)	542.87 (671.77)	111.53 (270.36)

*Numbers not in parentheses represent mean values; numbers shown in parentheses are standard deviations.

** For 1997 and 1998 Tidewater values, the encounter data set used as the source of data has quality issues; caution should be used in interpreting these results.

of managed care, inefficient Tidewater CMHCs (rightmost column of Table 7.5) are using 19 more units of crisis support for non-emergency care, 4 more outpatient assessments, 60 more instances of outpatient therapy, 273 more instances of outpatient medication management, 877 more clubhouse arrangements, and 112 more instances of case management, than their efficient counterparts do. In other words, other CMHCs with similar profiles use much fewer resources to provide similar outputs than the inefficient CMHCs do. The magnitude of the inefficiencies and the improvement needed for Richmond CMHCs are more even more dramatic than is the case for the inefficient Tidewater CMHCs.

7.4.4 Cone ratio model – Weight restrictions and practice patterns

DEA can also be used to direct provider behavior toward those practice styles found to be not only effective but also cost efficient. This can be accomplished either by utilizing a weight-restricted DEA model [11, 23], or by calculating preferred ratio constraints and restricting each CMHC's ratio to a particular level. Weight-restricted DEA calculations limit the use of certain virtual inputs and outputs, thereby creating efficiency scores relative to the frontier defined in the preferred efficient practice style.

Figure 7.1 is a conceptualization of a model with two inputs – use of case management and use of non-emergency crisis support – and one output – Medicaid SMI patients with SSI eligibility. - In the example illustrated in Figure 7.1, there are 12 CMHCs and three practice styles. Practice Style 3 can be defined as a case-management oriented model. Here case management is designated as the preferred type of treatment management, i.e. preferred over more expensive treatments. The ratio constraints can be defined as case management over non-emergency crisis support; outpatient medication management over outpatient therapy. When restricted by these preferred ratio constraints, the efficiency frontier includes only that section creating the preferred practice style.

To create the preferred ratio constraints that are used to define the practice styles, DEA weights (also referred to as prices or multipliers) are utilized. The desired ratio(s) are calculated using the input weights from each CMHC. Then, for each ratio created, the minimum, first quartile, median, third quartile, and maximum values are calculated. These values illustrate the distribution of the ratio and give the researcher choices about the level at which to restrict the distribution. How much restriction is placed on a particular ratio depends on the distribution level selected (usually median or third quartile values are selected initially); in the current analysis, we used third quartile values. These newly restricted ratios can be plugged back into the DEA model and will restrict the use of those selected inputs needed to reach the efficiency frontier.

A graphic conceptualization of a weight restricted model using the two-inputs-one-output model is shown in Figure 7.2, where the area identified as "Care Management Cone" exemplifies a balanced approach for efficient management of mental health patients.

We analyzed two CMHC practice styles, as shown in Table 7.6. The first model (Base Model) contains no ratio constraints, and illustrates practice as usual. The second model is a cone ratio model, which uses weight restrictions. This model includes the preferred ratios: case management over non-emergency crisis support, and outpatient medication management over outpatient therapy.

Figure 7.1 DEA conceptualization of CMHC Practice Styles

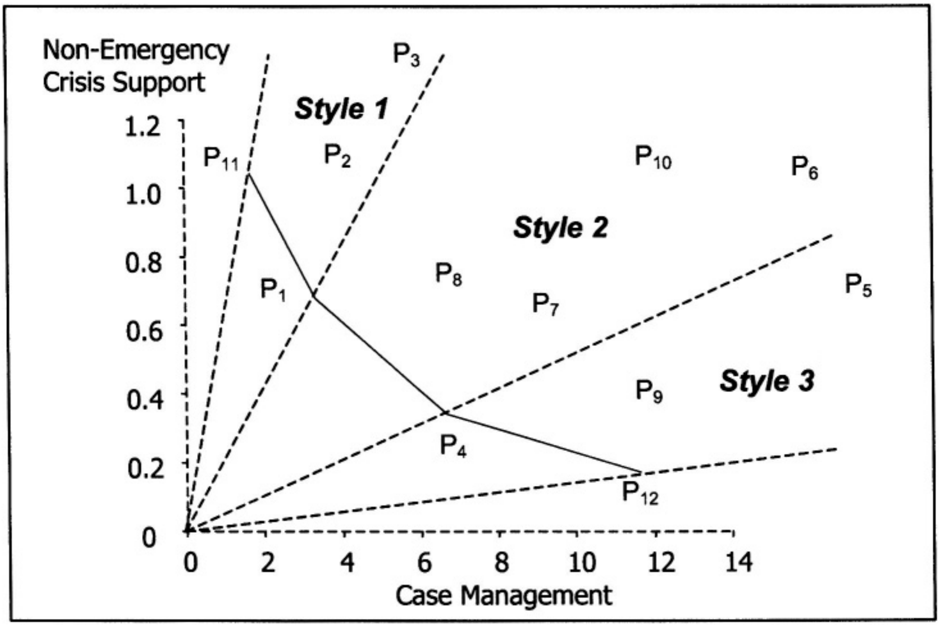


Figure 7.2 Conceptualization of super efficient CMHC model

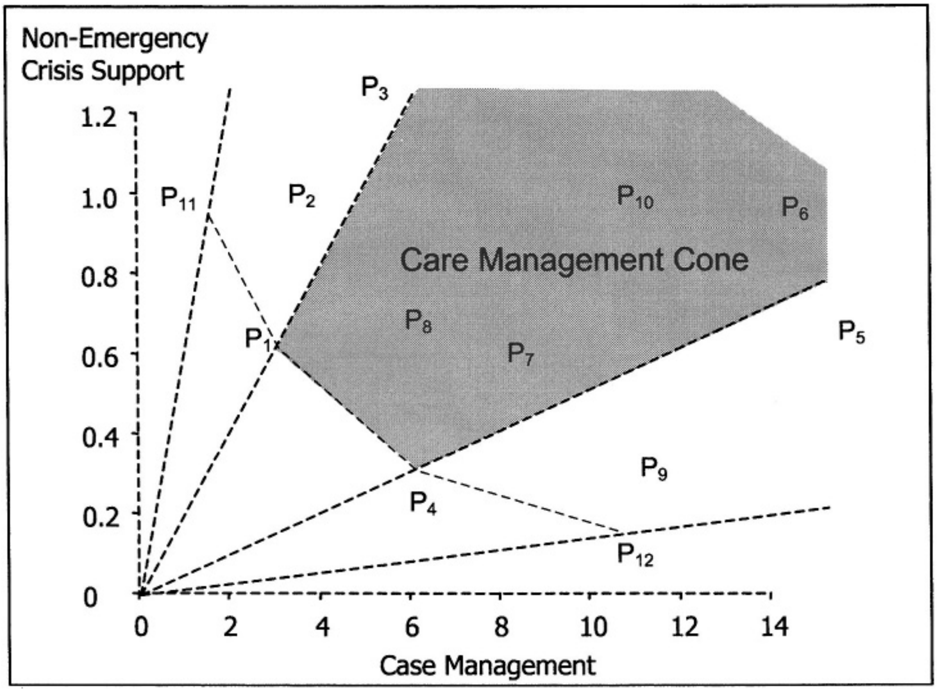
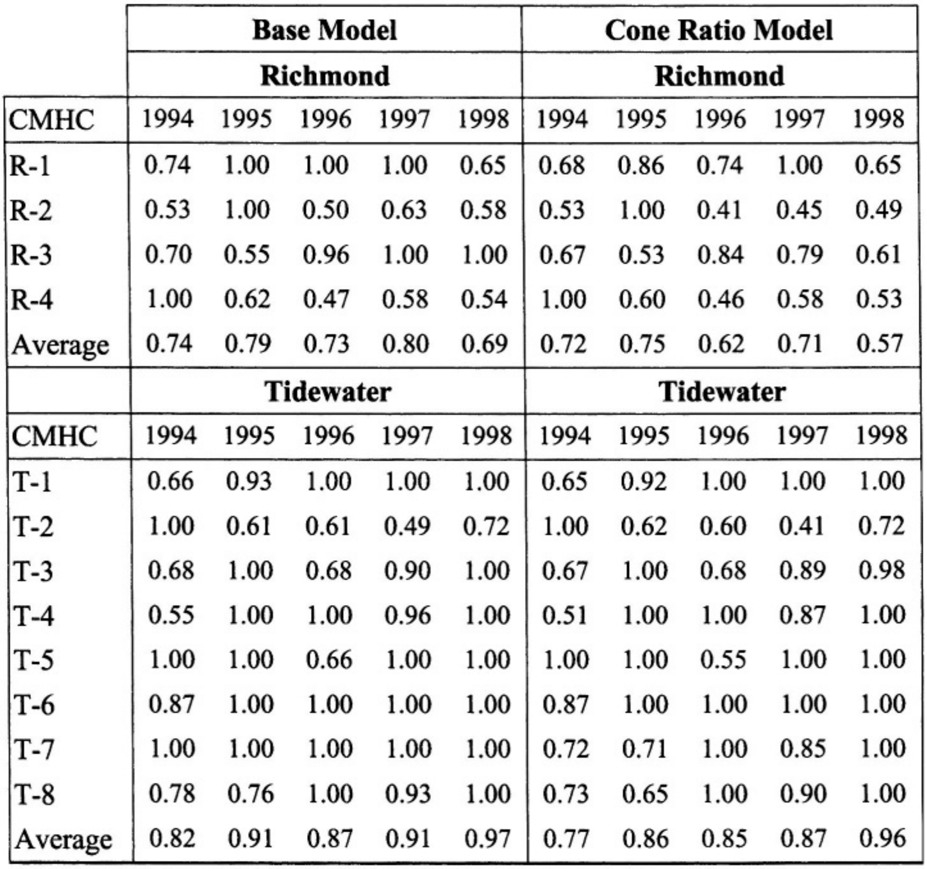


Table 7.6 Base- and weight-restricted model efficiency scores by year *



	Base Model					Cone Ratio Model				
	Richmond					Richmond				
CMHC	1994	1995	1996	1997	1998	1994	1995	1996	1997	1998
R-1	0.74	1.00	1.00	1.00	0.65	0.68	0.86	0.74	1.00	0.65
R-2	0.53	1.00	0.50	0.63	0.58	0.53	1.00	0.41	0.45	0.49
R-3	0.70	0.55	0.96	1.00	1.00	0.67	0.53	0.84	0.79	0.61
R-4	1.00	0.62	0.47	0.58	0.54	1.00	0.60	0.46	0.58	0.53
Average	0.74	0.79	0.73	0.80	0.69	0.72	0.75	0.62	0.71	0.57
	Tidewater					Tidewater				
CMHC	1994	1995	1996	1997	1998	1994	1995	1996	1997	1998
T-1	0.66	0.93	1.00	1.00	1.00	0.65	0.92	1.00	1.00	1.00
T-2	1.00	0.61	0.61	0.49	0.72	1.00	0.62	0.60	0.41	0.72
T-3	0.68	1.00	0.68	0.90	1.00	0.67	1.00	0.68	0.89	0.98
T-4	0.55	1.00	1.00	0.96	1.00	0.51	1.00	1.00	0.87	1.00
T-5	1.00	1.00	0.66	1.00	1.00	1.00	1.00	0.55	1.00	1.00
T-6	0.87	1.00	1.00	1.00	1.00	0.87	1.00	1.00	1.00	1.00
T-7	1.00	1.00	1.00	1.00	1.00	0.72	0.71	1.00	0.85	1.00
T-8	0.78	0.76	1.00	0.93	1.00	0.73	0.65	1.00	0.90	1.00
Average	0.82	0.91	0.87	0.91	0.97	0.77	0.86	0.85	0.87	0.96

*The Richmond CMHCs are denoted by R-1 through R-4; the Tidewater CMHCs are denoted by T-1 through T-8.

An input-oriented variable-returns-to-scale (VRS) DEA technique was employed in both models to identify the best set of practice patterns. An input-oriented model is preferred because CMHCs can change the number and type of inputs they use (relaxing the assumption of mandated services by the state government), but not the number of clients who visit them for treatment. Those CMHCs that did not exhibit efficient practice patterns were further analyzed and compared to their peers to inquire under what circumstances their practice patterns would mimic the best practice behavior.

Table 7.6 compares the results from the base and the cone ratio models. In the cone ratio model, the efficiency of CMHCs in Richmond is significantly less than in the base model. The Tidewater CMHCs' efficiency scores were reduced during the pre-managed care era; they were significantly higher after managed care. A perfect efficiency score is a score of 1.0. Richmond had only three perfectly efficient CMHCs in the cone ratio model, as compared to seven in the base model, yielding a 57.1% reduction in perfect efficiency. On the other hand, the number of instances of perfectly efficient CMHCs in Tidewater dropped to 20 in the cone ratio model from 24 in the base model, yielding only a 16.7% reduction. This shows the power of the cone ratio model, which produces more stringent efficiency outcomes.

7.5 DISCUSSION

Over the past decades, researchers have demonstrated differences in the amount of resources used for health care in this country due to varying patterns of care. The most probable cause for this variation is varying provider practice styles. There is a growing concern about the efficiency with which health care services are delivered, and about which of the varying practice styles are more efficient, and thus more appropriate. This chapter has described how we developed and applied a DEA methodology as a mechanism to identify the most efficient practice patterns for behavioral health care, and to evaluate the variations in resource use associated with different variations in practice. The data used in this chapter are useful for methodological development. However, the managed care data are from the early years of a new data system and there are known concerns regarding the data quality. Therefore the results should be used to understand the method, but should not be used to judge the actual efficiency of the organizations we analyzed.

Given that caution, several observations can be made. The efficiency score of providers in Tidewater increased after the implementation of managed care, but the efficiency score of providers in Richmond did not change. The differences in the efficiency score of the two regions are statistically significant after managed care was implemented. However, the differences in inefficiency scores between the two regions are not significant and are unchanged after the implementation of managed care. Despite the limitations of the data, our analysis demonstrated that providers practiced more efficiently – that is, they used fewer resources to produce similar outputs – under the managed care payment system. The two main areas that account for the efficiency differences between the two regions are the case management and non-emergency crisis support services.

In this study we were limited in the selection of input/output variables because of sparse data on certain service variables. Furthermore, there were some concerns about the quality of data for input variables with respect to outpatient assessment, outpatient therapy, and outpatient medication management.

We also recognize that the issue of the quality of care raises the question of the effectiveness of care by the CMHCs. We assumed that the quality of care is the same from all the providers. Further analysis is needed to identify how efficiency affects the quality of care.

References

[1] U.S. Department of Health and Human Services (1999). *Mental Health: A Report of the Surgeon General.* Substance Abuse and Mental Health Services Administration, Center for Mental Health Services, National Institute of Mental Health, Bethesda, MD.

[2] Lurie N., J.B. Christiansen, D.Z. Gray, W.G. Manning, Jr., and M.K. Popkin (1998). The effect of the Utah Prepaid Mental Health Plan on structure, process, and outcomes of care. In D. Mechanic (Ed.), *Managed Behavioral Health Care: Current Realities and Future Potential.* New Directions for Mental Health Services, No. 78 (pp. 99-106). Jossey- Bass, San Francisco, CA.

[3] Hausman. J.W., N. Wallace, and J.R. Bloom (1998). Managed mental health experience in Colorado. In D. Mechanic (Ed.), *Managed Behavioral Health Care: Current Realties and Future Potential.* New Directions for Mental Health Services, No. 78 (pp. 107-114). Jossey-Bass, San Francisco, CA.

[4] Dickey, B., E.C. Norton, S.T. Normand, H. Azeni, and W.H. Fisher (1998). Managed mental health experience in Massachusetts. In D. Mechanic (Ed.), *Managed Behavioral Health Care: Current Realties and Future Potential.* New Directions for Mental Health Services, No. 78 (pp. 115-122). Jossey-Bass , San Francisco, CA.

[5] Seiford, L.M. (1996). Data Envelopment Analysis: The evolution of the state of the art (1978-1995). *Journal of Productivity Analysis,* 7, 99-138.

[6] Sherman, H.D. (1984). Hospital efficiency measurement and evaluation: Empirical test of a new technique. *Medical Care,* 22, 922-938.

[7] Nunamaker, T. (1983). Measuring routine nursing service efficiency: A comparison of cost per patient day and Data Envelopment Analysis models. *Health Services Research,* 18, 183-205.

[8] Huang, Y.L. and C.P. McLaughlin (1989). Relative efficiency in rural primary health care: An application of Data Envelopment Analysis. *Health Services Research,* 24,143-158.

[9] Sexton, T.R., A.M. Leiken, A.H. Nolan, S. Liss, A. Hogan, and R.H. Silkman (1989). Evaluating managerial efficiency of Veterans

Administration Medical Centers using Data Envelopment Analysis. *Medical Care,* 27, 1175-1188.

[10] Chilingerian, J. and D.H. Sherman (1990). Managing physician efficiency and effectiveness in providing hospital services. *Health Services Management Research,* 3 , 3-15.

[11] Ozcan, Y.A. (1998). Evaluation of variation in physician practice behavior: DEA approach for the case of otitis media. *Health Care Management Science,* 1, 5-17.

[12] Ozcan, Y.A., H.J. Jiang, and C-W. Pai (2000). Physician efficiency in treatment of sinusitis: Do primary care physicians or specialists provide more efficient care? *Health Services Management Research,* 13, 90-96.

[13] Schinnar, A.P, E. Kamis-Gould, N. Delucia, and A.B. Rothbard (1990). Organizational determinants of efficiency and effectiveness in mental health partial care programs. *Health Services Research,* 25, 387-420.

[14] Tyler, L.H., Y.A. Ozcan, and S.E. Wogen (1995). Technical efficiency of community mental health centers. *Journal of Medical Systems,* 19, 413-423.

[15] Ozcan, Y.A., S.E. Wogen, and L.W. Mau (1998). Efficiency evaluation of skilled nursing facilities. *Journal of Medical Systems,* 22, 211-224.

[16] Ozcan, Y.A. and J.J. Cotter (1994). An assessment of area agencies on aging in Virginia through Data Envelopment Analysis. *Gerontologist,* 34, 363-370.

[17] Ozcan Y.A., R.D. Luke, and C. Haksever (1992). Ownership and organizational performance: A comparison of technical efficiency across hospital types. *Medical Care,* 30, 781-794.

[18] Ozcan, Y.A. and R.D. Luke (1993). A National study of the efficiency of hospitals in urban markets. *Health Services Research,* 28, 719-739.

[19] Byrnes, P. and V. Valdmanis V (1994). Analyzing technical and allocative efficiency of hospitals. in A. Charnes, W. W. Cooper, A. Lewin and L. Seiford, Eds. *Data Envelopment Analysis: Theory,*

Methodology, andApplication. Kluwer Academic Publishers, Boston, MA.

[20] Cooper, W.W., L.M. Seiford, and K. Tone K (2000). *Data Envelopment Analysis,* Kluwer Academic Publishers, Norwell, MA.

[21] Charnes, A., T. Clark, W. Cooper, and B. Golany (1985). A developmental study of Data Envelopment Analysis in measuring the efficiency of maintenance units in the U.S. Air Forces. *Annals of Operations Research,* 2,85-112.

[22] Harris, J., II, H. Ozgen, and Y.A. Ozcan (2000). Do mergers enhance the performance of hospital efficiency? *Journal of the Operational Research Society,* 51, 801-811.

[23] Pedraja-Chaparro, F., J. Salinas-Jimenez, and P. Smith (1997). On the role of weight restrictions in Data Envelopment Analysis. *Journal of ProductivityAnalysis,* 8, 215-230.

8 USING SIMULATION IN AN ACUTE-CARE HOSPITAL: EASIER SAID THAN DONE

Michael W. Carter[1] and John T. Blake[2]

[1] Department of Mechanical and Industrial Engineering
University of Toronto
Toronto, Ontario, Canada M5S 3G8

[2] Department of Industrial Engineering
Dalhousie University
Halifax, Nova Scotia, Canada B3J 2X4

SUMMARY

Simulation, as it is typically taught, is a rather mechanical process. Students are taught to follow a recipe: analyze a system, design a model, convert the model to computer code, collect data, verify, validate, and analyze the output. In practice, many analysts find that simulation is an odd combination of art, science, and marketing. Using this technique appropriately, in any industry, involves more than simply following the text book. In our experience, health care provides some rather unique challenges for the modeler. This chapter describes four different practical examples of using simulation to analyze a problem in an acute care hospital. The specific examples are not described in detail, since the applications have appeared in other publications. The emphasis here is to present some of the obstacles that were encountered and the lessons learned.

KEY WORDS

Simulation, Nursing human resources, Surgical schedules, Emergency department modeling, Drug order entry

8.1 INTRODUCTION

Simulation has a vast range of application in health care. Anyone who has ever visited a hospital emergency room, undergone surgery, or even visited their family doctor will recognize that the provision of health care is a complex, stochastic process with an overall structure analogous to a network of queues. The heterogeneity of customers in this system, the vast range of potential paths through the network, and the time-sensitivity of service make health care a “textbook” application for simulation.

The application of simulation in a health care setting is not always as simple and straight forward as one might think from reading the standard texts. In this chapter we present four simulation studies and describe lessons learned during the projects. The objective of this chapter is not to describe how to conduct a simulation study, or to provide all details for the four projects, since this material appears elsewhere in the literature. Our goal is to give analysts an idea of the issues that arise when an operations research technique is applied to a health care setting.

The projects described in this chapter include: a study to evaluate the link between inpatient census and the surgical schedule; a study to evaluate the causes of, and solutions to, emergency room wait time in a pediatrics hospital; a pharmacy ordering model; and a generalized simulation model for an acute care emergency department. In each instance, the problem is described, and an overview of the solution methodology is presented along with a summary of results. Each section concludes with a summary of the lessons learned during the project.

8.2 EVALUATING THE IMPACT OF THE ELECTIVE SURGERY SCHEDULE ON RESOURCE ALLOCATION

8.2.1 Description of the application

Nursing, like many regulated health care professions, tends to go through human resource availability cycles. The length of time required to fully train a doctor or a nurse (four years or longer in many jurisdictions) means that decisions made today regarding training spaces in universities and colleges only have a noticeable impact five to ten years later. Of course, in the period of time between when the plans are made and come to fruition, the demand for health care professionals may have changed. This is a common problem in almost all medical human resource planning.

Nursing, as a profession, has a number of unique characteristics that make human resource planning more difficult still. The profession is disproportionately female, and thus child rearing and family responsibilities have an impact on participation in the market place. As with other professions, general economic conditions, quality of work-life issues, and random fluctuations in the labor market also affect participation.

In 1989, Toronto experienced a shortage of qualified nurses. A good economy, combined with a rapidly rising housing market in the metropolitan Toronto area, caused a net outflow of nursing personnel from downtown to suburban institutions. To deal with this problem, nursing leaders from five urban Toronto hospitals collaborated to discuss possible ways to attract and retain nurses in their institutions. The number one nursing complaint, the amount of money paid to nursing staff, was not open to change. The second most important issue was the work week; in short, nurses wanted less weekend work.

One of the members of the committee facetiously suggested, "If we did surgeries on Monday for people with length of stay of four nights, and did only day surgery on Friday, we could empty the wards on the weekend, and give nurses more weekends off." The suggestion was clearly not practical, but the idea that we could change the surgical schedule to reduce the weekend ward census was thought to be interesting. A project was subsequently funded by a grant from the Ontario Ministry of Health and five Toronto teaching hospitals: Toronto Hospital for Sick Children, Toronto General Hospital, Sunnybrook Health Science Centre, Mount Sinai Hospital and Toronto Western Hospital. (Our co-investigator was Professor Linda O'Brien-Pallas, Faculty of Nursing, University of Toronto.)

The study lasted for two years in 1991-93 and involved developing a simulation model to use as a decision support tool [1, 2]. The model included the operating rooms, the recovery room, intensive care units and regular inpatient wards. We were primarily interested in surgical patients since 90% of all surgical patients were scheduled in advance and, therefore, were somewhat controllable. Conversely, it was felt that nothing could be done to control medical patients, since 90% of all medical patients were emergency admissions. In all of the hospitals in the study, operating room time was assigned on a "block booking" basis. Surgeons received blocks of operating room (OR) time (e.g., every Monday morning for three hours) and were free to schedule patients in any order within their assigned blocks. Typically, elective surgery took place Monday to Friday on the day shift with one or two rooms available nights and weekends for emergencies.

Given this arrangement, we concluded that by changing the weekly OR schedule, we could influence workload and census in the rest of the hospital. By extension, we argued, it should be possible to determine a schedule that would be optimal from a staffing perspective. Furthermore, because we did not anticipate making any changes to the number or length of assigned blocks, we assumed that our schedule would have no impact on patients or their clinical care. The only impact, as far as we could tell ahead of time, would be minor inconvenience to surgeons who might have their block time rearranged within the master surgical schedule.

With these assumptions in mind, we built a simulation model, a database and user interface for the simulation. The model included all scheduled surgical patients and allowed for emergency patients who could preempt elective surgery as well as medical patients competing for intensive care unit (ICU) beds. The database included an underlying nursing workload model that estimated the total hours in each ward given the patient mix and volume flowing through it. If there were no beds available or not enough nursing hours when a patient was to be admitted, elective surgery would be canceled. The model generally used a first-come-first-served logic for allocating scarce resources. A small percentage of patients were also canceled for other reasons.

We used a two-pronged approach to collecting data for the model. We spent several months in each hospital analyzing the process to understand how patients flowed through the facility, creating process flow charts and collecting unique site-specific data. We also took advantage of an existing database of discharge records, The Canadian Institute of Health Information (CIHI). CIHI is a third-party organization that stores a discharge summary of every patient admitted to a hospital in Canada. Institutions in Canada are required to contribute data to this source, which is used by hospitals for their own internal review as well as by federal and provincial authorities.

Through the interface, the user was able to set the surgical schedule, make adjustments to the surgeons' case mix, specify the number of beds and nurses on each ward and change a variety of control parameters. The simulation itself was driven on a data trace. Because of the often confounding factors relating to age, gender, disease, co-morbidity, treatment, and outcome, we reasoned that it would be more practical to dispense with the idea of developing and fitting distributions for key simulation variables such as length of stay, processing time, etc, since we could not assume independent and identically distributed observations. Instead, we decided to sample directly from a large list of patients available from hospital discharge records. Thus when we needed to "create" a patient in our model, we randomly selected a person from this existing list and simply associated all of that patient's demographic, treatment, and outcome data with the simulated patient. This mechanism, we felt, would make the simulation easy to port between sites and easy to validate.

After running random patients through the model for a two-week warm up period, we ran ten replications of two weeks. Upon the completion of the run, we produced summary statistics on estimated annual patient volumes, cancellations, emergencies, and patient census and nursing hours in each ward by day of week.

8.2.2 Challenges encountered

Timing/project cycle time Simulations typically look simple to build; or at least they look simple at the start of a project. Our project was originally designed for a two-year cycle. The pilot model took approximately 12 months to complete. Ports to other institutions, which were scheduled to take two months, took about four months apiece to complete. Thus, by the end of the project, more than four years had elapsed since its inception in 1989. The reality of 1993 was much different than the reality of 1989, particularly in the health care sector in Canada. While 1989 was the high point in an economic cycle, 1993 was a low point. Thus, by the time our program was ready, the government was cutting health care budgets, and hospitals were laying off nurses! Simply put, weekend workload and quality of work-life issues had dropped off the radar screen; people were much more interested in holding onto their jobs than getting the weekend off.

Fortunately, this unexpected turn of events did not detract from the value of the project. The model turned out to be very useful as a mechanism to balance the use of increasingly tight hospital resources. The simulation allowed users to experiment with various allocations of OR time and forecast the impact of ward census, nursing workload, ICU beds and recovery rooms. Several of the sites used our model to improve their operations.

For example, at the Toronto Hospital for Sick Children, in one ward, the census was double on Wednesday night compared to every other day of the week. By making a few minor adjustments, we were able to suggest an OR schedule that would balance the nursing workload over weekdays. As another example, we used the model to look at Christmas closing in 1994 after the Ontario government asked hospitals to close all elective surgery for two weeks as a cost reduction measure. Mount Sinai Hospital asked us to complete an analysis of residual demand for OR time and ward space due to emergency patients. We used the model to predict the staffing levels that would be needed to cover this demand for the two weeks. At Sunnybrook Health Sciences Centre the model was used in a number of planning scenarios, not to balance nursing workload, but to calculate production limits for their cardiovascular surgery program.

At one institution we ran into a problem validating our model. The simulation model suggested that the OR time currently allocated to the Ear, Nose and Throat (ENT) service could accommodate almost twice as many patients as they were actually serving. We searched for the cause of the discrepancy for several days in the simulation. Ultimately, we discovered that ENT had a habit of not always using all of their allocated time. Meanwhile, General Surgery was starving for OR time. Whenever ENT did not need the allocation, someone in General Surgery was happy to use it. The OR managers had not noticed the problem since all of the booked rooms were being fully utilized.

Data collection In any simulation, data collection, verification, and validation are major issues. In our experience in health care, no one ever had the right data in the form that we needed it. Health care information systems are typically designed to meet clinical requirements, not administrative needs. The CIHI data was a mixed blessing for our project. The CIHI data was universally available for all institutions, in a standard format, and from a single source, and was thus easy to access and import into the simulation. We did, however, find a number of weaknesses in the database which limited its applicability for a simulation study.

Since the discharge summary is only a summary of what happened to a patient, it was not always possible to entirely reconstruct a patient's process through the hospital from their discharge report. For example, a patient admitted as a medical patient for treatment of diabetes falls and breaks a hip during her hospitalization. If, at discharge the broken hip is considered to have contributed more to the patient's length of stay than the diabetes, the patient may then have been labeled as a surgical patient. Without complete access to a patient's record, reconstructing a patient's length of stay often involved some assumptions and some estimation.

Furthermore, we found that source data sent to CIHI was not always viewed by the institutions as reliable. (This is rather surprising given that the institutions themselves are responsible for abstracting and summarizing the data that is forwarded to CIHI.) Finally, the lag between when data was collected, abstracted, and made available to CIHI meant that we typically had to use patient abstracts that were at least a year old (and in one instance two years old) in the model. This led to a common complaint among potential users that the data was "too old" and "not representative of what we're doing now".

Every hospital was different Our model was designed to be flexible and to provide the ability to answer a wide variety of questions. We wanted to be able to test potential length-of-stay variations by age, disease, sex, etc. Furthermore, we wanted the model "portability process" to be as simple as possible. Our intention was to develop one generic model and simply move this model from place to place, plugging in new patient records and a small amount of site-specific data (i.e., the number of wards and the beds on each). In practice we found it very difficult to create a single, generic, general-purpose patient simulation. Each institution had a unique combination of services, programs, and unique "quirks" that made it difficult to directly move a model from one location to another. These quirks ranged from unique processing rules to arcane details of the physical plant.

For example, at The Toronto Hospital for Sick Children, the managers suspected that an old bank of elevators that frequently broke down significantly impacted transportation time! In this case, we included the elevator in our model.

When we initially developed the pilot simulation at The Toronto Hospital for Sick Children, we decided to restrict the model to patients who had only one surgical procedure. The number of cases of multiple surgeries there was quite small. However, Sunnybrook is a regional trauma center, and multiple procedures are relatively common. So, we needed to modify the Sunnybrook model to allow for multiple surgical procedures.

Stakeholders Getting the buy-in of all stakeholders is always a key component to any simulation project. However, when working in a health care setting, acceptance by all stakeholder groups is especially important. In this particular project, our assumption that the schedule rearrangement would be a minor issue turned out to be incorrect. Physicians, as a rule, control the creation of the master surgical schedule and guard it jealously. Schedule changes are almost never thought to be a matter of minor inconvenience.

In fairness to physicians, the schedule dictates both their income and their work schedule. That is why, in practice, the issue is so controversial that in most of the hospitals that we have worked in, the administration simply allocates total O.R. time to each service (cardiology, general surgery, orthopedics, etc.). The doctors in each service then decide among themselves how to allocate specific blocks of time. This solves some of their political issues but, as a consequence, the administration relinquishes any control over daily work flow balance.

8.3 CHILDREN'S HOSPITAL OF EASTERN ONTARIO (CHEO)

8.3.1 Description of the application

CHEO is a pediatrics teaching hospital affiliated with the University of Ottawa. In 1993, the hospital's Emergency Department expressed concern that up to 20% of patients were forced to wait at least two hours before being seen by a physician. The issue was one of quality of service rather than quality of care, since all patients are triaged promptly and urgent cases are seen right away. Long waits are generally associated with patients having "runny noses" and other minor complaints. However, with provincial budget cuts looming, managers at CHEO felt it important to maintain good public relations.

The Vice President of Ambulatory Care (VPAC) called us in May 1993. She had received eleven process improvement suggestions from staff members. Suggestions ranged from overhauling patient flow to making changes to the physical layout of treatment rooms. One suggestion called for installing video games in the waiting room so patients would not realize how long the wait was. While the VPAC thought that many of the suggestions were interesting, she needed a mechanism to provide quantitative analysis of the options.

To determine the impact of the various strategies on patient wait time a full-scale patient simulation model was developed [3]. The model included all of the major patient processes in the emergency department (ED): patient arrival, registration, triage, assessment, testing, treatment, and admission or discharge processes. Our main evaluation criteria were the average wait time and the distribution of these times for each of the four triage categories defined by the hospital: emergency, urgent, deferrable, and medical walk in.

In terms of modeling effort, the simulation itself was relatively simple. However, data collection, model validation, and output analysis required significant effort. One of the first things we discovered when we started collecting data at the hospital was the highly fractured nature of work in the ED. CHEO is a teaching hospital. The ED was staffed by one to three physicians, called Casualty Officers (COs), five to seven nurses, and a number of residents. Normally, each patient was seen by a nurse, a resident, and the CO who reviewed the resident's assessment. On any given shift there were ten patient treatment rooms available for use. Patients in these rooms were under the care of one CO who might also have had responsibility for providing medical education to one or two medical residents.

We noted immediately that it was extremely rare for any worker (physician, nurse, or resident) to complete a work cycle on one patient from start to end with no interruptions. More commonly, we observed that nursing and physician services were delivered in small discrete batches spaced over a fairly long time period. For example, a physician might assess a patient and order a test. A nurse might then collect a sample or transport the patient to another area. During the time the test was being performed, the physician would move on to treat other patients. When the results of the test became available, the physician would read the results, interpret them, and order a treatment or send the patient home.

Since a physician had five to ten patients "on the go" at any time, work cycles became quite fractured. Indeed, physician work cycles were nothing short of chaotic given the additional requirement of also providing medical education for students and residents. The physician was, for example, required to confirm the resident's diagnosis, provide him or her with background about a disease state or treatment option, and then confirm test results, treatment opinions, or patient instructions. Since the casualty officer was legally responsible for the patient's care, no part of the treatment process could occur without the permission of the CO. In fact, we later found that COs spent about as much time interacting with residents as with patients.

Once we had the model working and validated, we started a designed experiment. The factors that we varied included the number of COs on shift, the number of residents on shift, and the queuing discipline used to select patients from the waiting list. We did not find much of a factor effect for queue discipline, but we did note a strong negative effect for the number of COs on shift and a strong positive effect for the number of residents on shift.

As described earlier, resident education was a major component of work in the ED. Our experiment indicated that the work created by resident education was so great that eliminating all residents from the ED would substantially reduce patient waiting time! In fact, our model indicated that adding one additional CO, or eliminating all residents, would result in approximately the same improvement in waiting time.

Obviously, eliminating residents from a teaching hospital is not a practical alternative, but the results indicated that waiting time could be impacted by a number of different scenarios, including different numbers of physicians, different shift schedules, and/or the addition of a hospital "walk-in clinic" to treat patients with minor injuries.

These scenarios led us to one of our more interesting results. As part of our plan to rearrange physician schedules, we prepared a simple plot of patient arrival times for each day of the week. We compared this to the COs' shift schedule. We found that demand peak (patients) often occurred several hours before the staffing peak. For example, on Sundays, the peak patient arrival period was between 10 a.m. and 1 p.m., but the peak staffing levels were scheduled for 5 p.m. to 7 p.m. Needless to say, the wait times for patients arriving in the afternoon were extremely long because a queue had been building all day. We were able to make significant improvements simply by staggering the doctors' start times.

Other major recommendations that came from this project included adding an additional four hours of CO time daily to the main ED and implementing a fast-track clinic for low-acuity patients. We estimated that these improvements would reduce patient wait time by as much as 20%. Although the approval process took over a year, the hospital did eventually hire a new casualty officer due, in large part, to our analysis.

8.3.2 Challenges encountered

Data collection The fractured nature of work in the ED presented a data collection problem for us. While good theoretical and practical models of nursing workload are available, no corresponding workload standards exist for physicians. As a result, it was very difficult to determine, for example, the demand for physician time resulting from a patient presenting symptoms of asthma.

Furthermore, the highly fractured nature of work cycles made manual data collection a difficult task. For example, much of the work a physician performed on a patient's file was done when the physician was distant from the patient (e.g. reading x-rays, interpreting test results, discussing with nurses or residents). Thus, measuring physician contact time was not an entirely accurate method of determining workload.

"Job shadowing" also presented some difficulties. For example, the nature of patient confidentiality precluded an observer from direct access to many types of patient-physician encounters. All in all, identifying accurate physician workload was a difficult task. We were, however, able to satisfy our data requirements through a combination of statistical work sampling and job shadowing. One of the project team members undertook the work sampling procedure, which could be performed without the observer necessarily having to be in the vicinity of the patient and the physician. In addition, the hospital provided us with two nurse instructors, who performed a physician job shadow. As clinicians, both physicians and patients accepted the nurses. In the end, we were able to build a reasonable data sample using the two techniques.

In this application as well as most of the others, we discovered that the length of time required for any particular task is extremely variable. When things get busy in the ED, everyone tends to work a little faster. In particular, the casualty officers spend much less time teaching as the demand increases. This is not surprising, but it creates some serious modeling challenges. One way to avoid this issue, as we did, was to use process times based on data collected during the busy times. Our real objective in this study revolved around queue length during busy times. As a result, our simulated patients were treated faster than the real patients during relatively quiet times.

Time frame A key challenge we faced with this project was finding the time to collect data, build the model, and run a reasonable set of scenarios. While the project originally was envisioned to be a short term two-week project, in the end we spent almost a year working on the model and its various components. Building the actual simulation model, as it turned out, was not particularly difficult or time-consuming. In fact, it took us about two weeks to build. The time consuming aspect of the project was data collection. To complete data collection, it was necessary to: identify the data necessary to run the simulation, make appropriate simplifying assumptions, define the method by which this data should be collected, assign personnel to data collection, and then collect the data.

Once the model was up and running, we found it was not possible to simply complete a set of runs, write up the results, and put the project behind us. Management at the hospital viewed the model as a useful planning tool. As the planning process developed at the hospital, we were asked to run the model under different assumptions and scenarios. Coincidentally, as we developed and ran these scenarios, our understanding of the ED process increased and we were able to point out to management results we felt were interesting. This resulted in a collaborative arrangement between management and modelers which, while fruitful, extended the project completion date.

8.4 MODELING THE DRUG ORDER ENTRY PROCESS FOR INPATIENTS

8.4.1 Description of the application

Currently, in the vast majority of hospitals in North America, doctors still prescribe medications for hospital inpatients by scribbling notes on paper. In one study published in 1998, Ash, Gorman and Hersh [4] found that fewer than 2% of U.S. hospitals had Computerized Physician Order Entry (CPOE) completely or partially available and required its use by physicians. The initial cost of implementing CPOE is one major obstacle for hospitals. At Brigham and Women's Hospital, the cost of developing and implementing CPOE was approximately $1.9 million, with $500,000 in maintenance costs per year. Installation of even "off the shelf" CPOE packages requires a significant amount of customization for each hospital and can be very expensive [5]. Finally, there may be cultural obstacles to CPOE implementation. For example, many physicians resist the idea of ordering prescriptions via computer instead of by hand. Although summary results were not available, the Leapfrog Group hospital survey [6] indicated that most U.S. hospitals are in the process of implementing CPOE.

On the surface, the manual Medication Administration Process appears quite simple. The physician writes a prescription on paper at the bedside and puts the order in the patient's chart. The nurse retrieves the order, transcribes it onto the "Medication Administration Record" (MAR) and leaves a copy of the order in a tray in the ward to be picked up by pharmacy technicians at routine times throughout the day. A pharmacist reviews the order and transcribes it into a computer with access to electronic patient records and decision support capability. The order is prepared in the pharmacy and delivered to the ward. The nurse checks drugs against the MAR and administers to the patient. The nurse records the administration on the MAR.

What is wrong with this picture? The doctor relies on memory/knowledge to determine the dose of the medication, to think of patient allergies and to remember possible drug interactions. The nurse may not know that an order has been written or that the drug has arrived. The multiple transcriptions increase the possibility of error and are not value-added work. The physical transport of the order wastes time. If the nurse cannot read the order, s/he must check with the doctor. If the pharmacist has any questions about medication or dosage, s/he must page the nurse and/or the doctor and hold the prescription until the order is confirmed. We believe that the process could be greatly improved if the doctor entered the order directly into a computer, using a handheld device, at the bedside.

Dr. Glen Geiger is a physician in Internal Medicine at Sunnybrook & Women's College Health Science Centre in Toronto. In 1999, Glen initiated a study where he asked doctors and nurses in his service to record process times on the drug orders. He discovered that over 25% of the orders were not administered within the targeted time frame. Most failures were not even close. These were process errors; they do not include cases where patients received the wrong drugs. The results of this study were a surprise to hospital leaders and continue to surprise health care professionals from other areas. Many issues that we discovered at Sunnybrook are common to most manual drug order entry processes.

It is fairly obvious that physician order entry will dramatically reduce cycle time for the process and reduce the workload of all parties – with the possible exception of the physician. Thus, we needed to convince the doctors that the system would dramatically improve the process without significantly increasing their workload. We decided to use simulation to quantify the potential for process improvement. We believed that it would be an important tool for demonstrating the advantages to physicians.

In the summer of 2001, four students, including three industrial engineers and one medical student, were hired to perform a detailed analysis of the prescription process. The students spent two months documenting the current process through interviews and direct observation. They then conducted a two week data collection during which all drug orders for a thirty-six bed Internal Medicine ward were tracked to facilitate the creation of a simulation model. The results of the detailed tracking confirmed Dr. Geiger's earlier results. In particular, many medication orders were not administered to patients in a reasonable amount of time [7].

One of the surprising discoveries was that this seemingly simple process was actually quite complex. For example, a different process was used when a doctor phoned an order in to the nurse as opposed to when the order was written. The day and night processes are different because the pharmacy is closed at night. At night, instead of placing an order with the pharmacy, nursing staff can access a night cupboard for commonly required medications.

There are also communication issues. For instance, pharmacists regularly visit the ward and review patient charts. Pharmacists sometimes write a "P" on the order. Some of the nurses knew that the "P" meant the pharmacist had reviewed the order. Others thought it meant they had "Pulled" the order. Also we found some confusion surrounding a physical flag attached to the chart. When the doctor writes an order s/he puts the flag up. Unfortunately, there is only one flag on each patient chart, and it is used for all orders. When multiple orders (e.g. drugs, lab tests, imaging, etc.) are in the file, the possibility exists that the nurse will find only the first one and put down the flag. Nurses check the complete chart every two hours, but errors sometimes occur. One order was in the chart for two days before the students pointed it out to the physician.

Timeliness of medication orders can be measured in several ways. For example, suppose a doctor prescribes medication for a patient at 11 a.m. to be administered three times a day at 6 a.m., 2 p.m. and 10 p.m. We could consider the delivery to be late if it was not back in time for the 2 p.m. dose administration. Pharmacokinetic practice says that a dose of medicine can be administered up to four hours late (for example, at 6 p.m.), half of the dosage interval, and still be on time. From a process perspective, we estimated that a prescription should not take longer than two hours to fill. All three measures were used in the study for determining whether an order was filled on time.

8.4.2 Challenges encountered

Lack of control From a quality perspective, we were quite surprised with the apparent lack of control of the prescription process. Since no written documentation was available, to determine how the process worked we simply asked everyone what he or she did. There was no formal training for nurses or doctors. New staff members learned by word of mouth. Virtually everyone we spoke with had a different view of the process. Moreover, there is no standardization across the hospital; each ward had apparently developed its own set of procedures. We attributed this to the perception that the process was "simple" and therefore did not require formal documentation and training.

Need for greater modeling detail In the validation of our simulation model, we could not get the turnaround times for medication orders in the model to match the times that we observed in practice. Initially, the average time in the model was 225 minutes, while the true average from the data was 262 or 16% higher. This seemed odd since the distributions in the model were based on statistically fitting the same data.

A major cause of this discrepancy originated in the pharmacy portion of the model. Initially, we had assumed that the pharmacy part of the process would be fairly reliable once the orders arrived there, so we chose to model the pharmacy as a black box. Since the pharmacy was computerized, we expected the process would provide consistent results and that when an order was picked up, it would be processed expeditiously and delivered back to the ward. We were surprised to discover dramatic variations in turnaround times.

Several months after the initial study, and long after the summer students had returned to school, we concluded that we needed to expand the scope of the analysis and perform a detailed process analysis of the pharmacy area. We discovered several anomalies. There was an 8 a.m. rush in the pharmacy to fill all of the orders that had accumulated overnight. The pharmacy processed each ward as a group of orders, and the sequence of the wards varied daily. Many long delays occurred when a particular ward was left toward the end of the sequence. We also learned that the pharmacy's workforce was highly variable. The pharmacy could not tell us how many pharmacists were working at any one time; the assignments varied hour-by-hour and day-by-day. Furthermore, it was found that a major complication was created by orders requiring clarification. We discovered that over 13% of orders required the pharmacist to call the doctor. These orders would be set aside temporarily while the pharmacist paged the doctor. We were unable to collect meaningful statistics on how long it took to get an answer to a page; however, it appears that about 5% of orders took more than three hours, and many of these were not resolved in the same day. Moreover, many of the pharmacists processed the simpler orders first, and saved clarifications until later in the day when they had some spare time.

Technology implementation As described above, the motivation for the simulation was to be able to demonstrate the potential process improvements that accrue from using automated physician order entry. Sunnybrook has already purchased the software to implement the automated prescription entry process. However, the system is still in development and the user interface must be customized. We cannot complete the simulation without first performing experiments with the interface to determine the distribution for access time. We do not expect it to take long, but this is likely to be the central measure of success for the physicians. We expect to have a pilot version ready by Spring 2004.

8.5 THE CROWDED STUDY: CAUSES AND RELATIONSHIPS OF OVERCROWDING AND WAITING IN DIFFERENT EMERGENCY DEPARTMENTS

8.5.1 Description of the application

Waiting times and overcrowding in the Emergency Department (ED) have become increasingly serious problems over the past several years. In the United States, surveys of hospital directors have reported ED overcrowding in almost every state [8, 9]; ED overcrowding has also been reported in Europe [10]. In most hospitals Emergency Department overcrowding is a symptom, rather than a cause, of the problem. For example, overcrowding in the province of Ontario in Canada is often attributed to patients who have been admitted to hospital, but who are waiting in the ED until a ward bed becomes available. Beds are often blocked in the wards because of discharge delays (e.g. waiting for test results, waiting for nursing home space, rehabilitation beds or home care). Thus, to really understand how the ED functions and why it backs up, it is necessary to develop a detailed process analysis specifically focusing on the impact of bed blockers.

A number of people have done ED simulations in the past, but have generally assumed that the processes outside the ED have little direct impact on its overall operation. Jun et al [11] present an extensive survey of simulation applications in health care. In fact, several simulation studies have been conducted to specifically analyze the issue of overcrowding in the ED. A priority queuing model was developed in one study to evaluate the potential impact of adding a fast-track facility to an emergency department [12]. Simulation modeling has also been employed to examine the relationship between hospital bed capacity and emergency admissions rates [13], with the finding that bed shortages can be expected when average bed occupancy rates exceed 85%. Simulations have been successfully applied to investigate the impact in the ED of nurse scheduling on utilization and patient length of stay (LOS) [14-16]. Based on these studies, recommendations were made for changing policies on staff scheduling, triage procedures and nursing responsibilities. Using the simulation model, the potential savings from the proposed changes were quantified.

The study described earlier in this chapter [3] also modeled the flow of patients through an ED. For all of the ED simulations mentioned, the patient LOS in the ED is assumed to be an exogenous variable, sampled once for each patient from a statistical distribution based on historical data. This assumption is reasonable given the complexity associated with most emergency departments. One can usually construct and validate these models quite adequately. However, this method does not allow decision makers to investigate the impact of changing non-ED components on the overall process flow. For example, if the time required to complete an external consult was reduced, or the process for MRIs was improved, what impact would that have on wait times in the ED or throughout the entire hospital?

In fact, our analysis suggested that the ED is a very complex entity, referred to by some of the doctors on our study team as “organized chaos”. In 2002, a team including operations researchers, ED physicians, a statistician and an epidemiologist received funding for a two-year study to analyze the detailed processes in ten Ontario hospital emergency departments. The Causes and Relationships of Overcrowding and Waiting in Different Emergency Departments (CROWDED) study was designed to include detailed data to promote better allocation decisions for scarce resources such as doctors, nurses, and examination rooms. The hospitals were selected to represent a cross-section of geography and clientele. Three large teaching hospitals, four community hospitals, and three rural emergency departments were selected for inclusion in the study. Two full time research assistants were hired for one year to collect data by directly observing patients, doctors and nurses. We conducted three trips to each site. There was a pre-visit of 2-3 days to study the layout, understand the policies, meet people, and put up posters to educate and inform people about the study. Data collection was conducted in two separate one week periods at different times of year to get a sense of pattern changes over time. The project was designed to construct a generic model of an ED that can provide detailed decision support for a wide range of process flow issues.

8.5.2 Challenges encountered

Doctors are difficult to track As mentioned earlier in the CHEO study, it is often difficult to tie physician workload to a specific patient. Doctors consult on the phone, read x-rays, view images on-line, chat with nurses and residents, as well as performing many other activities; all of these activities are done in the course of a patient's treatment, but rarely happen when the physician is proximate to the patient. However, since doctors are probably the scarcest ED resource, it is important to determine accurate workload information for them. In the CHEO study, we chose to implement a work sampling method supplemented with a job shadow provided by a small group of nurses. In the CROWDED project, we had significantly more resources at our disposal, and we were determined to get very accurate workload information.

Many physician and patient processes could not be observed directly. The observers needed to use indirect means of observation, such as consulting the patient chart or the "white board" that keeps track of patient progress in the ED. In some study sites, we had access to the hospital's electronic order entry/patient tracking systems. This also helped the observers to track the patients' pathways. However, in both paper and electronic documentation, it was found that recorded times did not usually reflect the actual time or duration of a process. For example, nurses or ward clerks might log an order for blood work into the computer at a certain time, but might not collect the blood until much later. The time recorded on the chart frequently corresponds to the time the order was entered; there is no information about the actual start and end time of the process.

Missing data A related issue we discovered during the course of the project was that it was quite difficult to collect complete, accurate flow data on all ED patients. The observers estimated that some data was missing for approximately 10-15% of patients in the study.

For example, critically ill patients may stay in the ED for a long time. The CROWDED study did not employ 24 hour observation, so process flow data tended to be incomplete. The observers tried, wherever possible, to fill in blanks using the patient's chart, but it was difficult to get good time estimates. When patients remain in the ED outside of the period of direct observation, the patient pathway through the ED will always have some missing data. However, even if some patient data was missing they were usually able to record a minimum data set including admission or discharge time along with any other charted information. Patients remaining in the ED for longer than a single observation shift tended to be admitted patients, or patients that required lengthy observation.

Trauma cases were also difficult to track. Because treatment for trauma cases needs to be started immediately, charting is usually performed after the fact. Moreover, trauma cases are generally handled behind closed doors. On the assumption that it is inappropriate for data collectors to be inside a trauma room or that observation may impede patient care, it was decided to forego direct observation of trauma cases. Instead the points of time of "trauma begins" and "trauma ends" were used as a way to track the many processes that could not be directly observed.

Similarly, acute patients may also receive treatment or undergo tests according to medical directives behind closed doors/curtains. In these cases, many different processes may be happening. The observers used the charts after the fact to determine which processes had occurred. This usually provided reasonable results in terms of what happened, but not always when it was done. It was sometimes possible to estimate start and/or end times if, for example, the observers saw nursing staff gathering up supplies or equipment prior to a process, but there was a lot of guesswork.

In addition to "closed door" treatments carried out by staff on trauma and acute patients, another challenge was that processes for many patients happen simultaneously. The research assistants were only able to observe the processes of one patient at any given time. Sometimes in the case of an acute patient such as, for example, a heart attack victim, a team of nurses and doctors might perform a series of treatments until the patient is stabilized. To capture all these processes required observation of that particular patient for an extended period of time. During that time, other things could be happening to other patients which were not observed or recorded.

Layout issues The layout of the ED sometimes created problems for data collection. Some EDs were physically spread out which made it difficult to see what was happening to a patient or to observe the doctor/nurse treating the patient. In one ED in our study, the physical layout was divided into a "major" and a "minor" side. During peak times, doctors would be assigned exclusively to one side or the other. However, in off-peak times, one doctor would float between both. It became impossible to follow the doctor, the nurses and the patients simultaneously in this environment. At another site the ED had a number of separate areas. The segregation of the ED made tracking difficult for the observers.

Fast-track clinic Some study sites had an off-site "fast-track" clinic (FTC) or an "urgent care" center, separate from the ED. Again, this physical separation made it difficult to track patients.

At one site, the hospital had a fast-track clinic operating from 2pm - 10pm on weekdays. The FTC was in a separate area from the ED, but had four beds and was staffed by a Nurse Practitioner[1]. During its hours of operation, less acute patients came to the ED, saw the triage nurse, were registered by ED staff, but then headed to the FTC for treatment. The fact that the FTC is external makes it harder to observe the patient flow process. It was tempting to simply ignore patients who were sent to a FTC; however, we believe it is important to model it as an internal process, using ED resources. In particular, one of our model decision variables may be to consider adding two FTC physicians, or having some shared resources work in the FTC and the regular ED.

Wait time before triage When we consider the question of ED wait time, part of that measure involves patients waiting before triage or registration. Predictably, none of the study hospitals tracked or had data on "time before triage". While we believe serious patients are seen immediately and all patients are triaged expeditiously, we asked our observers to sit in the waiting room and conduct a separate study of "time-to-triage" to determine the magnitude of this issue. Observing time-to-triage; however, meant that observers could not track patients inside the ED due to layout and sight-line issues. The results of our preliminary studies indicated patients frequently line up to be triaged, but critically-ill patients were not overly delayed.

Unplanned critical events In any study, blind luck (good or bad) sometimes comes into play. In the CROWDED study the data collection process was facilitated by a custom designed PDA application. After the first few site visits were completed, the PDA programmer made some minor adjustments to the application. Subsequently, after three days of collection following the

[1] A Nurse Practitioner is a Registered Nurse who has taken a graduate level program and who can perform many of the functions that are commonly associated with doctors.

adjustment, the observers discovered that a bug in the revised program blocked the transfer of all patient demographic information (name, age, gender, ID number, etc.) to the production database. This was a serious issue in terms of validation and completion of missing data elements.

Additionally, a number of unforeseeable public health issues arose during the collection process. The ED at one site was closed for several weeks because of an outbreak of the Norwalk virus, which interrupted our data collection. To make matters worse, after three days of data collection at a different site the next week, one of the observers became ill with Norwalk like symptoms. She went into voluntary quarantine, and the other observer attempted to collect what she could for the remainder of the visit.

However, the worst setback occurred in March 2003 when Toronto was hit with the SARS virus (Severe Acute Respiratory Syndrome). We had to pull the observers out of all study hospitals for almost two months. Even hospitals distant from Toronto were closed to non-essential personnel. Moreover, patient volumes in EDs throughout the province decreased in response to patient fear of SARS. Things started to return to normal after a short period, but we needed to extend the data collection for two months, hire an extra observer, and adopt an aggressive visit schedule to make up for lost time.

Preemption When the ED gets busy, some processes can be preempted by more critical needs. When a doctor or nurse comes back to the interrupted process, they may have to start the entire process again. For example, while a nurse would not interrupt an IV start, he/she might interrupt an assessment. When the nurse later returns to complete the assessment, it is usually necessary to repeat some elements. One of our team members, an ED physician, believes the process can almost grind to a halt when things get busy. Physician assessments and nursing assessments are frequently interrupted. In our study the observers attempted to track starts and ends for all processes, even those that were incomplete, but this was an imperfect solution.

Administrative issues Despite our best efforts, staff members at the hospitals were often suspicious about the intentions of our study. It was perceived as a study created by the provincial government to streamline the costs of health care and reduce employment. Many staff members at hospitals believed the study would never be used to benefit health care or that the study was misguided. Our research assistants were conscientious in assuring the participating hospital staff that we were performing an independent study funded by CIHR, and not by their hospital administration or the provincial government, and that we were doing our best to accurately represent the processes in their departments so that patient flow could be improved without compromising patient care.

We also found that turnover in key management positions in the ED was a factor in our project. During our one year study we had the primary contact change at over half of the hospitals. Despite the fact that all of the study hospitals had agreed to be part of the project, we often discovered when we went to visit a site the current managers had no knowledge of our project, and we needed to begin the sales pitch again. In one case, we needed to reshuffle our data collection schedule because new managers did not know we were coming.

Security Data security was a very critical component of our study. During data collection, our team needed personal information to allow matches between paper and electronic hospital records. PDAs used for data collection were downloaded daily into the laptop computers and backed-up daily on a password protected CD-ROM, which was kept in a safe, secure location. Upon return to the lab, the data was copied from the laptop onto a master CD-ROM, which was kept in a locked drawer. The data on the laptop was then stripped of all personal identifiers (name, address, ID numbers, etc.) to ensure patient anonymity.

8.6 DISCUSSION/CONCLUSION

Health care is an enormous business offering a wealth of potential applications for simulation and other operations research techniques [17]. However, health care is a business unlike any other business. In our experience the context in which a decision making situation arises has a significant impact on the way in which it is solved. Nowhere is this truer than in health care. We believe that, because analysts and clinicians speak different languages, operations research has made fewer inroads into this field than in more traditional industries. However, our experience also suggests that OR techniques can be successfully applied in the health care setting. The secret is to understand the unique nature of the health care business and its impact on models, decision makers, and the development of implementable policies.

In this chapter we have used four simulation projects to highlight the practical lessons of applying operations research in health care. Analysts should remember that decision making in hospitals is characterized by multiple players; seeking the council and incorporating the objectives of all decision makers is vital in this environment. In this industry data collection systems may not be designed to provide administrative data; collecting data on patient flow and operational performance metrics may require some patience and may extend the project life cycle. Finally, while many processes and procedures are fundamentally similar regardless of the institution, there are usually enough local quirks to render multi-site "cookie-cutter" models infeasible.

Health care is a fascinating industry to work in. The authors have, over the past decade, devoted themselves to applying operations research to health care and have enjoyed the experience immensely. It is our desire that the lessons we learned will prove useful for others following in this field.

Acknowledgments

The authors would like to thank the Ontario hospitals that participated in the studies described in this chapter: The Toronto Hospital for Sick Children, the Toronto General Hospital, Mount Sinai Hospital, Sunnybrook and Women's College Health Sciences Centre, the Children's Hospital of Eastern Ontario (Ottawa), Markham-Stouffville Hospital (Markham), London Health Science Centre, Sudbury Regional Hospital, Stevenson Memorial Hospital (Alliston), South Muskoka Memorial Hospital (Bracebridge), Windsor Regional Hospital, Quinte Health Care Corp (Belleville), The Royal Victoria Hospital (Barrie) and Kingston General Hospital.

Thanks are also due to co-investigators and research assistants: Prof. Linda O'Brien-Pallas, Prof. Linda McGillis-Hall, Dr. Michael Schull, Dr. Glen Geiger, Prof. Greg Zaric, Dominic Fernandes, Yaron Derman, Cathy Wong, Carolyn Busby, Jessica Law, Tracy Ayow, Angelo Gentile Gosia Szymanski, Patricia Ricci, Aviv Gladman, Leila Peyravan and Rhea Plosker.

Finally, the authors would like to thank the Ontario Ministry of Health and Long Term Care, the Natural Sciences and Engineering Research Council of Canada, the Canadian Institute for Health Research and the Social Science and Humanities Research Council of Canada for their assistance funding these works.

References

[1] Blake, J.T., M.W. Carter, L.L. O'Brien-Pallas, and L. McGillis-Hall (1995). A surgical process management tool. *Proceedings of the 8th World Congress on Medical Informatics MEDINFO 95.*

[2] Carter, M.W., L.L. O'Brien-Pallas, J.T. Blake, L. McGillis, and S. Zhu (1992). Simulation, scheduling and operating rooms. *Proceedings of the 1992 Simulation in Health Care and Social Services Conference,* J.G. Anderson, Ed., Simulation Council Inc., San Diego, 28-30.

[3] Blake, J.T., M.W. Carter, and S. Richardson (1996). An evaluation of emergency room wait time issues via computer simulation. *INFOR,* 34, 263-273.

[4] Ash, J.S., P.N. Gorman, and W.R. Hersh (1998). Physician order entry in U.S. hospitals. *Proceedings of the AMIA Annual Symposium,* 235-239.

[5] Bates, D.W., L.L. Leape, D.J. Cullen, N. Laird, et al. (1998). Effect of computerized physician order entry and a team intervention on prevention of serious medication errors. *Journal of the American MedicalAssociation,* 280, 1311-1316.

[6] http://www.leapfroggroup.org, [online document] Accessed on June 29, 2003.

[7] Wong, C., G. Geiger, Y.D. Derman, C.R. Busby, and M.W. Carter, (2003). Redesigning the medication ordering, dispensing, and administration process in an acute care academic health science centre. *Proceedings of the 2003 Winter Simulation Conference,* S. Chick, P.J. Sánchez, D. Ferrin, and D.J. Morrice, Eds., New Orleans, LA, 1894-1902.

[8] Derlet, R.W. and J.R. Richards (2000). Overcrowding in the nation's emergency departments: Complex causes and disturbing effects. *Annals of Emergency Medicine,* 35, 63-68.

[9] Andrulis, D.P., A. Kellermann, E.A. Hintz, B.B. Hackman, and V.B. Weslowski (1991). Emergency departments and crowding in United States teaching hospitals. *Annals of Emergency Medicine,* 20, 980-986.

[10] Miro, O., M.T. Antonio, S. Jimenez, A. De Dios, M. Sanchez, and A. Borras (1999). Decreased health care quality associated with emergency department overcrowding. *European Journal of Emergency Medicine,* 6, 105-107.

[11] Jun, J., S. Jacobson, and J. Swisher (1999). Applications of discrete event simulation in health care clinics. *Journal of the Operational Research Society,* 50, 109-123.

[12] Siddharthan, K., W.J. Jones, and J.A. Johnson (1996). A priority queueing model to reduce waiting times in emergency care. *International Journal of Health Care Quality Assurance,* 9, 10-16.

[13] Bagust, A., M. Place, and J.W. Posnett (1999). Dynamics of bed use in accommodating emergency admissions: Stochastic simulation model. *British Medical Journal,* 319, 155-158.

[14] Kumar, A.P. and R. Kapur (1989). Discrete event application – Scheduling staff for the emergency room. *Proceedings of the 1989 Winter Simulation Conference,* MacNair, E.A., K.J. Musselman, and P. Heidelberger, Eds., IEEE, Washington, DC, 1112-1120.

[15] Rossetti, M.D., G.F. Trzcinski, and S.A. Syverud (1999). Emergency department simulation and the determination of optimal attending physician staffing schedules. *Proceedings of the 1999 Winter Simulation Conference.* Farrington, P.A., H.B. Nembhard, D.T. Sturrock, and G.W. Evans, Eds., Phoenix, AZ, 1532-1540.

[16] Kirtland, A., J. Lockwood, K. Poisler, L. Stamp, and P. Wolfe (1995). Simulating an emergency department is as much fun as .... *Proceedings of the 1995 Winter Simulation Conference,* Alexopoulus, C., K. Kang, W.R. Lilegdon, and D. Goldman, Eds., Arlington, VA.

[17] Carter, M.W. (2002). Health care management – Diagnosis: mismanagement of resources. *OR/MS Today,* April, 26-32.

9 ESTIMATING RISKS TO THE PUBLIC HEALTH

Rose Baker

Centre for Operational Research and Applied Statistics
University of Salford
Salford M5 4WT, United Kingdom

ABSTRACT

Risks to public health arise from infectious disease, exposure to toxic substances such as asbestos, environmental insults, and from lifestyle risks such as smoking. The risk assessment that must precede healthcare interventions or legislation requires probabilistic, statistical and computational methodologies. The introduction to this chapter discusses how our perception of the risks to public health is changing, and identifies some trends in the methodologies used for risk analysis. Risk assessment is largely characterised by likelihood-based statistical inference, using point-process models of disease intensity as a function of position in space and time. Conditional likelihoods such as Cox's partial likelihood and matched-pairs logistic regression are widely used to eliminate confounding variables. Two examples of the use of such conditional likelihoods are given. In the first, new tests for the space-time clustering of cases characteristic of infectious disease are derived and exemplified. In a second application of conditional likelihood, some research on risks of Shigella infection to schoolchildren arising at school or from playmates is presented. The original content of this chapter is two new tests of space-time clustering, and a case-study using an unusual conditional likelihood.

KEY WORDS

Epidemiology, Risk, Likelihood, Knox test, Shigella

9.1 INTRODUCTION

9.1.1 The major risks to public health today

Nowadays we are seeking to identify ever smaller risks to public health. To this end, huge volumes of data are being made routinely available to the epidemiologist, and both statistical methodologies and computing power are being pushed to the limit.

Public health analysts have traditionally been concerned with risks from infectious disease and from food poisoning. The familiar story of how in 1854 John Snow removed the handle of the Broad Street pump in St. James's parish, London, to prevent the spread of cholera exemplifies this. More recently, lifestyle risk factors such as smoking were identified for late-life diseases such as cancer and heart attacks, where the link between cause and effect was harder to establish. In the last 20 years, environmental public health has emerged as a major concern. Risks of exposure to toxic materials such as lead, asbestos and air pollutants have been much studied, and the resulting legislation has greatly ameliorated these hazards [1]. Environmental concerns also include the possible existence of disease clusters, either of infectious origin or around some 'environmental insult' such as a power station or toxic waste landfill site.

In the developed world, there is now great anxiety about the risks posed by human activity in general, and from technology in particular. This includes genetically modified foods (in Europe), radioactivity from power stations, electromagnetic effects from power lines, pollution from toxic waste, global warming, and environmental pollution in all its forms. On the other hand, many people resolutely continue to smoke despite its clearly proven ill effects, concern about cancer from mobile phones has not reduced their use, and obesity even among the young is increasing. A large subculture abuses hard drugs, with resulting high mortality. Risks posed by the actions of others are evidently perceived as more threatening than risks posed by one's own lifestyle [2].

Currently, the risk from ELF (extremely low frequency) magnetic fields, which has been studied with variable results for over 20 years [3, 4], has been firmly established. Only from 2000 onwards have large definitive studies and meta-analyses swung the weight of evidence firmly towards the existence of a real risk. ELF fields, which result from familiar technology widespread throughout the infrastructure of our cities and homes, are now known to greatly increase the risk of miscarriages [5, 6], and to increase the risk of childhood leukaemia [7]. They may also cause asthma and many other chronic illnesses [8]. ELF magnetic fields do not inspire the 'fear and dread' that radioactivity or genetically modified foods do, and are currently the concern mainly of local pressure groups opposing new power lines or cellular telephone masts. The situation here may change, and extensive litigation could result.

The traditional concern of epidemiology, infectious disease, is also re-emerging as a major risk factor. Globalisation, travel and population movements can result in outbreaks of locally unfamiliar diseases. New diseases such as AIDS and variant Creutzfeldt-Jakob Disease have appeared. In the near future infectious disease may be set to again take centre stage as a risk factor in the developed world, with the growth of multidrug-resistant strains of once familiar diseases such as multidrug-resistant tuberculosis [9], and with tropical diseases such as malaria increasing their range through global warming.

Healthcare provision requires the estimation of risks to health from all these hazards. This chapter is concerned with the probabilistic and statistical techniques used. Risk estimation is not, however, the end of the story: it must be followed by remedial action. This may require simply giving reassurance to the public, issuing guidelines on lifestyle, taking action to remove particular environmental 'insults', or pressing for changes in the law or in public policy. Snow in fact had to argue his case with the Board of Guardians of St. James's parish, and the pump handle was removed by them the following day, despite some doubts about the correctness of his case. The water board were then directed to improve the quality of the water. A modern analogue is the study led by Anto and Sunyer [10], in which asthma among residents of Barcelona was linked to soybean dust released when soybeans were unloaded to silos from ships in the harbour. This work led to the prompt installation of filters to prevent airborne dissemination of soybean dust in 1987 [10].

The first step in ameliorating health hazards is to demonstrate that they exist. The next section addresses this issue.

9.1.2 Statistical inference and its problems

Broadly, the aim of inference is to demonstrate an increased risk of disease or death arising from a particular risk factor, and then to quantify this risk. Several types of statistical analysis are relevant.

The fast-growing methodology of disease mapping is used to reveal geographical variations in risk [11]. Much work has also been devoted to the study of disease clusters, either in space, perhaps around a power station, or as the space-time clusters characteristic of infectious disease. Cluster alarms are frequently reported to public health authorities, and must be investigated, although very often the cluster is the result of random variation, and no action need be taken [12, 13]. Such alarms demonstrate the strength of public anxiety about health risks posed by human activity.

Ecological analysis measures explanatory variables and relates them to disease. Disease monitoring or surveillance seeks to detect outbreaks of disease very early in their progress [14] rather than studying outbreaks retrospectively.

There are many problems with risk assessment. Some are endemic to the science or art of statistics itself. For example, we may wish to show a causal relationship but can only demonstrate an association. Epidemiologists have long wrestled with this problem and have developed stringent criteria for showing causality [15]. In practice, epidemiologists may not be able to satisfy these criteria. However, the strength of an epidemiological case for an association, based on several different studies, eventually becomes so strong that rival explanations become increasingly implausible to all but cranks or those with vested interests [1].

Other problems stem from the limited nature of the available data, such as confounding and bias in general.

Confounding is best introduced by an example. If we wish to examine the association between drinking and cancer, smoking is a confounding variable. People who drink heavily also tend to smoke, so a naïve analysis would show a strong association between drinking and cancer. Correcting for smoking, by examining the drinking-cancer association separately for smokers and nonsmokers, shows the effect of drinking to be small.

In general, confounding variables (sometimes called nuisance variables in the statistical literature) are either sources of risk in their own right, or they may augment or potentiate the effect of other variables in which we are interested. Such variables need to be included in any model of risk, but may be unobserved or completely unknown. Much of the technical content of this chapter deals with attempts to overcome this problem.

Epidemiologists have long been aware of many different types of bias that 'can lead to conclusions that are systematically different from the truth'. Last [16] cites 27 different types of bias, of which confounding bias is one. They arise at all stages of a study from design and initial sample selection through interviews (recall bias), modelling and data analysis, to (finally) publication bias, where the picture is distorted by the nonpublication of negative or uninteresting results.

The ecological fallacy [16] is a major source of bias. This arises when variables are measured over a region and the aggregated variables are used to draw conclusions at the individual level. For example, Durkheim [17] found that the suicide rate was greater in regions where a greater percentage of the population was Protestant. There is an obvious explanation, but the data could also be explained if Catholics were more likely to commit suicide in regions where they felt beleaguered. Related to this, fitting nonlinear models also requires the use of special statistical methods when using aggregate level data [18].

In general, our lack of knowledge about the biological basis of some risks, such as ELF magnetic fields, leads to erroneous estimates of the exposure suffered by individuals to the hazard, and hence reduces our estimate of the risk posed, perhaps to the point where it does not attain statistical significance. , for example, wiring type has been used as a risk marker but turns out not to be closely related to risk [4]. Our not knowing which subset of the population is at risk also reduces estimated risk, through the dilution of the susceptible population with nonsusceptibles.

In addition to the many biases identified by epidemiologists, statisticians are becoming aware of a widespread tendency to understate the size of errors and confidence intervals. This bias appears empirically in meta-analyses of major trials. One cause is conditioning on the model finally selected. This means that one (rightly) chooses the model to best fit one's data after what may be a long process of model fitting and iterative refinement, but then (wrongly) acts for purposes of statistical inference as if the model had been decided on without any reference to the data. Modern computing power augments this problem by making it feasible to fit many models. Naturally, the model finally selected fits the data 'too well'. There is as yet no fully satisfactory solution to this difficulty. The problem is only partially alleviated by choosing models using such model-choice criteria as the corrected Akaike Information Criterion (AICc) [19]. Bayesian model-averaging is another, albeit computationally expensive, alternative [20].

Experimentally, the possibilities available to epidemiologists are limited. There have been a few supervised healthcare interventions where a randomised group was encouraged to change lifestyle, but such interventions cost millions of dollars [21]. Prospective or cohort studies are relatively bias free, and are the 'gold standard', given that people cannot generally be randomised to adopt different lifestyles as in a clinical trial, and certainly not while the effect of doing so is in doubt. Prospective studies may however take many years to complete and accumulate only a few cases. They cannot include currently unknown risk factors. Retrospective or case-control studies can be carried out quickly, are more cost-effective and are widely used [21].

9.1.3 Definitions

Some of the technical terms used in this chapter are now defined.

Risk refers to the probability that an event such as illness or death will occur. Relative risk or risk ratio is the ratio of risk in exposed individuals to risk in unexposed individuals. Attributable risk is the risk that could be removed if the exposure to the risk factor were eliminated. Last [16] gives full definitions of these concepts.

The hazard of an event (such as death) has a precise meaning in statistics. For a hazard *h(t), h(t)dt* is the probability that the event will occur in the time interval *(t, t+dt),* given that (conditioned on the fact that) it has not yet occurred by time t. When more than one event can occur, the *intensity p(t)* is defined such that *p(t)dt* is again the probability that the event will occur in the time interval *(t, t+dt),* but now conditioned on the previous history of such events. Intensity generalises the concept of hazard to repeatable events.

The likelihood function is the probability or probability density function (pdf) of observing the data given the model. Statistical inference is often likelihood-based. In particular, there is a class of powerful likelihood-based tests called *score tests* [22] where the test statistic is the derivative of the logarithm of the likelihood function with respect to a model parameter of interest, evaluated in the limit of 'no effect' when the parameter value is zero.

9.1.4 Current situation and trends in risk analysis

The methods of risk assessment currently used are mainly parametric models of the hazard or of the probability of disease or morbidity, fitted to data using likelihood-based methods. The widespread use of likelihood functions based on point-process models unifies this field methodologically.

While some likelihood functions are comparatively simple, such as those used in logistic regression, others are more complicated. These latter likelihood functions are derived from the theory of counting processes [23], and enable models of the intensity of disease as a function of spatial and temporal location of susceptible individuals and of 'environmental insults' to be fitted to data [24]. Thus the stochastic theory of counting processes is the probabilistic underpinning of modern risk assessment, and likelihood-based methods of inference are its statistical methodology. The 'executive arm' is the great availability of data and of computing power, with aids such as geographical information systems (GISs) replacing the older use of maps.

Satellites such as Landsat provide detailed information that can be organized by a GIS to facilitate epidemiological studies.

In statistics generally, there has been for some years a debate between Bayesians and frequentists. Frequentists regard the probability of an event as a statement about frequencies in many hypothetical trials, while Bayesians take a subjective view of probability. This enables Bayesians to write down 'prior probabilities' of events that reflect one's beliefs before the data are examined, and to use Bayes Theorem (which statisticians of all stripes accept) to construct a statement of the 'posterior' probability of the event given the evidence.

The Bayesian/frequentist debate has rumbled on for years. The computationally more expensive Bayesian methods have gained ground in the last decade, and some Bayesians have claimed that their approach constitutes a new scientific paradigm in the Kuhnian sense. Recently, there is some evidence that many statisticians are using an eclectic mixture of Bayesian and frequentist methods, in pragmatic attempts to find the best solutions to particular problems. Bayesian concepts such as prior probability and frequentist concepts such as confidence intervals may be mixed in the same article, and this is not now considered such a solecism as it was a few years ago. It is becoming 'horses for courses' as practitioners seek answers to practical problems, and leave the philosophy to take care of itself.

This trend can also be seen in epidemiology. Bayesian methods such as Markov-chain Monte Carlo models are now commonly used in disease mapping, where all available information must be synthesised, while frequentist methods predominate where it is necessary to present evidence of a hypothesis for public debate, independent of prior belief.

9.1.5 Conditional likelihood

Likelihood-based methods of inference are the most powerful, so tweaking the 'plain vanilla' or unconditional likelihood function in some way is an attractive option.

Many ingenious likelihood-based statistical methods have been developed to eliminate confounding variables, such as the use of Cox's partial likelihood [25] and the use of matched pairs in logistic regression [26, 27]. Both of these methods rely on conditioning the likelihood function in order to remove confounding variables. We need hopefully sacrifice only a little of the information in a dataset to get rid of confounding bias.

Thus in the matched-pairs case, we model the probability that one particular individual was a case, and *N* others were controls, given that (conditioning on the event that) one of the *N* individuals was a case. This approach enables risks due to known risk factors such as exposure to asbestos to be computed, while confounding variables are absent because of the matching of cases to similar controls [26, 27].

In Cox's partial likelihood approach [25], the hazard that an event such as death occurs to an individual is conditioned on death occurring to one of the individuals in the 'risk set'. Applying this conditional likelihood approach to the proportional hazards model, in which hazards from different sources multiply rather than adding, the 'baseline hazard' due to common confounding variables cancels out, leaving the dependence on risk factors alone in the likelihood function. Often in prolonged cohort studies, the baseline hazard of an outcome is expected to vary with time in an unknown way, and so the technique of Cox regression or partial likelihood is appropriate.

This chapter illustrates the state of the art of conditional likelihood methods, with two such likelihood-based approaches, but using less familiar conditional likelihoods.

The first study (Section 9.2) derived from an attempt to derive the well-known Knox test of space-time clustering [28, 29] as a score test. Here the hazard of infection is modelled as being elevated if close in space and time to a 'case' of the disease. Point-process models lead to a score test of infectious aetiology, and to estimates of the relative risk, when the population density $S(x, t)$ is known. When S is not known, but must be imputed from the locations and times of the cases, we are asking a lot of the data. By making reasonable assumptions about the dependence of S on space and time (that it factorises) it is possible to derive score tests based solely on the locations and times of infections. The Knox test can indeed be derived as a score test, and so can a 'corrected' version of the Knox test, which it is hoped may be more powerful than the Knox test itself.

In the second study (Section 9.3), which was motivated by an outbreak of dysentery in the North West of the UK, the risk of contracting dysentery *(Shigella sonnei)* from school toilets is investigated using a conditional likelihood related to the 'weird bootstrap' [23]. Data on infected individuals only are used. Intuitively, each individual acts as his or her own control; they will be infected at moments when the risk is high, and their younger selves who were uninfected at moments of lower risk play the part of controls.

This study is unusual in that rather than asking if there is an infectious aetiology, we ask if a particular mode of exposure to an infectious disease constitutes a risk factor. This study also models hazards of infection from independent sources as adding (as they should) rather than multiplying as they do in Cox's proportional hazard model. By fitting the model to data, we can estimate the relative risks of infection corresponding to the various risk factors, and also the attributable fractions of infection.

In this analysis, the device of *blocking* (matching) was also used. The transmission coefficient was assumed to be identical for children using the same toilet block, and to vary between toilet blocks. Children using the same block are demographically similar. Blocking makes it unnecessary to model the way in which the hazard of infection with *S. sonnei* depends on demographic variables.

9.1.6 Weird likelihoods

In general, infection is both a result of earlier infections (an effect) and also a cause of later infections. Some of the conditional likelihoods used in inference are derived by loosening this relationship, and imagining that cause and effect can be decoupled. We consider some other pattern of infection than the observed one, e.g. that those who were infected might have been infected at different epochs (Shigella study), or that infections might have occurred at any permutation of the observed space-time coordinates (space-time clustering study). We then condition the observed likelihood on the more general pattern of infection. However, in constructing this more general pattern of infection, the infections are only regarded as effects, and not as the originators of new infections. Since Anderson et al. [23] have referred to the Monte Carlo generation of random events from the general pattern of infection as the 'weird bootstrap', in this chapter the corresponding likelihoods, for want of a better term, are referred to as 'weird likelihoods'.

9.2 EXAMPLES OF METHODOLOGIES: IDENTIFYING SPACE-TIME CLUSTERS

Using conditional likelihoods derived from point-process models, methods are developed for testing whether cases have an infectious aetiology, and for estimating the relative risk arising from proximity to an infecter in space and time.

A problem here is the definition of 'proximity'. If we do not even yet know whether or not a disease has an infectious aetiology, we are unlikely to know the 'critical distances' defining spatial and temporal proximity. Use of the likelihood framework shows the way to a solution.

The likelihood function for n events arising from a point process occurring in space and time may be written as

$$L = \coprod_{i=1}^{n} \{p(x_i, t_i)S(x_i, t_i)\} exp\left(-\int p(x,t)S(x,t)dx\,dt\right), \tag{1}$$

where p is the intensity of the point process, S the population density, x_i and t_i are the space and time coordinates of the i^{th} infection, and the integral runs over a region of space (usually a surface) and time.

This likelihood function for a point process can be derived by writing the likelihood as a product of conditional probabilities for each small time step, where if an infection occurs, the conditional probability is $p(u)du$, and if no infection occurs, the probability is $1\text{-}p(u)du$. Equation (1) follows by taking the exponential term as the product-limit. This expression is well known [25, 26, 30].

A simple model of a point process with infection is to model p as

$$p(x,t) = \beta\left(1 + \alpha \sum_{j=1}^{n} f(x - x_j)g(t - t_j)\right) \tag{2}$$

where β is an unknown rate constant, and infection is increased by a factor $1+\alpha$ near an infected individual, i.e., the relative risk from proximity to an infecter is $1+\alpha$.

The definitions of f and g are:

$$f_{ij} = \begin{cases} 1 & if\ |x_i - x_j| \le d_{x,} \\ 0 & otherwise \end{cases}$$

$$g_{ij} = \begin{cases} 1 & if\ |t_i - t_j| \le d_{t,} \\ 0 & otherwise \end{cases}$$

where d_x and d_t are space and time critical distances. We also have $f_{ii} = g_{ii} = 0$, because a case cannot cause itself. Other definitions of f and g can be made, and much of the methodology will still carry through.

The rate constant β is unknown and is a nuisance parameter. It may be removed in three ways, as a marginal likelihood obtained by integrating β over its prior distribution with pdf $1/\beta$, by estimating β and plugging the estimate back in to obtain a profile likelihood, or by conditioning the likelihood on the 'weird' likelihood

$$L_n = \left(\int p(x,t)S(x,t)dxdt\right)^n exp\left(-\int p(x,t)S(x,t)dxdt\right)/n! \tag{3}$$

that *n* individuals are infected. In any case, to a constant factor, we have the conditional likelihood

$$L_c = n!\frac{\Pi_{i=1}^{n}\left\{\left(1+\alpha\Sigma_{j=1}^{n} f_{ij}g_{ij}\right)S(x_i,t_i)\right\}}{\left(N+\alpha\Sigma_{k=1}^{n} N_k\right)^n}, \tag{4}$$

where $N = \int S(x,t)dxdt$ and $N_k = \int S(x,t)f(x\text{-}x_k)g(t\text{-}t_k)dxdt$.

Assuming that the population density $S(x,t)$ is known, equation (4) can be used to estimate the relative risk $1+\alpha$ (e.g., as a maximum-likelihood estimate) and to derive confidence limits on α. The starting point is the log-likelihood from equation (4),

$$\ell = \sum_{i=1}^{n} log\left(1+\alpha\sum_{j=1}^{n} f_{ij}g_{ij}\right) - n\,log\left(1+\alpha\sum_{k=1}^{n} N_k/N\right) \tag{5}$$

Maximizing this with respect to α gives an estimate of the relative risk $1+\alpha$. A large-sample 95% confidence interval can be derived as the range of α for which ℓ exceeds $\ell_{max} - (1.96)^2/2$. Equation (5) can also be used to derive a score test of H_0 that $\alpha = 0$. The score statistic is

$$s = d\ell/d\alpha\Big|_{\alpha=0} = \sum_{i=1}^{n}\sum_{j=1}^{n} f_{ij}g_{ij} - n\sum_{k=1}^{n} N_k/N,$$

which is the number of close pairs of cases relative to the number expected if infections occur randomly. The variance of the score is estimated as

$$Var(s) = -d^2/d\alpha^2\big|_{\alpha=0} = \sum_{i=1}^{n}\left(\sum_{j=1}^{n} f_{ij}g_{ij}\right)^2 - n\left(\Sigma_{k=1}^{n} N_k/N\right)^2. \tag{6}$$

Hence a standardised score $z = d\ell/d\alpha|_{\alpha=0} / \left(-d^2\ell/d\alpha^2|_{a=0}\right)^{1/2}$ can be calculated, which is (for very large samples) normally distributed. Its reference distribution is better found by Monte Carlo simulation, i.e. by simulating random coordinates and times of infections a large number N (e.g., 10,000) times from the trivariate pdf $S(x,t)/\int S(x,t)dxdt$ and recalculating values Z of the standardised score. The p-value of the test is read off as the proportion of simulated Z values that exceed z.

The space and time critical parameters d_x and d_t are usually unknown. The test proposed here and the Knox test therefore have the unusual property of having nuisance parameters present only under H_1. The asymptotic theory of such tests is given in [31, 32]. Here however exact p-values are found using the Monte Carlo approach. The following argument derives the form of the test statistic for use with unknown critical parameters that gives the most powerful test.

By the Neyman-Pearson Lemma (e.g., [33]), particularising to the situation described here, asymptotically most powerful tests must be based on the difference of log-likelihoods

$$\ell\left(\hat{\alpha}, \hat{d}_x, \hat{d}_t\right) - \ell_o, \tag{7}$$

where $\hat{\alpha}$ is the maximum-likelihood estimate (MLE) of α, and $\hat{d}_x, \hat{d}_t$ denote the MLE of the nuisance parameters d_x, d_t. The log-likelihood ℓ_o under H_0 when $\alpha = 0$ does not depend on d_x, d_t.

From, e.g., [33], the asymptotic approximation of expression (7) is:

$$\ell(\alpha, d_x, d_t) - \ell_0 \cong s\alpha - (1/2)Var(s)\alpha^2. \tag{8}$$

Maximizing this expression for α and d_x, d_t yields

$$\ell\left(\hat{\alpha}, \hat{d}_x, \hat{d}_t\right) - \ell_o = (1/2) \sup_{d_x, d_t} z^2, \tag{9}$$

$$z = s/(Var\, s)^{1/2}. \tag{10}$$

As we do not wish to reject H_0 if $\sup z < 0$ (this is a 1-tailed test), $\sup z^2$ is replaced by $\sup z$. Tests based on the statistic $\sup z$ are therefore, for large samples, the most powerful possible. In practice, $\sup z$ is computed both for the sample, and for N simulations, and the p-value found as the proportion of simulations for which $\sup Z \geq \sup z$. We now have a test for space-time clustering when the population density is known as a function of space and time, but the critical distances are unknown.

The population density $S(x,t)$ may however not be known. The Knox test [28, 29] has been widely used in this situation. Here the test statistic is simply the number of close pairs. Its reference distribution is best found by permuting the labels of either space or time, thus making the Knox test into a permutation test. Such permutation is justified if the population density $S(x,t) = A(x)B(t)$. Here the population grows or decays uniformly throughout the region. The Knox test is known to give spurious results if this assumption is not met, for example if a population migration takes place. Kulldorff [34] suggests that the Knox test be tried, and only if it gives a significant result is there the need to acquire population data and to carry out any more sophisticated test.

The Knox test can be derived as a score test using a 'weird' likelihood, under the assumption that the population density factorises. Restricting the 'weird' likelihood in equation (3) to cases whose space or time coordinates are permutations of the observed cases, and conditioning the likelihood in equation (1) on it, the likelihood becomes

$$L_w = \frac{\coprod_{i=1}^{n}\left(1+\alpha\sum_{j=1}^{n} f_{ij}g_{ij}\right)}{\sum_{perms} numerator},$$

where the terms in S have cancelled out.

It is easy to see that the Knox test follows as the score test based on the statistic $d\ell/d\alpha|_{\alpha=0}$ from this likelihood function, where $\ell_w = \log L_w$. This formulation of the test enables the relative risk $1+\alpha$ to be estimated as the MLE of $1+\alpha$, and for a confidence interval on α to be estimated. For computational purposes, the denominator would be replaced by a large number of randomly chosen permutations.

Besides the large-sample confidence interval based on the Normal limit of the likelihood function, an exact confidence interval for α can be computed by exploiting the relationship between a statistical test and a confidence interval, that the confidence interval is the set of values α_0 of α for which the hypothesis $\alpha = \alpha_0$ can not be rejected. The score test can be done for any value α_0 and its p-value found as $p = \sum_{perms\ where\ \ell_w > s} L_w$, where s is the observed score.

Another benefit of the formulation of the Knox test as a score test is that it makes clear how covariates such as age and gender should be treated. Suppose that there are m classes of individuals. The risk α now becomes a 'who was infected by whom' matrix, where α_{kl} is the risk that a class l individual infected a class k individual. The population density S now also sprouts a class suffix. The likelihood becomes

$$L_w = \frac{\prod_{i=1}^{n}\left(1 + \sum_{j=1}^{n} \alpha_{c(i),c(j)} f_{ij} g_{ij}\right)}{\sum_{perms} numerator},$$

where c(i) denotes the class of the ith individual.

It is interesting to see what happens if we do not condition on a 'weird' likelihood. On conditioning the likelihood in equation (4) on the event that the cases are drawn from some permutation of space-time labels of the actual cases, we obtain the conditional likelihood

$$L_w = \frac{\prod_{i=1}^{n}\left(1 + \alpha \sum_{j=1}^{n} f_{ij} g_{ij}\right) exp\left(-\alpha \sum_{k=1}^{n} N_k\right)}{\sum_{perms} numerator}.$$

The corresponding score is now not quite the Knox statistic, as the exponential term does not cancel. The expression $\sum_{k=1}^{n} N_k$ would only be invariant under permutation if either A or B were a constant. The exponential term gives the probability that no other cases were infected besides the n cases observed, and under H_1 that $\alpha > 0$ this varies between permutations. Essentially we have the score statistic from equation (6), to be evaluated by permuting space or time labels. The second term is the expectation of the first, and varies much less strongly with permutation.

Estimating S=AB as proportional to a product of sums of delta-functions

$$S(x,t) = \sum_{i=1}^{n} \delta(x - x_i) \sum_{j=1}^{n} \delta(t - t_j)$$

at the space and time coordinates of the observed cases gives a score statistic which after a slight adjustment becomes the modified Knox statistic

$$T = \sum_{i=1}^{n}\sum_{j=1}^{n} f_{ij}g_{ij} - \frac{1}{n-1}\sum_{k=1}^{n}\sum_{l=1}^{n}\sum_{j=1}^{n} f_{kj}g_{lj}\,. \tag{11}$$

The first term is the Knox statistic, twice the number of pairs of cases that are close both in space and time. The second term is the expectation of the first under H_0. This follows because $\sum_{k=1}^{n} f_{kj}$ is the number of cases close in space to the jth case. A fraction $\sum_{l=1}^{n} g_{lj}/(n-1)$ of cases are close in time to the jth case, so if space and time distributions of cases are independent, we should expect the number of close cases to the jth to be $\sum_{k=1}^{n}\sum_{l=1}^{n} f_{kj}\, g_{lj}/(n-1)$, and twice the total number of close cases to be $\sum_{k=1}^{n}\sum_{l=1}^{n}\sum_{j=1}^{n} f_{kj}\, g_{lj}/(n-1)$.

The proposed test is thought likely to be more powerful than the unmodified Knox test, following an argument from Lehmann [35]. A test statistic such as the number of close pairs can only take integer values, whereas the statistic in equation (11) breaks this degeneracy and so can take many more values. Consider a set of simulated values of the test statistic. In moving from the Knox test to the proposed test, the fraction of simulated values greater than or equal to the sample value will decrease, making the p-value of the test smaller and the test more powerful, as all the previously lumped values now span a range.

Baker [36] gave a version of the Knox test for use when time and space critical distances are unknown. The same procedure follows for the score statistic in equation (11). The test statistic sup z is evaluated by evaluating z at a large number of grid points that cover all 'reasonable' values of the space and time critical distances. Its reference distribution under H_0 is found by treating a large number of permuted datasets identically, i.e. we find sup Z for each permuted dataset. The variance of T is needed for this, and can be found from the simulations, but it is more convenient to compute it using a formula.

The derivation of this follows the method for calculation of permutational variances of Knox-like statistics set forth very clearly by [37]. The calculations are straightforward but somewhat tedious, and so only the result is quoted here. Its correctness has been verified by simulation studies.

$r_j = \sum_{i=1}^{n} f_{ij}$, $r = \sum_{i=1}^{n} r_i$, $s_j = \sum_{i=1}^{n} g_{ij}$ and $s = \sum_{i=1}^{n} s_i$

$$
\begin{aligned}
Var(s) = {} & \frac{2rs}{n(n-1)} + \frac{4\left(\sum_{i=1}^{n} r_i^2 - r\right)\left(\sum_{i=1}^{n} s_i^2 - s\right)}{n(n-1)(n-2)} \\
& + \frac{\left(r^2 - 4\sum_{i=1}^{n} r_i^2 + 2r\right)\left(s^2 - 4\sum_{i=1}^{n} s_i^2 + 2s\right)}{n(n-1)(n-2)(n-3)} \\
& - \frac{2\left(r^2 - 2\sum_{i=1}^{n} r_i^2\right)\left(s^2 - 2\sum_{i=1}^{n} s_i^2\right)}{n(n-1)^2(n-3)} - \frac{\left(4\sum_{i=1}^{n} r_i^2\right)\left(\sum_{i=1}^{n} s_i^2\right)}{n(n-1)^2} \\
& + \frac{\left(r^2 - \sum_{i=1}^{n} r_i^2\right)\left(s^2 - \sum_{i=1}^{n} s_i^2\right)}{n(n-1)^3} + \frac{\left(\sum_{i=1}^{n} r_i^2\right)\left(\sum_{i=1}^{n} s_i^2\right)}{n(n-1)^2}
\end{aligned}
\tag{12}
$$

As an example of the use of this modified test, McHardy et al. [38] gave grid coordinates of homes and year of onset of 22 cases of Kaposi's sarcoma, in the West Nile district of Uganda. The authors state that computer analysis using the Knox [28] and Barton and David [39] techniques showed no significant space-time clustering. Time criteria varied from less than 1 to less than 5 years, and space criteria up to 24 kilometers (km). There are two pairs of cases for which onset was in the same year, and who lived within 2 km of each other.

A reanalysis using a Knox test with $d_x = 2$ km, $d_t = 0$ months gave $p = 0.042$, with 50,000 simulations. The modified test described here gave $p = 0.0176$. The distribution of the test statistic T from equation (11) is shown in Figure 9.1.

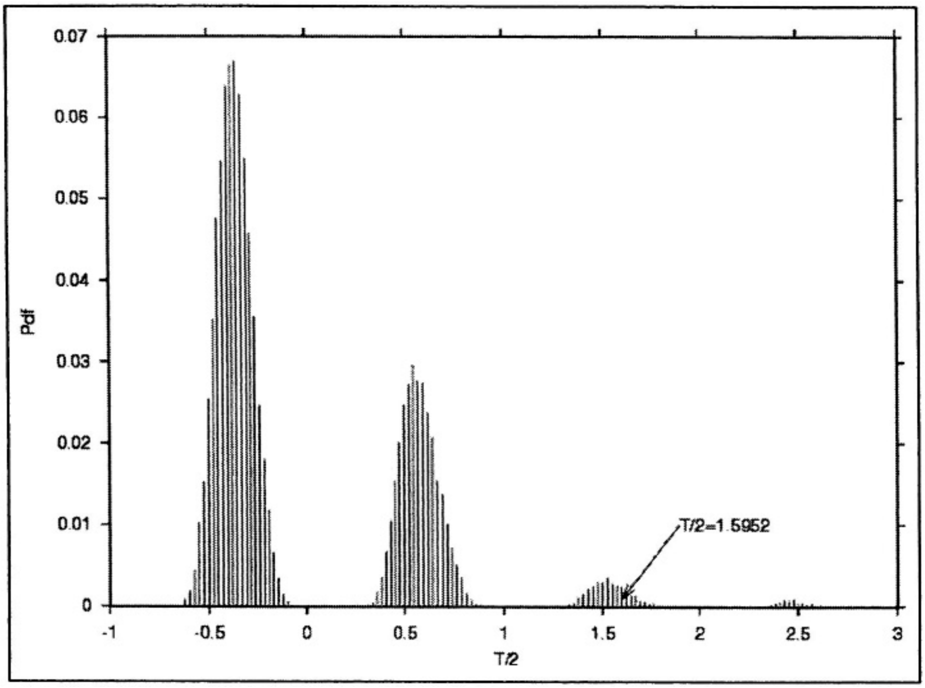
Using a grid of five time values from 0 to 4 months, and 16 space values from 0 to 15 km described in Baker [36] gave a significance level of $p = 0.109$, with $\hat{d}_t = 0$ years, and $\hat{d}_x = 0$ km. The corresponding test proposed here gave $p = 0.0609$ over the same grid. The ratio of observed to expected counts was 12.8.

These tests look promising, and it is hoped that they will be applied by practitioners. A FORTRAN95 program for the modified Knox test and details of the algorithms used are available from the author.

9.3 EXAMPLES OF METHODOLOGIES: THE 'WEIRD BOOTSTRAP' AND *SHIGELLA SONNEI*

The reported incidence of sonnei dysentery increased throughout the UK in the early 1990's, especially in the North-West, and was particularly high in

Figure 9.1 The distribution of the test statistic *T* from equation (11)



Salford. A retrospective study was carried out of dysentery transmission between children attending four Salford schools, to address the question of where infection was occurring.

A model was formulated in which the hazard of infection was a function of risk factors indicative of high contact rates, such as infecter and contact living close to each other, attending the same school, etc. Relative risks and attributable fractions were estimated by the method of maximum likelihood. This was carried out numerically, using a FORTRAN95 program written by the author, and which in turn used the NAG [40] function minimiser E04UCF.

The analysis showed that transmission of dysentery from contact with infected school toilets was not a major cause of infection in schools implementing PHLS guidelines [41], and neither was contact in the classroom. The analysis supported the view that closing schools down during dysentery outbreaks was not a useful control measure.

9.3.1 The shigella problem

Shigella sonnei has been responsible for over 90% of isolates of bacillary dysentery in the UK in recent years. Young children aged 5-8 years are at greatest risk. A typical case presents with diarrhoea lasting 2-3 days after an incubation period of 1-3 days. Abdominal cramps, vomiting and fever may occur. Susceptibility is general and immunity following infection is short-lived. Infection is transmitted by the faecal/oral route from human cases or asymptomatic excreters.

In 1991, there were approximately 9,200 laboratory reports and/or notifications of *S. sonnei* in the UK, representing the highest rates recorded for twenty years. There were nearly 17,000 in 1992, and thereafter the annual total has fallen steadily from below 7,000 to under 2,000 today. This study is of the 1991-1992 epidemic.

The isolation rate of *S. sonnei* in the North West Regional Health Authority rose sharply from 6.2/100,000 in 1990 to 51.3/100,000 in 1991. The outbreak commenced in September 1990. The total number of isolates reported to North Western Public Health Laboratory (NWPHL) in the period 1990-92 was 940. The peak annual isolation rate (March 1991 to March 1992) was 230 per 100,000. Of the 940 cases, 15.1% of isolates were from 0–2 year olds; 61.4% of isolates were from 3-11 year olds; and 22.7% of isolates were from people aged 12 and over.

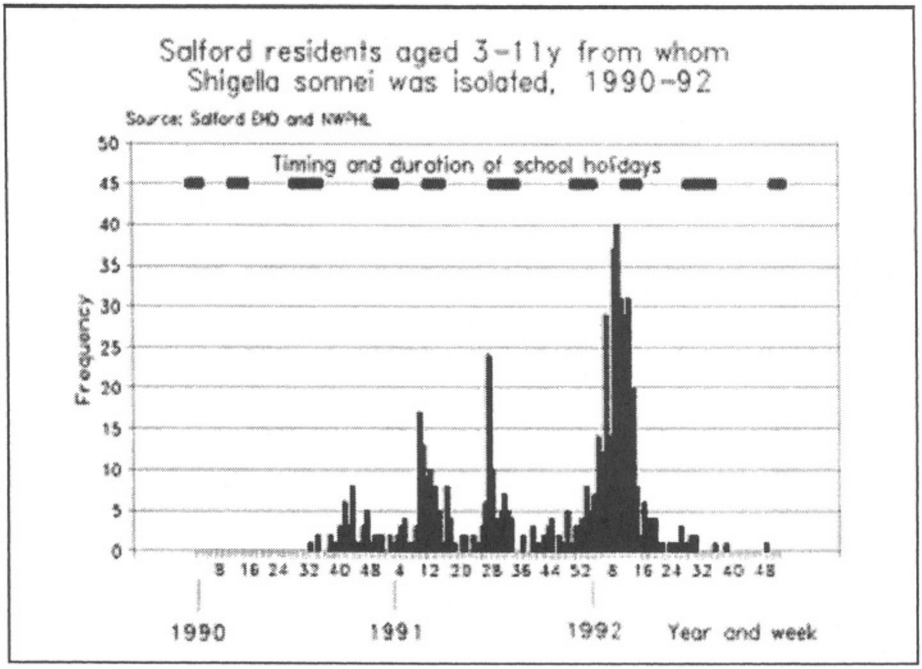
Children aged 3-11 years attending 54 out of 101 primary and nursery schools in Salford were involved. The highest number of affected children attending a particular school was 49 (isolation rate 19%). More than five children were affected in nine schools, and the isolation rate exceeded 5% in six schools.

Ascertainment and control of cases was undertaken by Salford Environmental Health Department (EHD) in close liaison with schools. Because of the protracted and serious nature of the outbreak, head-teachers of schools in affected areas were asked to report cases of diarrhoea and/or unexpectedly large numbers of absent pupils to the EHD. If two or more cases of sonnei dysentery were confirmed within one week, control measures were applied.

School infection control policy consisted of: emphasising the importance of handwashing to teachers; inspecting toilets to verify reasonable hygienic standards and the presence of adequate warm water, soap and disposable towels; exclusion of pupils for 14 days following the onset of symptoms; and, thrice daily, thorough cleansing of toilets with disinfectant.

Figure 9.2 shows the four-weekly incidence of confirmed cases of sonnei dysentery in 3-11 year old Salford residents from January 1990 to December 1992.

Figure 9.2 Confirmed cases of sonnei dysentery in the Salford epidemic



9.3.2 Data collection

The study period was defined retrospectively from January 1, 1992 to July 31, 1992. Figure 9.2 shows that this period forms a discrete episode within the epidemic curve of the general Salford outbreak.

The study population was defined using Salford EHD outbreak investigation records which gave the school attended by all cases and contacts. It consisted of children from the four Salford primary schools with the highest numbers of isolates during the study period.

EHD records contained name, address, age and date of onset of symptoms of affected individuals. Sex was imputed from first name, and grid references from home addresses. It was thus possible to calculate the distance between the homes of any two children. Schools provided details of class membership and class size during the study period for the entire study population. These data were collected in September 1992.

Two of the schools studied were only a few kilometers apart. To examine the effect of infections arising from children not attending the same school, data were extracted from the EHD records on all children up to age 11 living in that area who were infected during the study period.

All four schools had several toilet blocks. In many instances, a particular block was said by school staff to be used exclusively by members of a particular group of classes. Class membership was therefore associated with a unique toilet block. Staff were closely questioned about the possibility of children using toilet blocks other than the one associated with their class (especially during playtime when they were more mobile within the school premises). Where indiscriminate usage was thought to occur, individuals were not assigned to a toilet block; these individuals comprised only 4% of the total.

Notification to the EHD of cases of diarrhoea in schools continued throughout outbreaks. It is likely therefore that the majority of affected children came to the attention of the EHD. The policy of Salford EHD during the study period was to visit and obtain faeces samples from all contacts of known cases of *S. sonnei* dysentery.

The index case in a household or series of community contacts was taken to be the first case notified to the EHD. On microbiological investigation however, some of the contacts were found to be infected prior to the index case or to be co-primary cases. The EHD outbreak investigation records do not state whether cases were incident or revealed through contact tracing. The accuracy of onset dates of cases revealed through contract tracing mainly relies on the memory of parents ('recall bias'). Most onset dates are likely to be accurate to within two days, however.

9.3.3 Risk factors for dysentery transmission

The following risk factors were modelled:

1. Toilet block: This is a plausible source of infection.

2. Pupils' age: This may affect transmission of *S. sonnei* in several ways. First, children of 5-6 years are at greatest risk of infection. Also, children might tend to be at greater risk of infection from members of their peer group, who would naturally tend to be of similar age.

In the analysis, the device of blocking copes with the first effect, and the second effect is modelled explicitly.

3. Pupils' sex: Thomas and Tillet [42] show that there are slightly more isolates of *S. sonnei* from male primary school children, a result also seen here (male-female ratio 0.55). Most toilet blocks were single sex. Males are therefore more likely to use the same toilet block and more likely to have an early onset date, if a significant fraction are infected. It is also possible that the sexes segregate during play, so that for example males are more likely to acquire infection from males than from females.

The device of blocking prevented the wrong imputation of this sex-based effect to infection acquired from school toilets.

4. Pupils' class-membership: Person-to-person transmission or environmental contamination within the classroom could lead to increased transmission rates.

5. Infection from siblings: This is known to be common.

6. The infection of pupils from school friends who are not classmates: Person-to-person transmission between friends who use the same toilet block could occur during break-time or after school hours. Here an enhanced transmission between peers could be wrongly ascribed to use of the same toilet block. This effect means that estimates of infection arising from use of a common toilet block may overstate the amount of infection acquired from contamination of the toilet.

7. Pupils' attendance at a particular school.

8. Proximity: The infection of pupils from contact with an infecter living nearby, e.g. by playing together outside school hours.

It can be seen why it is important to model the effect of all these factors simultaneously. For example, many UK children are born within two years of a sibling, and if siblings tend to use the same toilet block, and the sibling effect were not modelled, we would erroneously ascribe sibling infection to infection from contaminated toilets. Modelling the main effects described above reduces error from confounding bias.

9.3.4 Model assumptions

The model assumptions are:

1. From DuPont [43], upon infection with *S. sonnei* there is a mean latent period of 1.4 days before the recipient becomes infectious, followed by an infectious period of 2.6 days. Onset dates were known only to the nearest day, and so it was considered that infection could have occurred at any time up to 4 days prior to onset. It is in theory possible to determine these parameters from the data; however with the values quoted, the results were not sensitive to changes in the parameter values.

2. After onset of the disease, children are removed from school until recovered. Hence it was assumed that after the onset date, they are not a source of infection to children attending school.

3. On return to the school, children are immune for the duration of the epidemic. Hence, after the onset date they cannot be reinfected. This assumption is reasonable, as Keusch and Bennish [44] conclude that for several months after infection there is immunity to *Shigella* reinfections with the original serotype.

4. Transmission is homogeneous throughout the whole population, with the exception of transmission attributable to risk factors modelled. Interaction terms between these effects were also studied.

5. The relative risk of infection is the same for each block, and similarly for each class, etc.

Tables 9.1 and 9.2 and Figure 9.2 give a general picture of the four schools in the study.

Table 9.1 gives some demographic details of cases in the four schools, and shows that the mean age and standard deviation of ages of pupils from whom *S. sonnei* was isolated are comparable between schools. Table 9.2 shows the age ranges and number of classes using toilet blocks. There was an average of 30 pupils per class.

9.3.5 Statistical modeling

In this section we derive maximum-likelihood estimates and standard errors of the relative risks of infection attributable to various risk factors and of the corresponding attributable fractions of infections.

Table 9.1 The numbers and age distribution of the study population

School	Staff	Pupils	Siblings	Mean Age of Pupils	Standard Deviation of Age of Pupils
1	3	49	10	6.38	1.93
2	0	31	3	5.77	2.12
3	0	33	4	6.94	2.22
4	0	56	13	5.78	1.81
Totals	3	169	30	6.19	2.05

Table 9.2 Usage of toilet blocks by classes in the four schools studied

School	Toilet Block	Sex	Age Range	Number of Classes
1	A1	Mixed	3-4	1
	A2	Mixed	4-7	4
	A3	Mixed	7-11	5
2	B1	Mixed	3-4	1
	B2	Male	5-11	9
	B3	Female	5-11	9
3	C1	Male	5-8	2
	C2	Female	6-8	2
	C3	Male	3-9	6
	C4	Female	3-11	8
4	D1	Male	3-4	1
	D2	Female	2-4	1
	D3	Male	4-5	6
	D4	Female	5-7	4
	D5	Male	8-10	4
	D6	Female	8-11	4

Risk markers and relative risks

Let the onset of morbidity for the N infected individuals occur at successive epochs $t_1 \ldots t_N$. These onset times start at the beginning of the infectious period. We are interested in the likelihood of observing some subset n of these (those who attend one of the four schools studied), and in the others purely because of their role as causative events of infection. Each individual can have any of q types of spatial proximity to infected individuals. Define risk markers

$$f_{kj}^{m} = \begin{cases} 1 & \text{if kth person is in type m proximity to the jth person} \\ 0 & \text{otherwise} \end{cases}$$

For example, proximity to the house of an infected child is a risk marker, because children play together, and we can take $f = 1$ if the Euclidean distance between their houses is less than some distance d. Define also

$$g(t_k, t_j) = \begin{cases} 1 & \text{if the jth individual could have infected jth} \\ 0 & \text{otherwise} \end{cases}$$

More precisely, for $g = 1$ we require that $t_k - t_j \geq \Delta$. Here Δ was taken as 4 days. When two infections occur simultaneously, it is not known who infected whom, and so we set $g(t_k, t_j) = g(t_j, t_k) = 1/2$. The problem of tied onset dates is discussed in detail later.

The hazard of infection of the kth individual at some epoch u is written

$$h_k(u) = \beta h_k'(u),$$

where β is the (unknown) transmission coefficient for dysentery, and the reduced hazard h' is modelled using the linear model:

$$h_k'(u) = \sum_{j=1}^{N} g(t_k, t_j)\left\{1 + \sum_{m=1}^{q} \alpha_m f_{kj}^{m}\right\}, \tag{13}$$

where the terms $\alpha_1+1, \ldots, \alpha_q+1$ are relative risks, unity if the corresponding risk marker (closeness to an infecter) has no effect.

This model assumes that infected individuals cause infection independently. The values of f and g will depend on one or more of several critical values such as the distance *d,* used to define proximity, and the duration of infection Δ.

9.3.6 The likelihood function

Following the method of derivation of likelihood functions outlined earlier, the likelihood function of observing infections at epochs $t_1 \ldots t_n$ is thus

$$L = \prod_{k=1}^{n} \{\beta h'_k(t_k)\} exp\left\{-\beta \sum_{l=1}^{n} \int_{\tau_l}^{t_l} h'_l(u)du\right\} \times P, \tag{14}$$

where *P* is the probability that no other individuals in the population are infected. Once the *k*th person becomes infected, data are retrospectively available over the period τ_k to t_k. The data are for the point when symptoms occur. This will be true for cohort and retrospective cohort or 'trohoc' studies, and for case-control studies. Here τ_k will be some epoch prior to the start of the epidemic.

Similarly, the 'weird' probability that the observed number *n* of infections happen to the particular individuals who were observed to be infected, but at any epoch within the period of observation, is

$$L_n = \beta^n \left\{\sum_{l=1}^{n} h'_l(u)du\right\}^n \Big/ n! \times exp\left\{\int_{\tau_l}^{t_l} h'_l(u)du\right\} \times P \tag{15}$$

The conditional likelihood $L_c = L/L_n$ is

$$L_c = \frac{n! \prod_{k=1}^{n} h'_k(t_k)}{\left(\sum_{l=1}^{n} \int_{\tau_l}^{t_l} h'_l(u)du\right)}. \tag{16}$$

The nuisance variables β and *P* have disappeared. One can also obtain equation (16) as a profile likelihood $sup_\beta L_n(\beta)$, by estimating β from equation (14) and substituting it back into equation (14). The estimate of β is

$$\hat{\beta} = n \Big/ \sum_{l=1}^{n} \int_{\tau_l}^{t_l} h'_l(u)du,$$

and *n*! is replaced by its large-sample approximation $n^n \exp(-n)$. Yet again, equation (16) can be derived by giving β an (improper) prior distribution with pdf $1/\beta$. On integrating over β from zero to infinity, we regain equation (16), but with $(n-1)!$ replacing $n!$. The constant is of course irrelevant for the subsequent inference.

Here L_c is the pdf that infections were experienced at the observed epochs rather than at some other possible epoch. In the denominator of equation (16) although infections as outcomes can occur at any epoch in the range, the infections as causative events are fixed at their observed epochs. Cause and effect have been decoupled, and we imagine infections occurring at new epochs, and their associated infecters still fixed as they were.

The logic leading to the partial likelihood (Cox regression) method [25] is very similar. There, one can condition the likelihood so as to remove the unknown function (in our notation) $\beta(t)$. The conditional likelihood used here is simpler, and loses less of the information from the likelihood in equation (15). It is appropriate for the short period of observation considered here.

We now develop equation (16). Evaluating the integrals, taking logs, and discarding the $n!$ factor, we obtain $\ell = \log L_c$ as

$$\ell = \sum_{k=1}^{n} log \sum_{j=1}^{N} g(t_k, t_j)\left\{1 + \sum_{m=1}^{q} \alpha_m f_{kj}^m\right\} - n\,log\left(1 + \sum_{m=1}^{q} p_m \alpha_m\right) - n\,log \sum_{l=1}^{n}\sum_{j=1}^{N} \omega_{lj} \tag{17}$$

where

$$\omega_{lj} = \int_{\tau_1}^{t_1} g(u, t_j)\,du, \tag{18}$$

and

$$p_m = \sum_{l=1}^{n}\sum_{j=1}^{n} \omega_{lj} f_{lj}^m \bigg/ \sum_{l=1}^{n}\sum_{j=1}^{n} \omega_{lj}.$$

Since g is a 0–1 function, $\omega_{lj} = \Delta$, the duration of infection of the jth infecter, unless the lth infection occurs before the infection period has finished, or the interval commences after it has begun to operate. It is zero if the jth infection occurred too late to have caused the lth infection. Thus p_m, which can be calculated from the data, is the probability that a random cause is 'close' in attribute-space to a random infection that it preceded, weighting causes by their periods of operation.

Maximising l with respect to α can be carried out numerically, and yields MLEs of relative risks. Analytically,

$$\partial l/\partial \alpha_r = 1/\alpha_r \times \sum_{k=1}^{n} F_{kr} - n/\alpha_r \times F_r,$$

where

$$F_{kr} = \frac{\sum_{j=1}^{n} g(t_k, t_j)\alpha_r f_{kj}^r}{\sum_{j=1}^{n} g(t_k, t_j)\left(1 + \sum_{m=1}^{q} \alpha_m f_{kj}\right)},$$

and

$$F_r = p_r \alpha_r \Big/ \left(1 + \sum_{m=1}^{q} p_m \alpha_m\right). \qquad (19)$$

Setting $\partial L/\partial \alpha_r = 0$ gives

$$1/n \sum_{k=1}^{n} \hat{F}_{kr} = \hat{F}_r. \qquad (20)$$

The $\hat{F}_{kr}$ are estimated fractions of the kth infection due to the rth risk marker, or estimated attributable fractions. Equation (20) is defining the estimated attributable fraction due to the rth risk marker as a sample mean, and the MLE of relative risk is the solution of the set of equations (19).

The covariance matrix for α may be estimated as the inverse of

$$-\partial^2 L/\partial \alpha_r \partial \alpha_s \big|_{\alpha=\hat{\alpha}},$$

which is trivially calculable. Since the score

$$\partial L/\partial \alpha_r = (n/\alpha_r) \times \left\{ \sum_{k=1}^{n} F_{kr}/n - F_r \right\},$$

and the covariance matrix of the score [25] is

$$-\partial^2 L/\partial \alpha_r \partial \alpha_s \big|_{\alpha=\hat{\alpha}},$$

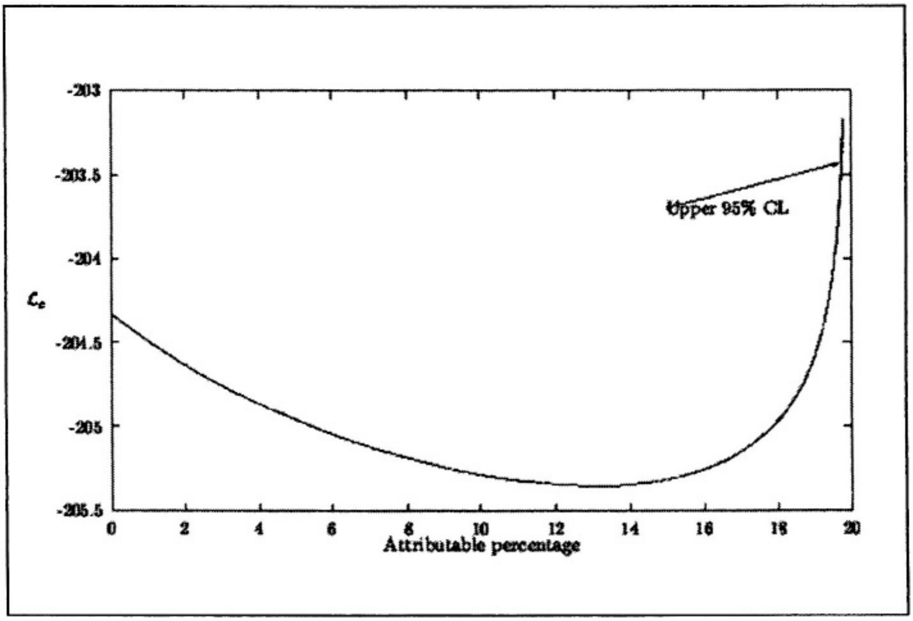
the MLE of the covariance matrix of the attributable fraction $\hat{F}_r$ is

$$-\left(\hat{\alpha}_r\hat{\alpha}_s/n^2\right)\partial^2 L/\partial\alpha_r\alpha_s\big|_{\alpha=\hat{\alpha}}=\left\{\sum_{k=1}^{n}\left(\hat{F}_{kr}-\hat{F}_r\right)\left(\hat{F}_{ks}-\hat{F}_s\right)\right\}\Big/n^2,$$

which turns out to be just the usual formula for the sample variance.

To obtain accurate confidence intervals on relative risks, for large samples one can plot the profile likelihood (all other parameters except the 'α' of interest are varied to maximise the likelihood). The 95% confidence limit is at the point where twice the log-likelihood has decreased by $(1.96)^2 = 3.92$. Confidence limits on the attributable fraction of infections from toilet blocks are shown in Figure 9.3.

Figure 9.3 Confidence limits on the attributable fraction of infections from toilet blocks



For small samples, the simplest (and most computer-intensive) method is to carry out bootstrap resampling to obtain a series of MLEs, $\hat{\alpha}'$. They are sorted, and the confidence interval taken between percentiles of the resulting sample distribution. These methods simultaneously give confidence intervals on the attributable fractions, i.e. by calculating these latter and sorting them, and proceeding as before.

When individuals differ in susceptibility because of age or other demographic variables, so that β varies, similar individuals may be grouped into n_b blocks, and conditional likelihood $L_c^{(i)}$ found for each block. The total conditional likelihood is $L_c = \prod_{i=1}^{n_b} L_c^{(i)}$. In this study, toilet blocks were taken as the unit of blocking, as one block is only used by children of similar age, and usually of the same sex.

Having estimated α by maximising the total conditional likelihood, the estimation of attributable fractions is also straightforward. The score is the sum over block scores, etc. Equation (20) stands, but equation (19) becomes

$$F_r^{(i)} = \left(n^{(i)}/n\right)\sum_{i=1}^{n_b} p_r^{(i)}\alpha_r \Big/ \left(1 + \sum_{m=1}^{q} p_m^{(i)}\alpha_m\right)$$

where all block-specific quantities have been given an upper suffix in parentheses. Thus the $p^{(i)}$ are now calculated using only infections in block *i*.

9.3.7 Results

The first question considered was the extent to which infection is acquired from schoolmates rather than simply from other children living in the same area. The model was fitted to onset dates for children attending Schools 1 and 4, which are just over 1 km apart. Infections could be from children attending the same school, the other school, or from 70 additional cases among children living in the area. Transmission of infection could be enhanced if the infecter lived within 3 km of the contact, attended the same school, or was a sibling of the contact.

Table 9.3 shows the results. The model parameter is ‘relative risk minus unity’; thus, it is zero if that risk factor has no effect. To illustrate the meaning of the table: the column labelled ‘*z*’ shows the relevant model parameter divided by its standard deviation; parameters with $z > 1.65$ will correspond to risk factors that significantly enhance disease transmission (one-sided test, 5% significance level). The second to the last column shows the estimated percentage of cases attributed to the corresponding risk factor.

Thus, children living in the same area were 185.8 times more likely to infect a contact than infecters who did not, and this effect is significant ($z = 3.91 > 1.65$). 45.7% of infections were attributed to this risk factor. Nearly 36% of infections arise because the infecter attended the same school as the contact. Hence more than half the infections do not occur at school.

Table 9.3 Results*

Risk Marker	RR-1	SE**	z	Attributed Fraction	SE**
Sibling	8945	1943	4.60	17.8%	4.6%
Same area	184.8	47.2	3.91	45.7%	4.4%
Same school	450.3	155.8	2.89	36.3%	3.2%

* RR denotes relative risk. The '*z*' value is *RR-1* divided by its estimated standard error. This table is based on 105 infections in Schools 1 and 4 from school-attenders and potential neighborhood contacts.
** SE denotes standard error.

Table 9.4 Results*

Risk Marker	RR-1	SE**	z	Attributed Fraction	SE**
Toilet block	1.65	1.06	1.56	13.0%	1.3%
Class	1.45	1.78	0.81	3.3%	0.7%
Sibling	33.8	7.36	4.60	13.2%	3.2%
Similar age	1.67	1.02	1.64	15.5%	1.5%
Live nearby	1.92	0.56	3.46	31.2%	2.5%

* This table is based on 170 infections among school-attenders caused by schoolmates.
** SE denotes standard error.

The next step is to examine infections acquired purely from schoolmates in all four schools. Table 9.4 shows the results of fitting the model to onset dates classified by school, toilet block, class, and neighborhood.

Extra infections attributed to sharing a toilet block with an infected individual are only 13% of the total infections acquired from schoolmates, and presumably, as mentioned earlier, some of this effect may not be due to direct contact with infected toilets. Figure 9.3 shows the upper 95% confidence limit on the attributable fraction obtained from the profile likelihood using the factors in Table 9.4. The lower 95% limit would lie below zero. This effect is not in fact statistically significant, and so may even be entirely absent. Hence the level of hygiene currently prevailing during outbreaks is certainly adequate, and little would be gained by extra spending on disinfection.

The effect due to sharing a classroom is also small at 4% and not quite statistically significant.

The sibling effect is significant at more than four standard deviations, and siblings are very many times more likely to infect each other than non-siblings. However, less than 14% of infections arise in this way, as each contact has on average few siblings, but many schoolmates.

Again, the parameter measuring increased transmission between children within a year of the same age also hovers near statistical significance. Such an effect if present would account for 15.5% of infections.

Enhanced transmission between children who live close to each other certainly occurs ($z > 1.65$). This suggests that infection does not occur only on school premises.

Finally, Table 9.5 shows the results of including those factors identified as important, returning to data from only the two schools in the same area first considered. Again, infections acquired from school toilets and from classmates comprise only 4% of the total. There may be increased contact with other children of the same age and attending the same school, and one fifth of all infections arise from schoolmates living nearby (within 0.75 km). This increased risk did not occur with children from other schools living nearby.

It therefore seems possible that relatively few infections actually happen on school premises, but that attending a common school facilitates contact between children outside school hours, e.g. near their homes. From Table 9.5 one could attribute an upper bound of 3.8 + 2.3 = 6.1% of infections to contacts occurring on school premises, excluding contacts with children of the same age, or 3.8 + 2.3 + 21.1 = 27.2% including the latter. Not all of these necessarily occur on school premises, but at most about a quarter of infections can occur there. Hollins [45] reached a similar conclusion, that much infection is spread outside the school environment between neighboring families.

9.3.8 Conclusion on the role of the schools in dysentery transmission

The role of schools in spreading infection is of great interest to the public. Out of the total of 36% of infections that could be attributed to schools, perhaps only 6-27% of infections may arise on school premises. Since the four schools with the greatest incidence of *S. sonnei* were studied, the true attributable fraction of infections may well be even less than this figure.

Table 9.5 Analysis using only those factors identified as important*

Risk Markers	RR-1	SE**	z	Attributed Fraction	SE**
Toilet block	5.0	16.3	0.31	3.8%	0.46%
Class	9.7	21.6	0.45	2.3%	0.58%
Sibling	301	68.0	4.43	17.2%	4.5%
Across sexes	4.2	3.1	1.37	15.3%	1.7%
Same area	1.85	1.23	1.51	13.2%	1.8%
Similar age and school	22.3	14.5	1.54	21.1%	2.3%
Live nearby and same school	12.2	7.4	1.65	19.8%	2.0%

* The table is based on 105 infections in Schools 1 and 4 from school-attenders and potential neighborhood contacts.
** SE denotes standard error.

The majority of infections are spread between children living in the same area, between siblings, or between schoolmates living close to each other. This last finding suggests that school attendance mediates children's social contacts; children do not often acquire infection from children attending other schools, even if they live nearby.

Closing schools during outbreaks of dysentery causes considerable economic loss. It would at best reduce the rate of infection by a quarter, and might even increase it, if children spend the freed time playing with neighboring children.

School toilets are often thought by the public to be the root of all evil where dysentery infection is concerned. In this study, no statistically significant amount of infection could be attributed to this source. This finding thus does not support the conclusion of Hutchinson [46] that toilets represent a major risk factor.

There would seem little point in attempting to reduce infections occurring within the classroom. It seems that sharing a classroom with an infected child is not a risk factor. This is perhaps not surprising, as there will be little body contact in the classroom; it suggests that environmental contamination via fomites, etc. is negligible. However, one fifth of infections arise because of contact with those living nearby (within 0.75 km). It might be worthwhile encouraging parents to keep their children away from possible infecters during outbreaks, or to discourage their infected children from playing with others.

9.4 AVENUES FOR FURTHER RESEARCH

The practical aim of epidemiological research is to identify risks to public health, and to present reasoned evidence such that preventative action will be taken.

We are moving into a time when great amounts of data and great computing power will be available to the researcher. There is a need for the science of statistics itself to develop in the area of model choice when very many models are considered by the 'data miner'. It is also beginning to be necessary to evaluate the many different models and tests available in order to produce 'good practice' guidelines.

The research presented in this chapter is based firmly on statistical principles. The rallying cry of 'back to orthodoxy' will never be popular, but we must not lose sight of basic principles in the welter of new possibilities opened up by increased computing power.

Imaginative methods of modelling and of carrying out statistical inference, such as Cox's partial likelihood, are required, in order to continue to make valid inferences about risk in the presence of confounding variables. Sophisticated numerical methods and algorithms are also needed to enable the rapid computation needed for epidemiological studies in the 21st century.

References

[1] Bates, D.V. (1994). *Environmental Health Risks and Public Policy.* University of Washington Press, Seattle, WA.

[2] Peña, M. and J. Bacallao (2000). *Obesity and Poverty: A New Public Health Challenge.* Pan American Health Organization Publications, Washington, DC.

[3] Viel, J.F., S, Wing and N. Hoffmann (1999). Environmental epidemiology: Public health advocacy and policy. In A. Lawson, A. Biggeri and D. Bohning, *Disease Mapping and Risk Assessment for Public Health,* Wiley, New York.

[4] Smith, C.W. and R.D. Baker (1982). Comments on the paper 'Environmental Power-Frequency Magnetic Fields and Suicide' (Letter to the editor). *Health Physics,* 3, 439-441.

[5] Li, D., R. Odouli, S. Wi, T. Janevic, et al. (2002). A population-based prospective cohort study of personal exposure to magnetic fields during pregnancy and the risk of miscarriage. *Epidemiology,* 13, 9-20.

[6] Lee, G.M., R.R. Neutra, L. Hristova, M. Yost, and R. A. Hiatt (2002). A nested case-control study of residential and personal magnetic field measures and miscarriages. *Epidemiology,* 13, 21-31.

[7] Ahlbom, A., N. Day, M. Feychting, et al. (2000). A pooled analysis of magnetic fields and childhood leukaemia. *British Journal of Cancer,* 83, 692-698.

[8] Beale, I.L., N.E. Pearce, R.J. Booth and S.A. Heriot (2001). Association of health problems with 50-Hz magnetic fields in human adults living near power transmission lines. *Journal of Australasian College of Nutritional and Environmental Medicine,* 9-12.

[9] Dye, C., M.A. Espinal, C.J. Watt, C. Mbiaga, and B.G. Williams (2002). Worldwide incidence of multidrug-resistant tuberculosis. *Journal of Infectious Diseases,* 185, 1197-1202.

[10] Anto, J.M., J. Sunyer, C. E. Read, J. Sabria, et al. (1993). Preventing asthma epidemics due to soybeans by dust-control measures. *New England Journal of Medicine,* 329, 1760-1763.

[11] Lawson, A., A. Biggeri, and D. Bohning, Eds. (1999). *Disease Mapping and Risk Assessment for Public Health.* Wiley, New York.

[12] Stein, C.E., S. Bennett, S. Crook, and F. Maddison (2001). The cluster that never was: Germ warfare experiments and health authority reality in Dorset. *Journal of the Royal Statistical Society Series A,* 164, 23-28.

[13] Steward, J. and G. John (2001). An ecological investigation of the incidence of cancer in Welsh children for the period 1985-1994 in relation to residence near the coastline. *Journal of the Royal Statistical Society Series A,* 164, 29-44.

[14] Kulldorff, M. (2001). Prospective time periodic geographical disease surveillance using a scan statistic. *Journal of the Royal Statistical Society Series A,* 164, 61-72.

[15] Hill, A.B. (1971). *Principles of Medical Statistics (9th Ed.).* Oxford University Press, New York.

[16] Last, J.M. (1995). *A Dictionary of Epidemiology (3rd Ed.).* Oxford University Press, Oxford.

[17] Durkheim, E. (1897). *Le Suicide: Etude de Sociologie.* Alcan Publishers, Paris.

[18] Wakefield, J. and R. Salway (2001). A statistical framework for ecological and aggregate studies. *Journal of the Royal Statistical Society Series A,* 164, 119-138.

[19] Burnham, K.P. and D.R. Anderson (1998). *Model Selection and Inference: A Practical Information-Theoretic Approach.* Springer, New York.

[20] Draper, D. (1995). Assessment and propagation of model uncertainty (with discussion). *Journal of the Royal Statistical Society Series B,* 57, 45-97.

[21] Breslow, N.E. (1996). Statistics in epidemiology: The case-control study. *Journal of the American Statistical Association,* 91, 14-28.

[22] Tarone, R.E. (1985). Score tests. In S. Kotz, and N. L. Johnson (Eds.), *Encyclopaedia of Statistical Sciences,* Wiley, New York.

[23] Andersen, P.K., O. Borgan, R.D. Gill and N. Keiding (1993). *Statistical Models Based on Counting Processes.* Springer, New York.

[24] Lawson, A.B. (2001). *Statistical Methods in Spatial Epidemiology.* Wiley, New York.

[25] Cox, D.R. and D. Oakes (1984). *Analysis of Survival Data.* Chapman and Hall, London.

[26] Breslow, N.E. and N.E. Day (1987). *Statistical Methods in Cancer Research. Volume 2 – The Design and Analysis of Cohort Studies.* Oxford University Press, New York.

[27] Collett, D. (1991). *Modelling Binary Data.* Chapman and Hall, London.

[28] Knox, E.G. (1964). The detection of space-time interactions. *Applied Statistics,* 13, 25-29.

[29] Knox, E.G. (1964). Epidemiology of childhood leukemia in Northumberland and Durham. *British Journal of Preventive Social Medicine,* 18,17-24.

[30] Becker, N.G. (1989). *Analysis of Infectious Disease Data.* Chapman and Hall, London.

[31] Davies, R.B. (1977). Hypothesis testing when a nuisance parameter is present only under the alternative. *Biometrika,* 64, 247-254.

[32] R.B. Davies (1987). Hypothesis testing when a nuisance parameter is present only under the alternative. *Biometrika,* 74, 33-43.

[33] Cox, D.R. and D.V. Hinkley (1974). *Theoretical Statistics.* Chapman and Hall, London.

[34] Kulldorff, M. and U. Hjalmers (1999). The Knox method and other tests for space-time interaction. *Biometrics,* 55, 544-552.

[35] Lehmann, E.L. (1975). *Nonparametrics: Statistical Methods Based On Ranks.* Holden-Day, San Francisco, CA.

[36] Baker, R.D. (1996). Testing for space-time clusters of unknown size. *Journal of Applied Statistics,* 23, 543-554.

[37] Mantel, M. (1967). Detection of disease clustering and a generalised regression approach. *Cancer Research,* 27, 209-220.

[38] McHardy, J., E.H. Williams, A. Geser, G. Dethe, E. Beth, and G. Giraldo (1984). Endemic Kaposi's sarcoma: incidence and risk-factors in the West Nile district of Uganda. *International Journal Of Cancer,* 33, 203-212.

[39] Barton, D.E. and F.N. David (1966). Random intersection of two graphs. In F.N. David (Ed.), *Research Papers in Statistics,* Wiley, New York.

[40] Hopkins, T. and C. Phillips (1988). *Numerical Methods in Practice Using the NAG Library.* Addison-Wesley, Wokingham, UK.

[41] PHLS working group on the control of *Shigella sonnei* infection (1993). Revised guidelines for the control of *Shigella sonnei* infection and other infective diarrhoeas. *CDR Review,* 5, R69-R70.

[42] Thomas, M.E.M. and H.E. Tillett (1973). Sonnei dysentery in day schools and nurseries: An eighteen-year study in Edmonton. *Journal of Hygiene,* 593-602.

[43] DuPont, H.L., R.B. Hornick, A.T. Dawkins, M.J. Snyder, and S.B. Formal (1969). The response of man to virulent *Shigella flexneri* 2a. *Journal of Infectious Diseases,* 119, 296-299.

[44] Keusch, G.T. and M.L. Bennish (1991). Shigellosis. In Evans, A. S. and Brachman, P. S. Eds., *Bacterial Infections of Humans (2nd ed.),* Plenum Publishers, New York.

[45] Hollins, F.R. (1970). Sonne dysentery in primary schools. *Medical Officer,* 346-349.

[46] Hutchinson, R.I. (1956). Some observations on the method of spread of sonne dysentery. *Mon. Bulletin of Health,* 110-118.

10 MODELING HEALTH OUTCOMES FOR ECONOMIC ANALYSIS

Thitima Kongnakorn[1] and François Sainfort[1]

[1] School of Industrial and Systems Engineering
Georgia Institute of Technology
Atlanta, GA 30332

SUMMARY

Measuring health outcomes is critical for individual and societal decision making. This chapter briefly reviews the field of health outcomes modeling in general and provides detailed theoretical background for one specific class of such models, the Quality-Adjusted Life Years model, which is primarily grounded in operations research and utility theory. The chapter describes methodological issues and concludes with a discussion of promising areas for further research.

KEY WORDS

Health-related quality of life, Medical decision analysis, Cost-effectiveness analysis, Utility theory

10.1 INTRODUCTION

The measurement of health outcomes is a critical matter in medical decision making. When clinicians and patients make clinical decisions such as choosing among alternative medical treatments, they base at least part of their judgment on their perceptions of relative gains or losses in future health. The existence of a good metric, or quantitative system, for measuring future health resulting from alternative treatments would greatly facilitate the process of making such decisions. The ultimate goal of medical treatment is not to improve a particular clinical parameter, to eliminate particular symptoms, or to cut costs, but to improve health of patients. There is little dispute that improving health, in medicine, involves two main components: increasing life expectancy or "length of life" and increasing "quality of life" of patients [1]. Clinical outcomes defined in terms of mortality or physiological measures such as blood pressure or intermediary diagnostic test results, are often necessary, but insufficient, for making a final treatment decision. Patients' preferences for health outcomes need to be captured and explicitly included when contrasting and evaluating alternative treatments for making medical decisions. Any health outcome measure would need to account, in some way, for both length and quality of life.

Similarly, at the population level, capturing and aggregating those preferences is also often deemed necessary for evaluating new treatments, health services or medical technology. Failure to include such information may result in suboptimal decisions that do not conform to individual or societal preferences. For example, in cost-effectiveness analysis, a standard tool used in health economics, the costs and benefits of one health intervention are compared with costs and benefits of another by calculating the incremental cost-effectiveness ratio, which expresses the cost per additional unit of health benefit conferred for one intervention compared to another [2]. In such a model, the complete elicitation and estimation of relevant costs and the most representative and accurate measure of health benefits, or effectiveness, are needed. If a goal is to permit comparisons across diseases or conditions, health benefits can be expressed in generic terms such as "health-adjusted life years" (HALYs), as opposed to disease- or condition- specific terms (such as number of specific cases averted).

HALYs can be viewed as a large field encompassing a number of measurement systems, which differ in at least three overall dimensions: (a) disease-specific versus generic measures; (b) non-preference versus preference-based measures; and (c) use for individual versus societal decision making. As mentioned before, a generic measure permits comparison of health benefits across diseases or conditions and is not naturally tied to a certain disease or condition (as would be the case with physical measures such as blood pressure or total cholesterol level or a condition-specific rating scale such as a scale measuring back pain). As noted by Fryback [1], another fundamental difference between measurement systems is whether the numbers generated reflect individual preferences for different health states and thus are derived from human judgment about the relative desirability of being in one health state versus another, or are derived in a manner not directly related to preferences. For example, the eight scales of the short-form health survey SF-36™ [3] produce numbers that do not reflect preferences. Utility-based models such as the Health Utility Index [4], on the other hand, are specifically designed to reflect preferences. Finally, it is important to note that measures designed to support individual decision making may or may not lend themselves to aggregation across individuals in a population to assist in societal decisions. Thus, in terms of the applicability and validity of measurement systems, it is important to consider the viewpoint being adopted. Nord et al. [5], for example, have identified a number of limitations in aggregating individual measurements of health-related quality of life for assessing the societal value of health care investments and have proposed adjustments for dealing with such problems.

A number of measurement systems have been developed by researchers from many different disciplines. In this chapter, we primarily review selected contributions of operations researchers, economists and psychologists who developed one of the most widely used, and criticized, class of HALYs – the *quality-adjusted life years* (QALYs). The QALY model is a generic, preference-based measurement system designed to assist in individual decision making. It is widely used for societal decision making, provided that its limitations are properly dealt with [5, 6]. In this chapter, we review some of the literature, present major methodological issues, and identify promising areas of research.

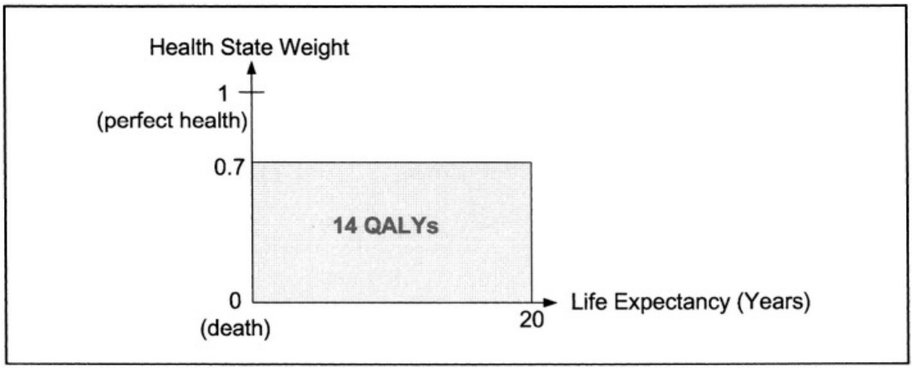
10.2 QALY MODEL – THEORETICAL CONSIDERATIONS

10.2.1 Background

The concept and techniques of utility theory have been applied for health outcome measurement in order to incorporate patients' preferences and risk attitudes. Such utility measurement techniques have been developed and applied, to a large degree, within the context of "chronic health states". A chronic health state is generally defined as a health state that stays constant over a relatively long period of time (typically more than one year). Most real-life situations, however, challenge the assumption of a *constant* health state. Chronic diseases, even when treated, are generally not stable but lead to health status deterioration over time. Health states generally do not remain at the same level over lengthy periods of time even in healthy individuals, for whom decrements are expected with the normal aging process.

The most widely applied model for health outcome measurement in medical decision analysis is the quality-adjusted life year (QALY) approach. The QALY model has emerged as the gold standard for health outcome measurement [7]. Both life expectancy and quality of life are taken into account in a QALYs measure. The number of QALYs is typically obtained by multiplying life expectancy by a numerical weight associated with a constant health state experienced during the remaining life expectancy. The weight is a number between 0 and 1 where 0 is defined as "death" and 1 as "perfect health". On this scale, the weight associated with a health state represents the health-related quality of life (HRQOL) of such health state. The product of the HRQOL weight and the life expectancy is a measure of the desirability of the health state experienced during the life expectancy. For example, as shown in Figure 10.1, the health of an individual who has a life expectancy of 20 years with a disease that has a HRQOL weight of 0.7 is valued at 20 × 0.7 = 14 QALYs.

Figure 10.1 Illustration of QALYs in the case of constant health state

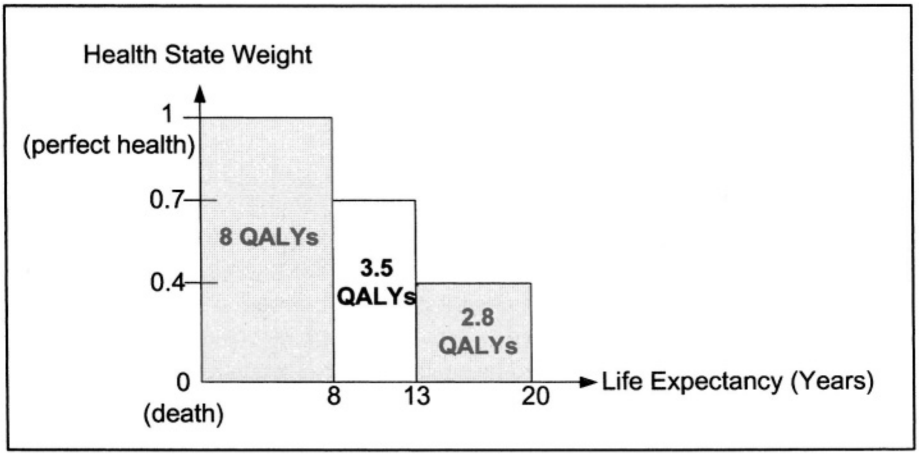


Extending the approach to sequences of chronic health states such as the sequence shown in Figure 10.2, one typically calculates the desirability of such a sequence by taking the sum of all products of duration and health weight corresponding to the health states in that sequence. For example, an individual with a health profile shown in Figure 10.2 would value that sequence at [(1×8) + (0.7×5) + (0.4×7)] = 14.3 QALYs.

10.2.2 Theoretical foundation – Risk neutral QALY model

The quality-adjusted life year (QALY) model is a measurement technique for health outcomes that takes into account both quality and quantity of life.

Figure 10.2 Illustration of QALYs in the case of non-constant health profile



It is the product of life expectancy and a utility-based measure of quality of life of the remaining years of life. The QALY method was developed in the 1970's [8]. The original theoretical properties of the QALY measure are summarized in a paper by Pliskin et al. [9]. They show that QALY is a valid utility function, which represents individual preferences, if three conditions hold. These conditions are as follows.

1. Mutual utility independence (MUI) of life years *(T)* and health state (*Q*) This assumption means that preferences for gambles over either one of the two attributes, with the other attribute held at a fixed level, do not depend on the particular level of that other attribute. For example, an arthritis patient does not judge his own health state differently because he has five or 20 years remaining in his life. If MUI holds, one can construct a multiattribute utility model for the health profile *(Q,T)* as follows:

$$U(Q,T) = a \cdot U(Q) + b \cdot U(T) + (1\text{-}a\text{-}b) \cdot U(Q) \cdot U(T)$$

where *U(Q,T)* is the utility of health profile *(Q,T); U(Q)* is the utility of health state *Q*; *U(T)* is utility of life years *T*; *a* and *b* are scaling constants.

2. Constant-proportional tradeoff property This requires that the proportion of the remaining life that one would trade-off for a specified quality improvement is independent of the actual amount of the remaining life. For instance, consider the situation where one asks an individual to trade off an amount of time of his/her remaining years of life in order to have perfect health versus the poorer health state. If he/she gives up 10 years out of 20

remaining life years, he/she would equally give up 5 out of 10, 2.5 out of 5, and so on. Thus, the proportional trade-off is constant, in this case always exactly half of his/her remaining life years.

3. Risk neutrality over life years This assumption means that the utility function for life years is linear. If risk-neutrality over life years holds in all health states, MUI and constant proportional trade-off will also hold [10].

10.2.3 Theoretical foundation – Risk adjusted QALY model

The above three assumptions are the requirements for the standard QALY model which assumes risk neutrality with respect to life duration and hence assures linearity of the component utility function over life years. However, the assumption of linearity is not empirically realistic. For example, McNeil, Weichselbaum, and Pauker [11] found that patients with bronchogenic carcinoma had moderate risk aversion over life years. Stiggelbout et al. [12] found mild risk aversion in male patients with testicular cancer. Additionally, Verhoef et al. [13] conducted a study with healthy women and found risk aversion over life years, but risk-seeking preferences over gambles involving short durations. On the contrary, in a different health context, Mehrez and Gafni [14] found risk aversion when the length of the durations increased. Thus, the violation of risk neutrality in the standard QALY model would lead to invalidity of QALY as a representation of an individual's preferences.

However, QALYs can be defined in either a risk-neutral (standard QALY model) or a more general risk-adjusted form (generalized QALY model), depending on whether the decision maker is risk neutral or not with respect to uncertainty regarding life years. If the decision maker is risk neutral with respect to life years, then QALYs will be decomposed in the following form:

$$\textbf{Risk-neutral QALYs} = U(Q,T) = H(Q) \times T$$

The more general risk-adjusted QALYs are defined as follows:

$$\textbf{Risk-adjusted QALYs} = U(Q,T) = H(Q) \times [T]^{r}$$

In both formulations, *H(Q)* is the quality weight, measured on a scale between 0 (death) and 1 (full health), and *r* is the risk parameter that defines the shape of the utility function for quantity of life. If the subject is risk neutral, $r = 1$. If mutual utility independence and constant proportional trade-off hold, then risk-adjusted QALYs, as defined by Pliskin et al. [9], are a valid utility function representing preferences over constant health states [10].

10.2.4 Theoretical foundation – The zero-condition

Instead of the three assumptions established by Pliskin et al. [9], which require knowledge of concepts from utility theory, Bleichrodt et al. [15] suggested a more elementary and fundamental characterization of QALYs that can relax Pliskin et al.'s [9] assumptions. They found that risk neutrality together with the "zero-condition" are sufficient to imply the existence and validity of the QALY model. The "zero-condition" indicates that all health state levels are equivalent, from a quality-of-life perspective, to a zero duration of life years. The zero-condition seems unavoidable in the medical context. Thus, the only assumption that is needed to imply the existence and validity of the QALY model is the risk neutrality for *all* health states.

However, there is ample empirical evidence showing a violation of risk neutrality as previously described. A generalized QALY model that can relax the assumption of risk neutrality has been established to solve the risk neutrality issue. A generalized QALY model has the following form:

$$U(Q,T) = V(Q)W(T),$$

where $U(Q,T)$ is the utility of the health profile (Q,T), $V(Q)$ is the value or utility function over health states Q, and $W(T)$ is the function that values life duration and can be nonlinear, with $W(0) = 0$. Instead of the risk neutrality condition, Miyamoto et al. [16] suggested another condition, "standard gamble invariance" (SG invariance). SG invariance basically says that, if Q and Q' are unequal to death and p is the probability equivalent of (Q,T) with respect to (Q,Y) and (Q,Z), then p is also the probability equivalent of (Q',T) with respect to (Q',Y) and (Q',Z) [16]. Without risk neutrality, a generalized QALY model holds if and only if both zero-condition and SG invariance hold.

10.3 METHODOLOGICAL ISSUES

The QALY model is subject to a number of methodological issues. Those include both theoretical and practical issues. Practical issues, especially in terms of development and use of alternative utility assessment methods suitable for eliciting and constructing the utility function over health states Q, have been discussed elsewhere (see, for example, [17]). Current popular methods include the visual analog scale (VAS), time-tradeoff (TTO), and standard gamble (SG) techniques. In the following sections, we primarily focus on theoretical issues.

10.3.1 Validity of MUI and constant proportional tradeoff

Besides challenging the risk neutrality assumption, several studies on the validity of the other assumptions have been preformed. For example, Miyamoto and Eraker [18] tested the mutual utility independence assumption and found empirical support for this assumption. Bleichrodt and Johannesson [19] performed empirical tests on both the utility independence and constant proportional tradeoff assumptions. They found that without adjustment for imprecision of preference (imprecision adjustment was suggested because of the unfamiliarity of the subjects regarding both the health states and the elicitation methods), 22.8% of the subjects satisfied the constant proportional tradeoff assumption, 13.4% satisfied utility independence, and 5.8% satisfied both assumptions. However, with the imprecision adjustment, 90.1%, 75.8% and 88.8% of the subjects satisfied constant proportional tradeoff only, utility independence only, and both assumptions, respectively. The authors concluded that the constant proportional tradeoff holds roughly and utility independence holds, but in a much weaker way. Pliskin et al. [9] reported 25 pairs of time-tradeoff responses from 10 subjects in hypothetical questions concerning the relief of different levels of anginal pain. They found that only four out of 25 pairs were consistent with the constant proportional tradeoff assumption.

10.3.2 Validity of utility theory

Expected utility theory or the von Neumann-Morgenstern expected utility theory is the foundation for most health outcome assessment and measurement techniques. A utility function exists when certain axioms hold. Three axioms of expected utility theory [20], known as normatively compelling rules for rational decisions under uncertainty, are as follows. Here, X is a set of outcomes; $\Delta(X)$ is the set of probability distributions over X; $\succ$ denotes an individual's preference relation over probability distributions; and ~ denotes the indifference relation over probability distributions.

1. Weak order

$\succ$ is asymmetric ($p \succ q \rightarrow$ not $[q \succ p]$) and both $\succ$ and ~ are transitive (if $p \succ$ (or ~) q and $q \succ$ (or ~) r then $p \succ$ (or ~) r for all $p, q, r \in \Delta(X)$).

2. Independence

For all $p, q, r \in \Delta(X)$ and any $\alpha \in [0, 1]$, then $p \succ q$ if and only if $\alpha p + (1\text{-}\alpha)r \succ \alpha q + (1\text{-}\alpha)r$.

3. Continuity Axiom

For all $p, q, r \in \Delta(X)$ such that $p \succ q \succ r$, then there exist α and $\beta \in [0, 1]$ such that $\alpha p + (1-\alpha)r \succ q \succ \beta p + (1-\beta)r$.

Thus, in addition to the three required assumptions of QALYs described previously, how adequately QALYs represent preferences over health states also depends on whether QALYs are consistent with von Neumann and Morgenstern's expected utility theory. If the axioms of von Neumann and Morgenstern's expected utility theory hold true, decision makers should be able to make decisions that are consistent with their underlying preferences. However, in medical decision making, as in many other application domains, violations of all three axioms have been shown and are well known [21]. These include Allais' paradox and Ellsberg's phenomenon and are not reviewed here. Instead, we focus our attention to a more important problem, developing a proper decomposition for multistate health profiles as shown in Figure 10.2.

10.3.3 QALYs for multistate health profiles

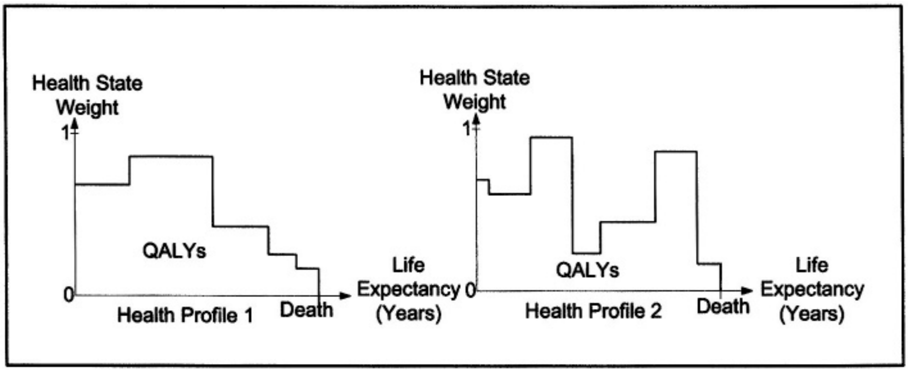
In the case of multistate health profiles, QALYs are generally calculated as the sum of all products of duration and health preference weight for all health states representing the health profile. Bleichrodt [22] has shown that for such decomposition to hold, the assumption of additive independence must hold. In essence, additive independence requires that the preference for one health state be independent of preference for other health states in the multistate health profile.

For example, consider the two multistate health profiles depicted in Figure 10.3. Both Health Profile 1 and Health Profile 2 have been designed to produce the same amount of QALYs (the area under each curve). The two profiles are clearly different, yet the QALY model would rate them as equally preferred. Some individuals, however, may have a preference for one pattern over another. Many potential factors that define the pattern of health profiles might affect an individual's preferences. The QALY framework currently fails to account for these factors.

10.3.4 Violation of additive independence

Several empirical studies have explored the validity of the additive independence assumption. Richardson et al. [23] examined the validity of the additive QALY model in a 16-year post-mastectomy health profile represented by a gradual deterioration and three health states: moderate side effects during the first five years, mild side effects for the next 10 years, but then breast cancer would recur and the patient would experience severe side effects during the last one year. Sixty-three female respondents participated in the study. Rating scale, time-tradeoff and standard gamble techniques were used to assess utility for each health state and the holistic utilities for the health profiles. Preference scores from constituent states were combined to estimate scores for the health profile using a discount rate of 3% and 9%. They found that holistic preferences for the multistate health profile (whether assessed with a rating scale, time-tradeoff, or standard gamble) were significantly different from composite preferences derived from the constituent health states, irrespective of the discount rate applied.

Figure 10.3 Two multistate health profiles with equal amount of QALYs



Kupperman et al. [24] also investigated whether preferences for multiphase health states can be approximated by preferences from constituent health states. One hundred and twenty-one female subjects were asked to assess their preferences for eight health profiles, each composed of three to four health states, in the context of prenatal diagnosis choices (chorionic villus sampling and amniocentesis), by using visual analog scaling and standard gamble techniques. The authors explored whether a different statistical formulation could be derived to predict preference scores for health profiles from their constituent health states preference scores. They found that a duration-weighted additive model, as used in the conventional QALY model, was not predictive. A multiple regression model that derived from statistically inferred weights predicted the preferences for the profiles better than the duration-weighted model.

MacKeigan et al. [25] used the time-tradeoff technique to compare preference scores for the same lifetime paths between holistic and composite assessment. One hundred and one participants with type 2 diabetes assessed their preferences regarding four hyperglycemic treatment profiles lasting 30 years, composed of eight discrete treatment states. The authors failed to find any significant differences between holistic and composite scores, which conflicted with the results from the studies by Richardson et al. [23] and Kupperman et al. [24]. However, the health profiles used in MacKeigan et al. [25] were different in that they consisted of progressive minor deteriorations in states while the health profiles in the other studies consisted of critically different health states. The authors noted that another reason why they found no difference between composite and holistic scores was because the profiles in the study were too similar. They recommended that future research be repeated with profiles that are more distinct and with sequencing effects that are more pronounced.

In Spencer's study [26], three health states defined with the EuroQol classification system [27] were used in each multistate health profile. Each health profile in the study had a 10-year duration and contained three different health states with durations of three, three, and four years respectively. Two tests were conducted: a test of additive independence and a test of the overall additive model. Twenty-nine subjects participated in the study. The violation of additive independence was found in the first test. However, in the additive model test, only one of the two versions resulted in a rejection of the additive model. Thus, Spencer could not conclusively reject the additive model. The author suggested that a larger sample size might allow the test to be able to detect significant differences in the results. Also, comparisons of utilities based on holistic elicitation procedures and constituent states elicitation were performed. The results showed that two out of the seven profiles exhibited a significant difference between holistic and constituent states elicitation, which implied that the additive independence assumption was violated.

10.4 FUTURE RESEARCH DIRECTIONS AND CONCLUSION

The studies previously described clearly show the violation of the additive independence assumption. Thus, the additive decomposition for the multistate health profile does not work and does not come close to an acceptable estimation. Therefore, it is critical to investigate and formulate an alternative decomposition.

A number of studies (some within the health domain, others in different domains) have explored or identified characteristics that affect people's preferences for multistate profiles. These influential factors could lead to, and partially explain, the violation of the additive independence assumption. A review of the studies exploring such influential factors is given below.

10.4.1 Empirical studies

Rate of Change Hsee and Abelson [28] performed experiments to find a relationship between satisfaction (utility) and *rate of change* of the outcomes or what they called *velocity* in the contexts of gambling (probability of winning the game), class rank (the percentile standing in a hypothetical class), and stock (a hypothetical stock price). They found satisfaction to be positively related to actual outcome position and rate of change (or velocity) of the outcomes over time.

Chapman [29] rated ten sequences that had five different slopes for two overall trends (increasing or decreasing) in health and money domains using a 0 to 100 visual analog scale. Slope (rate of change) was found to be one of the significant factors impacting their rating scores. Subjects preferred gradually increasing or decreasing sequences to those with steep slopes.

These results were in conflict with the findings by Hsee and Abelson [28], which suggested that subjects preferred steep slopes for increasing sequences but small slopes for decreasing sequences. However, Hsee and Abelson did not control for the total number of units of outcome over a specific period of time while Chapman did. Thus, preference for higher rate of change in positive outcome in the findings by Hsee and Abelson might be the result of a higher amount of outcomes received within the specific period of time. Ariely [30] also found a significant effect of rate of change in a study of retrospective pain evaluation in the experience of heat stimuli on the forearms. The results showed evidence of a rate-of-change effect, as the subjects reported experiencing higher pain when the intensity steeply increased than when it gradually increased.

Trend Several empirical studies found a significant impact of *trend* of the overall profile (improvement versus decrement) on preferences [29-36]. For example, Chapman [34] explored preferences for improving or declining sequences in the domains of headache pain, athletic ability, facial acne and facial wrinkles. Those sequences were designed so that, if the additive assumption held, they should be equally preferred. She found that subjects strongly preferred the improving sequences to the declining ones. Moreover, Chapman [29] explored preferences for both sequences of health and monetary outcomes and found that subjects preferred improving sequences for both health and money for short sequences (1 year) whereas for the long sequences (lifetime), subjects preferred decreasing sequences for health but increasing sequences for money. She explained that the subjects preferred the decreasing sequences for lifetime health since they used their expectation as a reference point and exerted judgment by considering how close the profile in question was to their reference point. When considering a long time horizon such as a lifetime, subjects expect perfect health early and gradually declining health as they get older.

Loewenstein and Sicherman [31] also showed that a majority of subjects preferred an increasing sequence of wage profiles over a five-year period to a declining sequence. In a very different domain, Loewenstein and Prelec [32] found that a majority of subjects who reported having a preference for a French restaurant over a Greek restaurant also reported a preference for a dinner at a Greek restaurant first and at a French restaurant later, thus showing a preference for an improvement trend.

Spreading of Outcomes Loewenstein and Prelec [37] found that decision makers prefer outcomes that are *spread* across the time interval considered. For example, the majority of the subjects who were offered two free dinners preferred to distribute the two dinners across the time interval. This preference for spreading was confirmed by Chapman [38] who performed a study involving scenarios including both gains and losses in the contexts of monetary outcomes (win a prize or pay a fine), dinner (pleasant or unpleasant dinner), and health-related events (a painful trip to the dentist or a pain-relieving trip

Peak, Final Outcome, and Duration of the Profiles In medical decision making, retrospective pain evaluation is an important matter since it reflects patients' memories of how painful the treatment was and could impact their decisions regarding future treatments. A number of empirical studies have demonstrated that retrospective pain evaluation is influenced by the *peak* and the *final moment* of the experience and not significantly impacted by the overall duration of the painful experience itself [39-45]. For example, Varey and Kahneman [39] asked 46 subjects to evaluate different discomfort profiles ranging from 15 to 35 minutes. They found that subjects' evaluations were significantly impacted by peak and final intensity but not by duration.

The same phenomenon was also found in the retrospective evaluation of watching pleasant and unpleasant video clips [41] and in patients' retrospective evaluations of experiences in undergoing colonoscopy and lithotripsy [42]. Kahneman et al. [40] performed an experiment whereby thirty-two subjects immersed one hand in 14°C water for 60 seconds and immersed the other hand at 14°C for 60 seconds. Then the temperature was gradually increased to 15°C in another 30 seconds (total duration was 90 seconds). The majority indicated that the long trial had less overall discomfort, showing final intensity effect and duration neglect.

<u>Timing of Health Outcomes (Health Discounting Behavior)</u> When evaluating health outcomes in the future, values of the outcomes are usually discounted. In cost-effectiveness analysis, discount rates are typically applied in order to deal with this time preference issue. Numerous studies have explored individuals' discounting behavior. For example, the finding that discount rates decrease as delays increase has been found in the context of back pain [46], colostomy, blindness, and depression [47], health and money [48-49]. In addition, the magnitude of the outcomes was found to impact health discounting behavior as well. Smaller outcomes were discounted at a higher rate than larger outcomes [48-49]. Another finding was an effect of the sign of the outcomes. Delayed gains were discounted more than delayed losses [50]. Ganiats et al. [51] studied health discounting for five different disease conditions (chicken pox, Parkinson's disease, tropical disease, migraine headache, and sterilization) and found that discount rates were sometimes very high (up to 116%) and varied markedly across disease conditions.

10.4.2 Conclusion

The results of the studies described above can help researchers and decision makers understand the nature of the violation of the additive independence assumption and should assist in uncovering a more suitable decomposition. While those studies provide an excellent starting point, more empirical work needs to be performed. More importantly, we need to interpret the results in such a way that they can lead to, and be incorporated into, a new aggregation structure. At the same time, we need to develop a new theoretical foundation for the decomposition of multistate health profiles. It is necessary to extent the applicability of the QALY model to handle multistate health profiles appropriately, especially if one wants to apply such models to chronic conditions.

References

[1] Fryback, D.G. (1998). Methodological issues in measuring health status and health-related quality of life for population health measures: A brief overview of the "HALY" family of measures. In Field, M.J. and M.R. Gold, Eds., *Summarizing Population Health – Directions for the Development and Application of Population Metrics,* National Academy Press, Washington DC.

[2] Gold, M.R., J.E. Siegel, L.B. Russell, and M.C. Weinstein (1996). *Cost-effectiveness in Health and Medicine.* Oxford University Press, New York.

[3] Ware, J., Jr., and C.D. Sherbourne (1992). The MOS 36-item short-form health survey (SF-36). I. Conceptual framework and item selection. *Medical Care,* 30, 473-83.

[4] Torrance, G.W., M.H. Boyle, and S.P. Horwood (1982). Application of multi-attribute utility theory to measure social preferences for health states. *Operations Research,* 30, 1043-1069.

[5] Nord, E., J.L. Pinto, J. Richardson, P. Menzel, and P. Ubel (1999). Incorporating societal concerns for fairness in numerical valuations of health programmes. *Health Economics,* 8, 25-39.

[6] Nord, E. (1999). *Cost-Value Analysis in Health Care: Making Sense out of QALYs.* Cambridge University Press, Cambridge, U.K.

[7] Russell L., M. Gold, J. Siegel, N. Daniels, and M.C. Weinstein (1996). The role of cost-effectiveness analysis in health and medicine: Panel on cost-effectiveness in health and medicine. *Journal of the American Medical Association,* 276, 1172-1177.

[8] Fanshel, S. and J.W. Bush (1970). A health-status index and its application to health-services outcomes. *Operations Research,* 18, 1021-1066.

[9] Pliskin, J.S., D.S. Shepard, and M.C. Weinstein (1980). Utility functions for life years and health status. *Operations Research,* 28, 206-224.

[10] Johannesson, M. (1994). QALYs, HYEs and individual preferences - A graphical illustration. *Social Science and Medicine,* 39, 1623-1632.

[11] McNeil, B.J., R. Weichselbaum, and S.G. Pauker (1978). Fallacy of the five-year survival in lung cancer. *New England Journal of Medicine,* 299, 1397-1401.

[12] Stiggelbout, A.M., G.M. Kiebert, J. Kievit, J.W.H. Leer, G. Stoter, and J.C.J.M. de Haes (1994). Utility assessment in cancer patients: adjustment of time tradeoff scores for the utility of life years and comparison with standard gamble scores. *Medical Decision Making,* 14, 82-90.

[13] Verhoef, L.C.G., A.F.J. de Haan, and W.A.J. van Daal (1994). Risk attitude in gambles with years of life: Empirical support for prospect theory. *Medical Decision Making,* 14, 194-200.

[14] Mehrez, A. and A. Gafni (1987). An empirical evaluation of two assessment methods for utility measurement for life years. *Socio-Economic Planning Sciences,* 21, 371-375.

[15] Bleichrodt, H., P. Wakker, and M. Johannesson (1997). Characterizing QALYs by risk neutrality. *Journal of Risk and Uncertainty,* 15, 107-114.

[16] Miyamoto, J.M., P. Wakker, H. Bleichrodt, and H.J.M. Peters (1998). The zero-condition: A simplifying assumption in QALY measurement and multiattribute utility. *Management Science,* 44, 839-849.

[17] Stiggelbout, A.M., G.M. Kiebert, J. Kievit, J.W.H. Leer, G. Stoter, and J.C.J.M. de Haes (1994). Utility assessment in cancer patients: adjustment of time tradeoff scores for the utility of life years and comparison with standard gamble scores. *Medical Decision Making,* 14, 82-90.

[18] Miyamoto, J.M. and S.A. Eraker (1988). A multiplicative model of the utility of survival duration and health quality. *Journal of Experimental Psychology,* 117, 3-20.

[19] Bleichrodt, H. and M. Johannesson (1996). The validity of QALYs: An experimental test of constant proportional tradeoff and utility independence. *Medical Decision Making,* 17, 21-32.

[20] von Neumann, J. and O. Morgenstern (1947). *Theory of Games and Economic Behavior.* Princeton, NJ, Princeton University Press.

[21] von Winterfeldt, D. and W. Edwards (1986). *Decision Analysis and Behavioral Research.* Cambridge University Press, Cambridge, U.K.

[22] Bleichrodt, H. (1995). QALYs and HYEs: Under what conditions are they equivalent? *Journal of Health Economics,* 14, 17-37.

[23] Richardson, J., J. Hall, and G. Salkeld (1996). The measurement of utility in multiphase health states. *International Journal of Technology Assessment in Health Care,* 12, 151-62.

[24] Kuppermann, M., S. Shiboski, D. Feeny, E.P. Elkin, and E. Washington (1997). Can preference scores for discrete states be used to derive preference scores for an entire path of events? *Medical Decision Making,* 17, 42-55.

[25] MacKeigan, L.D., B.J. O'Brien, and P.I. Oh (1999). Holistic versus composite preferences for lifetime treatment sequences for Type 2 diabetes. *Medical Decision Making,* 19, 113-21.

[26] Spencer, A. (2003). A test of the QALY model when health varies over time. *Social Science and Medicine,* 57, 1697-1706.

[27] The EuroQol Group (1990). EuroQol - a new facility for the measurement of health-related quality of life. *Health Policy,* 16, 199-208.

[28] Hsee, C.K. and R.P. Abelson (1991). Velocity relation: Satisfaction as a function of the first derivative of outcome over time. *Journal of Personality and Social Psychology,* 60, 341-7.

[29] Chapman, G.B. (1996). Expectations and preferences for sequences of health and money. *Organizational Behavior and Human Decision Processes,* 67, 59-75.

[30] Ariely, D. (1998). Combining experiences over time: The effects of duration, intensity changes and on-line measurements on retrospective pain evaluations. *Journal of Behavioral Decision Making,* 11, 19-45.

[31] Loewenstein, G.F. and N. Sicherman (1991). Do workers prefer increasing wage profiles? *Journal ofLabor Economics,* 9, 67-84.

[32] Loewenstein, G.F. and D. Prelec (1991). Negative time preference. *American Economic Review,* 81, 347-52.

[33] Krabbe, P.F.M. and G.J. Bonsel (1998). Sequence effects, health profiles, and the QALY model: In search of realistic modeling. *MedicalDecision Making,* 18, 178-88.

[34] Chapman, G.B. (2000). Preferences for improving and declining sequences of health outcomes. *Journal of Behavioral Decision Making,* 13, 203-218.

[35] Ariely, D. and G. Zauberman (2000). On the making of an experience: The effects of breaking and combining experiences on their overall evaluation. *Journal ofBehavioral Decision Making,* 13, 219-232.

[36] Ariely, D. and Z. Carmon (2000). Gestalt characteristics of experiences: The defining features of summarized events. *Journal of BehavioralDecisionMaking,* 13, 191-201.

[37] Loewenstein, G.F. and D. Prelec (1993). Preferences for sequences of outcomes. *Psychological Review,* 100, 91-108.

[38] Chapman, G.B. (1998). Sooner or later: The psychology of intertemporal choice. *The Psychology ofLearning and Motivation,* 38, 83-113.

[39] Varey, C. and D. Kahneman (1992). Experiences extended across time: evaluation of moments and episodes. *Journal of Behavioral DecisionMaking,* 5, 169-85.

[40] Kahneman, D., B.L. Fredrickson, C.A. Schreibner, and D.A. Redelmeier (1993). When more pain is preferred to less: Adding a better end. *PsychologicalScience,* 4, 401-5.

[41] Fredrickson, B.L. and D. Kahneman (1993). Duration neglect in retrospective evaluations of affective episodes. *Journal of Personality andSocialPsychology,* 65, 45-55.

[42] Redelmeier, D.A. and D. Kahneman (1996). Patients' memories of painful medical treatments: real-time and retrospective evaluations of two minimally invasive procedures. *Pain,* 66, 3-8.

[43] Ariely, D. and G. Loewenstein (2000). When does duration matter in judgment and decision making? *Journal of Experimental Psychology,* 129, 508-523.

[44] Langer, T., R. Sarin and M. Weber (2000). The retrospective evaluation of payment sequences: Duration neglect and Peak-and-End-Effects. Working Paper, University of Manheim.

[45] Baumgartner, H., M. Sujan, and D. Padgett (1997). Patterns of affective reactions to advertisements: the integration of moment-to-moment responses into overall judgments. *Journal of Marketing Research,* 34, 219-232.

[46] Bleichrodt, H. and M. Johannesson (2001). Time preference for health: A test of stationarity versus decreasing timing aversion. *Journal of Mathematical Psychology,* 45, 265-282.

[47] Redelmeier, D.A. and D.M. Heller (1993). Time preference in medical decision making and cost-effectiveness analysis. *Medical Decision Making,* 13, 212-217.

[48] Chapman, G.B. and A.S. Elstein (1995). Valuing the future: Temporal discounting of health and money. *Medical Decision Making,* 15, 373-88.

[49] Chapman, G.B. (1996). Temporal discounting and utility for health and money. *Journal of Experimental Psychology: Learning, Memory and Cognition,* 22, 771-791.

[50] MacKeigan, L.D., L.N. Larson, J.R. Draugalis, J.L. Bootman, and L.R. Burns (1993). Time preference for health gains versus health losses. *Pharmacoeconomics,* 3, 374-386.

[51] Ganiats, T.G., R.T. Carson, R.M. Hamm, S.B. Cantor, W. Sumner, S.J. Spann, M.D. Hagen, and C. Miller (2000). Population-based time preferences for future health outcomes. *Medical Decision Making,* 20, 263-270.

11 DUCT TAPE FOR DECISION MAKERS: THE USE OF OR MODELS IN PHARMACOECONOMICS

Anke Richter

Defense Resource Management Institute
Naval Postgraduate School
Monterey, CA 93943

SUMMARY

Operations research (OR) provides an excellent set of tools for decision makers who regulate the use of new treatments or medications. The decision about whether to use a new treatment must typically be made well before long-term trials or database studies can be conducted. However, large amounts of information about new treatments are available from the clinical trials required for drug registration. OR models can synthesize this information and use it to predict expected costs and benefits of long-term treatment use within a given population. Such analysis provides valuable additional information for the decision maker when a novel treatment is initially being considered. These analyses are like duct tape for the decision maker: they are designed to make use of the best currently available information to help current decisions, thereby bridging the gap until better information becomes available.

KEY WORDS

Health economics, Pharmacoeconomics, Cost-benefit models, Cost-effectiveness models

11.1 INTRODUCTION

In a general sense, health economics and health outcomes research (which are both encompassed in pharmacoeconomics) are the study of the impact of new medications on society. They attempt to capture both the health benefits of a new medication such as improved survival (either through a cure or by slowing disease progression), lessened pain, improved functioning, and improved quality of life, as well as the economic impacts of a new medication. The preferred perspective is the societal point of view. However, other perspectives may be adopted, including that of a governmental or private health insurer or a patient. By capturing both benefits and costs, such methods illuminate the trade-offs involved in decisions about whether to use a medication in specific populations.

New medications are brought to market after undergoing a series of rigorous clinical trials, which are designed to test the safety and efficacy of the new medication. These clinical trials are typically of comparatively short duration (such as six months) due to the costs incurred in running them and the difficulty in tracking patients over time. However, many diseases are chronic conditions that progress over time. For such diseases, it may be possible to demonstrate the efficacy of new treatments within the time frame of the clinical trial by examining disease markers (such as blood pressure for hypertension, viral load for HIV, or FEV [forced expiratory volume, a lung function test] for asthma) or by showing a lessening of symptoms or acute attacks (such as number of myocardial infarctions in heart disease, or pain with arthritis). However, the potential long-term impacts of new medications cannot be determined from short-term clinical trials.

The economic consequences of new treatments are also not directly available from clinical trials. Beyond drug acquisition costs, many factors influence the overall economic impact of a new treatment. Acquisition costs can be offset by lessening the number of serious disease-related events that require hospitalization (such as a heart attack or stroke) or urgent care in an emergency room (such as severe asthma attacks or severe bowel disorders). A new treatment may reduce the number of medications that patients must consume. However, some costs may increase the economic burden of a new treatment. Side effects of a medication may require treatment in order for the patient to be able to continue to take the medication. Serious adverse events may even require hospitalization and medical procedures. The overall economic impact of a medication can be estimated by combining the observed results from the clinical trial with treatment patterns and costs of care seen in clinical practice.

The best methods for determining the long-term impact of new medications would be to either conduct naturalistic clinical trials after approval of a new medication or to conduct database studies of longitudinal data collected as the medication is used in practice. Such studies would provide the best and most reliable answers to questions about the future costs and benefits of a new medication. Unfortunately, this approach is of no use to decision makers who are faced with the questions now, at the time the medication is released.

A decision maker may be asking the following questions:

- How many future undesirable health consequences will the medication prevent?
    - How many hospitalizations will be avoided?
    - How many years of pain or diminished quality of life will be averted?
    - Will these adverse events be avoided or merely delayed?
    - Will the treatment have better/worse results in certain populations?
- How many patients under my care will receive the medication?
- What other treatment options exist, and how effective is each?
- What adverse events are associated with the medication?
    - How often will such events occur?
    - What types of treatment do such events necessitate?
- What is the budgetary impact of the medication?
- Is the acquisition cost offset by other cost factors such as reductions in
    - Emergency room visits or hospitalization?
    - Use of other medications?
    - Administration costs?
- What tradeoffs are made in choosing whether to accept this medication?

Decision makers must find some way to convert the available short-term clinical trial information into information about the long-term impacts of a new medication. Operations research (OR) models are ideally suited for this task. They can model disease progression and the impacts of new medications. Depending on the available data, the models can range from simple decision trees to complex simulations. Three case studies of analyses that have been conducted using OR techniques are described in detail below.

11.2 DECISION TREE MODEL OF ANTIHYPERTENSIVE MEDICATIONS

11.2.1 Application

Hypertension is a chronic condition that can lead to very serious cardiac complications. Because of these complications, there is general agreement on the need to treat hypertension and on the economic benefits of antihypertensive therapy [1]. While exercise and diet are the preferred first-line treatment, many patients have only marginal success in following such recommendations. Second-line treatments include antihypertensive medications. Physicians have numerous pharmacologic options for hypertension therapy: over 100 antihypertensives spanning eight classes of therapy are on the market worldwide. If one therapy does not work, a patient can simply be switched to another [2].

A new antihypertensive medication, an angiotensin-II inhibitor, was released in 1999. The new drug had similar or slightly improved efficacy to existing medications, cost more than most existing medications (many of which are generic), and had a different and mild adverse event profile. Clinical trials of this new drug were typically at most six months in duration. Decision makers wanted to know how this new drug would affect the managed care system. To answer this question, a decision analytic model was developed. The model was designed to explore the costs and consequences of treating mild-to-moderate uncomplicated hypertension starting with an angiotensin–II (A-II) inhibitor, relative to four other drugs – a diuretic, a beta-blocker, an angiotensin converting enzyme (ACE) inhibitor, and a calcium channel blocker (CCB).

11.2.2 Methodology

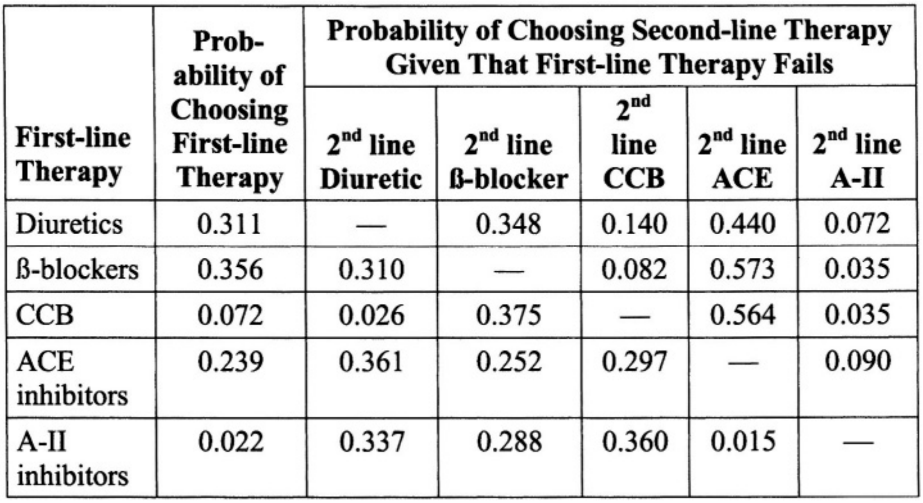
Current hypertension treatment patterns were ascertained from a literature review and a physician survey [3]. Key model data obtained from these efforts included the following:

- A patient with uncontrolled mild-to-moderate hypertension is seen monthly.
- Determining whether or not a specific drug is effective at lowering blood pressure can take up to three months.
- During these three months, patients may increase the dosage of their medication or switch to another therapy.
- Therapies may be switched either because
  - The patient has experienced intolerable adverse events, or
  - The drug has failed to control hypertension.

- Once hypertension has been controlled, the patient is seen every three months.

The physician survey also provided information that was used to estimate the probability that a particular drug is chosen as first-line therapy and the probability of the choice of second-line therapy (given that the first-line therapy fails). These probabilities are shown in Table 11.1. The model assumes that the remaining drugs have an equal likelihood of being chosen for third-, fourth-, and fifth-line therapies.

Table 11.1 Physicians' probability of choosing hypertension medications



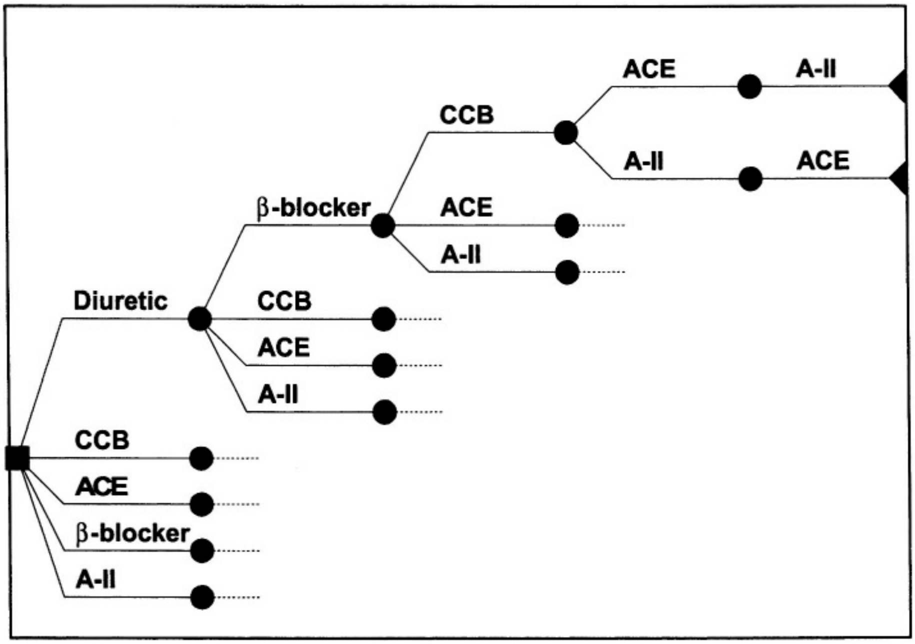
First-line Therapy	Probability of Choosing First-line Therapy	Probability of Choosing Second-line Therapy Given That First-line Therapy Fails				
		2nd line Diuretic	2nd line ß-blocker	2nd line CCB	2nd line ACE	2nd line A-II
Diuretics	0.311	—	0.348	0.140	0.440	0.072
ß-blockers	0.356	0.310	—	0.082	0.573	0.035
CCB	0.072	0.026	0.375	—	0.564	0.035
ACE inhibitors	0.239	0.361	0.252	0.297	—	0.090
A-II inhibitors	0.022	0.337	0.288	0.360	0.015	—

Copyright 2001 Medicom International. Reprinted with permission.

This information was used to construct a series of decision trees. The main decision tree (Figure 11.1) determines the outcomes associated with a specific sequence of drugs. All drug sequences are enumerated and the tree can be rolled back to any level using the probabilities shown in Table 11.1. Each branch is evaluated in terms of expected time to control and cost of choosing that particular sequence of drug therapies.

For each medication considered along the branches of the main decision tree, it was necessary to determine the medication's probability of achieving hypertension control, its adverse event rate and its cost. Since patients are typically started at low doses of medication and then have their doses titrated upwards as necessary for hypertension control, each medication can be given at varying doses. Each dose of medication has an associated probability of achieving hypertension control, adverse event rate and cost. The probability

Figure 11.1 Main decision tree: Drug sequences



Copyright 2001 Medicom International. Reprinted with permission.

of each drug being titrated was available from the comparative clinical trials of the new A-II inhibitor, as were drug dose efficacy and adverse event rates. Therefore, a series of additional decision trees using the titration likelihood and the effects of a given drug at each dosage level were used to calculate the drug's overall probability of achieving hypertension control, its overall adverse event rate and its overall cost.

Adverse event treatment algorithms were designed for each adverse event based on hypertension severity level. However, it was believed that patients suffering from continual moderate and severe adverse events would not remain on drug therapy, whereas those with mild adverse events would. Therefore, adverse events (and hence adverse event costs) were divided into two types—first-quarter adverse event costs, which include all of the adverse events experienced during the clinical trials and are incurred only during the first three-month period when a patient is placed on a new medicine, and maintenance adverse event costs, which include only costs resulting from mild adverse events and are incurred quarterly while a patient remains on medication. The model incorporates costs of drugs, physician visits, and adverse event treatments. Table 11.2 summarizes the efficacy and cost inputs that the model requires.

11.2.3 Results

Combining the cost and efficacy information presented in Tables 11.1 and 11.2 in the decision tree model structure shown in Figure 11.1, it is possible to roll the decision tree back to the choice of initial therapy. Therefore, the baseline results, shown in Table 11.3, are the weighted average of all pathways possible for each initial drug prescribed. For example, results reported for the diuretic are the weighted average of the results of all drug sequences in which the diuretic is initially prescribed (24 possible sequences, as shown in Figure 11.1).

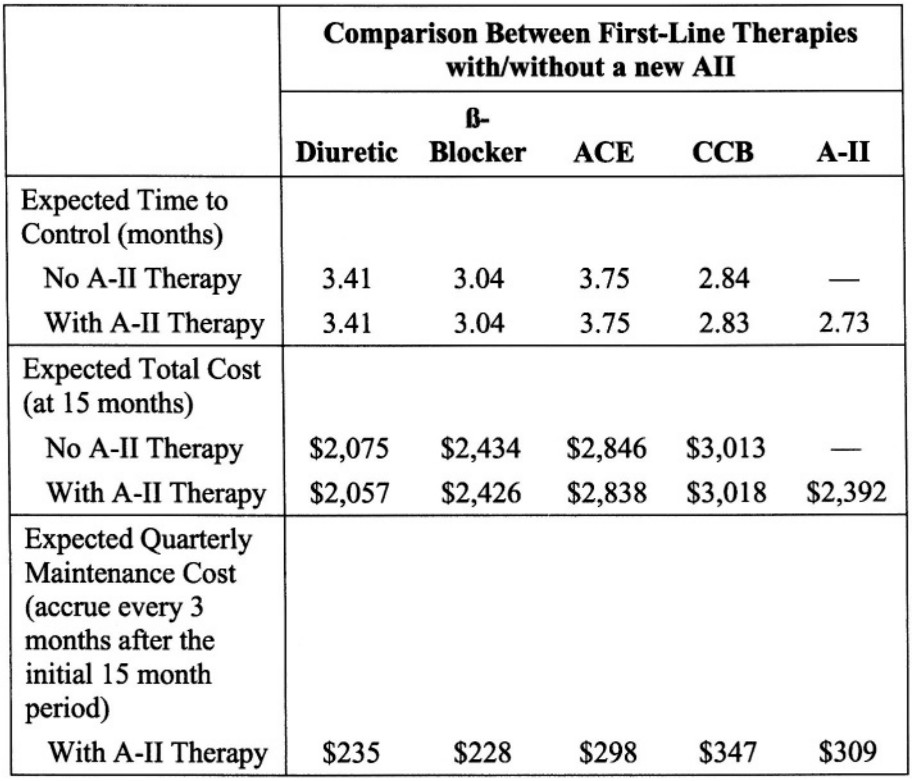
Table 11.2 Efficacy and cost data for decision tree in Figure 11.1

	Efficacy	**Costs**		
First-line Drug Therapy	**Probability of Hypertension Control**	**Expected Quarterly Drug Cost (given successful treatment)**	**First Quarter Expected Adverse Event Costs**	**Quarterly Expected Maintenance Adverse Event Costs**
CCB	0.78	$113.40	$850	$104
ß-Blocker	0.70	$4.37	$804	$39
ACE Inhibitor	0.53	$106.35	$639	$57
Diuretic	0.71	$1.09	$301	$56
A-II Inhibitor	0.72	$127.11	$332	$43

Copyright 2001 Medicom International. Reprinted with permission.

The model time horizon is initially set at 15-months, the longest time it would take to cycle through all possible first line medications. The measure of efficacy (expected time to hypertension control) is not dependent on the model time horizon, however, the cost results are dependent on the time horizon as they are continually accruing. Therefore, in Table 11.3, the initial expected total 15-month cost is presented as well as the expected total costs that would be accrued every three months thereafter. In this manner, a decision maker can choose a time horizon of interest and calculate the total costs over the entire time horizon from these results.

Table 11.3 Efficacy and costs results (expected time to control and expected costs) from the decision tree analysis



	Comparison Between First-Line Therapies with/without a new AII				
	Diuretic	ß-Blocker	ACE	CCB	A-II
Expected Time to Control (months)					
No A-II Therapy	3.41	3.04	3.75	2.84	—
With A-II Therapy	3.41	3.04	3.75	2.83	2.73
Expected Total Cost (at 15 months)					
No A-II Therapy	$2,075	$2,434	$2,846	$3,013	—
With A-II Therapy	$2,057	$2,426	$2,838	$3,018	$2,392
Expected Quarterly Maintenance Cost (accrue every 3 months after the initial 15 month period)					
With A-II Therapy	$235	$228	$298	$347	$309

For any given initial therapy except the CCB, the inclusion of the A-II inhibitor in the subsequent therapeutic options reduced the expected costs (at 15 months). The reduction in cost was mainly due to the lower initial adverse event costs of the A-II inhibitor and the ability to avoid using the CCB, which is by far the costliest drug in terms of drug acquisition and adverse event costs.

Initiating therapy with the A-II inhibitor (which is not currently common practice) is the second least expensive option. However, the savings over the other therapies that the model predicts during its 15-month time horizon will be reduced over time given the A-II inhibitor's drug acquisition costs and expected quarterly maintenance costs.

Extensive sensitivity analyses were conducted to test the stability of the model and its results. These analyses and the full model are reported in the literature [4-7].

The information gained from this analysis provided cost and efficacy estimates to decision makers before the new drug was used in clinical practice. In addition, it provided important input for two different policy questions: Should the drug be available as a second or later therapy once the standard initial therapy has failed? Should the drug be made available as a potential first-line therapy? The goal of the analysis was to provide decision makers with information that can be used to improve the therapeutic options of hypertension care.

11.3 MONTE CARLO SIMULATION OF HIV/AIDS VIRAL LOAD TESTING

11.3.1 Application

This case study examines a question that arose when viral load testing was still novel. However, it demonstrates a technique that remains pertinent with the introduction of any new monitoring/testing method. The question of interest is the following: Given a new method for testing how well a patient is responding to medications, how frequently should the test be conducted?

This question is of great importance with the human immunodeficiency virus (HIV) infection since the virus mutates rapidly over time and can become resistant to medications. When a patient's virus becomes resistant to the medication, viral load levels in that person increase rapidly. High levels of viral load damage a patient's immune system and lessen the person's ability to fight off common infections. Therefore, it is important to catch the point of viral resistance to medication as soon as possible so that the patient may be placed on a new medication.

In HIV, the matter is further complicated by the question of adherence to medications. The medications used to treat HIV cause numerous unpleasant side effects and are difficult to take. The medications must be taken continuously every 8, 12, or 24 hours. When a patient ingests subtherapeutic levels of medication due to repeated "drug holidays" or other nonadherence to the medication regimen, such as skipping doses or only taking certain drugs in a combination therapy, the virus develops resistant mutants more rapidly. This shortens the period of beneficial effects from the medication and allows for the development of multi-drug-resistant HIV strains. Current treatment guidelines strongly emphasize the importance of good adherence to medication, and delineate the possible negative consequences of nonadherence [8-10].

For HIV, the testing question is thus the following: Considering both the inherent progression of the disease and the possibility of non-adherence to medications, what is the optimal viral load testing frequency?

The frequency of viral load testing determines how quickly viral rebound is detected and how soon a patient is switched to the next therapeutic option. Thus, the frequency of viral load testing may affect the cost of treatment, the pattern of antiretroviral drug use, and (possibly) the quality of life and life expectancy for HIV-infected individuals. Annual costs of care and the lifetime cost per person may be affected by differences in the duration of highly active antiviral therapy (HAART) drug regimens, how soon patients are placed on more expensive (four-drug) therapies, and the cost of treatment for opportunistic infections and other medical care for individuals at different levels of immune suppression. Patterns of therapy are affected because different monitoring frequencies may cause regimens to be administered for different lengths of time. Patient outcomes may also be affected by different progression rates induced by the varying durations of suboptimal therapy.

A Monte Carlo simulation was designed to examine the question of optimal testing frequency. The simulation captured HIV disease progression in the presence of medications and their varying efficacy and levels of medication adherence. Using data on costs and consequences of HIV disease, the model estimates health outcomes and costs for patients undergoing three different frequencies of viral load testing (every month, every three months, and every six months). Four hypothetical populations, described by disease stage and rate of disease progression, are examined. These groups are patients with:

1. Moderate disease stage, average disease progression;
2. Moderate disease stage, fast disease progression;
3. Moderate disease stage, slow disease progression; and
4. Severe disease stage, average disease progression.

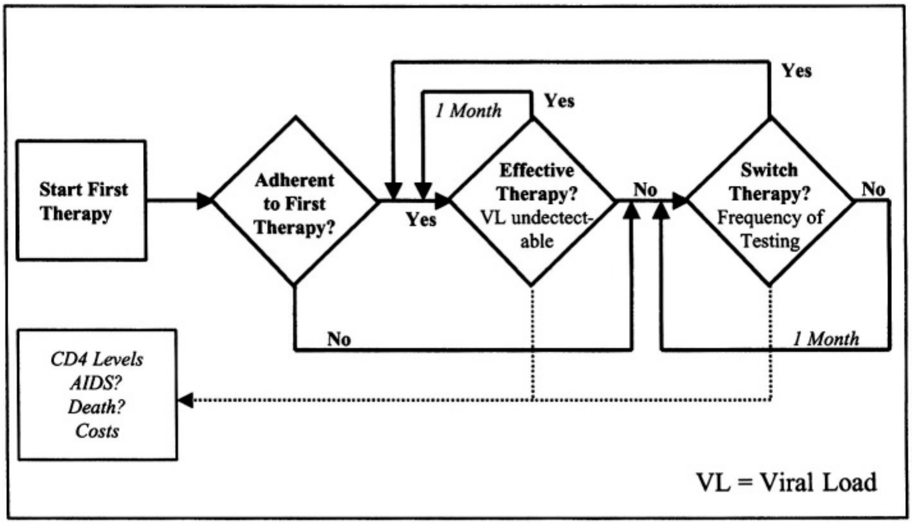
These population groups are analyzed under varying assumptions about adherence to medication. This disaggregate analysis is performed to capture the possible influence of each of these factors (disease stage and rate of disease progression) on the impact of viral load testing frequency.

11.3.2 Methodology

A Monte Carlo simulation is performed for each population group, tracking the disease progression of individuals for five years. The population groups are distinguished by their initial average CD4 cell counts (a measure of immune system function – higher numbers of these cells are better) and their initial average baseline viral loads. These two parameters provide estimates of how advanced the disease is and how fast individuals are expected to progress, and thus define the four groups described above.

Figure 11.2 provides a schematic of the Monte Carlo simulation. During each simulation, individual patients are simulated and their baseline viral loads and CD4 cell counts are determined randomly, according to the population's probability distributions. The results reported here represent outcomes for a simulated population of 5,000 individuals. The model tracks on a monthly basis each simulated patient's CD4 cell count, viral load, AIDS status, possible death, testing costs, drug therapy costs, and medical care costs.

Figure 11.2 Schematic for Monte Carlo simulation


Start First Therapy
Adherent to First Therapy?
Yes
No
1 Month
Yes
Effective Therapy? VL undectect-able
No
Yes
Switch Therapy? Frequency of Testing
No
1 Month
CD4 Levels
AIDS?
Death?
Costs
VL = Viral Load

Patients are treatment naïve at the start of the model (that is, their viral strains are not resistant to any of the available medications). Patients are followed for five years, during which time they are treated with a sequence of combination drug regimens. The regimens are chosen from the consensus statement of the International AIDS Society – USA Panel [10]. When a therapy is first effective, viral load is undetectable and CD4 cell counts increase. As a therapy continues to be effective, the viral load remains undetectable and no CD4 cells are lost. Once the patient's viral load becomes detectable, his/her CD4 cell count declines at a rate determined by the initial viral load. If the patient's viral load is detectable when he/she is tested, the patient is switched to the next drug regimen.

Patients have a 40 percent likelihood of being nonadherent to antiretroviral medications during the first combination therapy. When a patient is nonadherent, the viral load level rebounds sooner (as determined by a probability distribution) and the patient must switch to a new drug regimen.

Within the model, the following parameters are simulated for each individual from a probability distribution (in parentheses):

- Initial viral load and initial CD4 cell count (truncated normal and uniform distribution, respectively);
- Rate of CD4 cell count decline given a viral load level [11] (uniform distribution);
- Monthly probability of progressing to AIDS, depending on CD4 cell count (discrete distribution);
- Monthly probability of death, depending on CD4 cell count (discrete distribution);
- Probability that a therapy will be effective (varies with therapy type and order) (discrete distribution);
- Duration of effectiveness of a given therapy (varies with therapy type and order) (truncated normal distribution);
- Increase in CD4 cell count given an effective therapy (uniform distribution);
- Probability of patient nonadherence during the first therapy (discrete distribution); and
- Monthly probability that resistance develops due to nonadherence (discrete distribution).

The following parameters have set values for all individuals:

- Cost of drug therapy [12];
- Testing cost;
- Other medical care costs (dependent on CD4 count and AIDS status); and
- Salvage therapy costs once the antiretroviral medications have been exhausted.

11.3.3 Results

Results for each of the four populations are presented in Table 11.4. The outcomes assume that each population is composed of 5,000 individuals. (This number was chosen so that the simulations would have time to converge.) Smaller populations, as would be seen in clinical practice, will have results distributed around the mean of the larger population, depending on the variance about the mean and the actual size of the smaller population considered. This distribution implies that actual practice experience may be different from the population's true mean.

Table 11.4 shows the incremental results of implementing viral load testing every month, every three months, and every six months. Results are expressed in terms of incremental quality-adjusted life years (QALYs) gained and incremental costs. In Populations 1, 2, and 3 (slow, average, and fast progressors, respectively, in a moderate disease state), the decrease in antiretroviral drug costs and decreases in other medical care costs offset the increase in testing costs when testing frequency is increased from every six months to every three months. Increasing testing frequency from every three months to every month increases costs (due to the additional testing costs), but yields no appreciable gain in QALYs. Thus, this option is not cost-effective. In Population 4 (advanced disease state in an average progressor), lowering the testing frequency from every six months to every three months also results in a net cost savings. Lowering testing frequency further to every month increases costs due to the increased testing costs. However, since Population 4 is in an advanced disease state, there is a small gain in QALYs. The incremental cost-effectiveness ratio is $23,400 /QALY. This value is low compared with many currently accepted interventions, and it can easily be argued that this option is cost-effective.

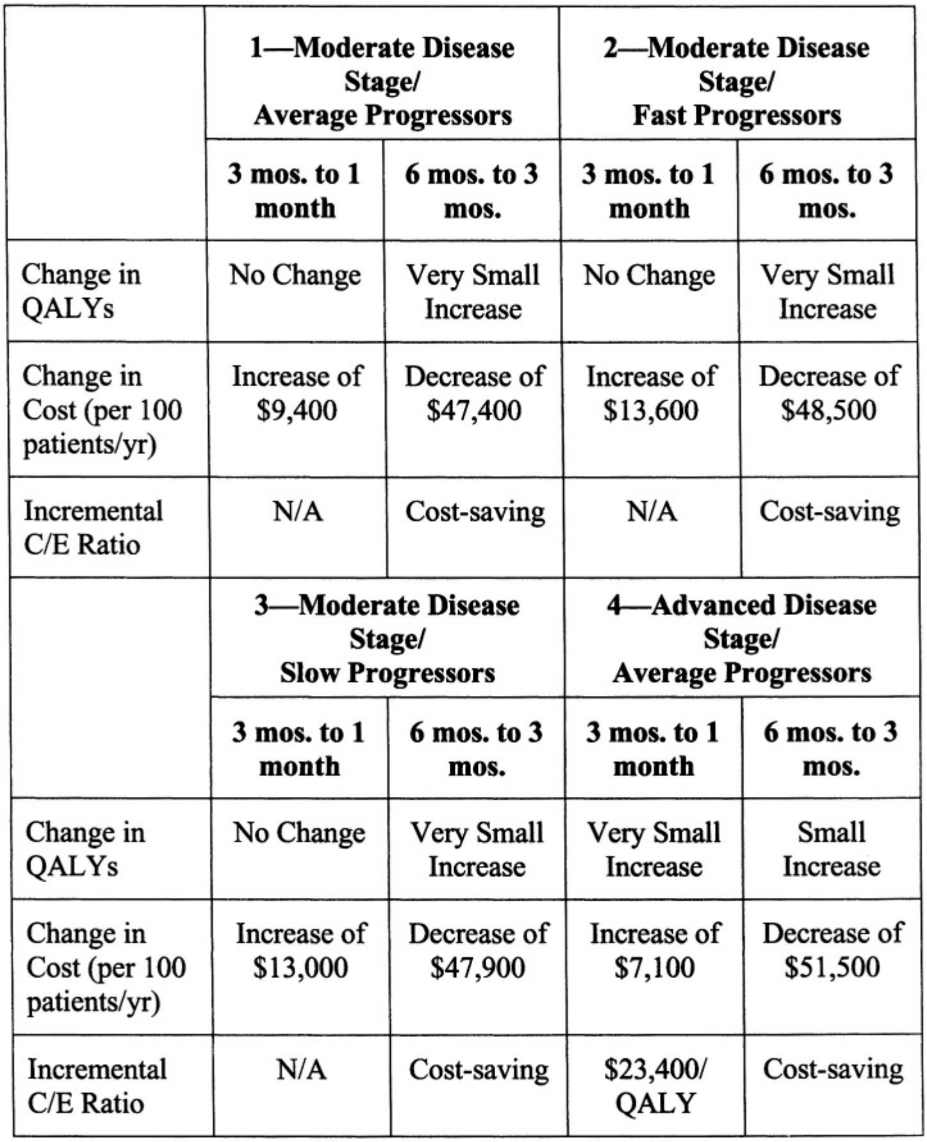
This analysis permitted an investigation that was not possible at the time of the original question. Given the best available data at the time, the analysis provided insight into a series of decisions that the managed care companies were facing when they first included HIV viral load testing in their benefit packages. As new tests – both for HIV and other diseases – become available, a similar type of model can be constructed to provide insight into the most appropriate testing frequency.

11.4 MARKOV MODEL OF CANCER TREATMENT MEDICATIONS

11.4.1 Application

This case study demonstrates an application of OR techniques for comparative analysis of new medications still in development. This is the most hypothetical case study since few clinical trials exist from which to gather efficacy and safety data. However, even in the earliest stages of new drug development, it is possible, and frequently advantageous, to examine the drugs in terms of their potential benefits and costs to the end users (and end decision makers).

Table 11.4 Incremental cost and QALYs by testing frequency



	1—Moderate Disease Stage/ Average Progressors		2—Moderate Disease Stage/ Fast Progressors	
	3 mos. to 1 month	6 mos. to 3 mos.	3 mos. to 1 month	6 mos. to 3 mos.
Change in QALYs	No Change	Very Small Increase	No Change	Very Small Increase
Change in Cost (per 100 patients/yr)	Increase of $9,400	Decrease of $47,400	Increase of $13,600	Decrease of $48,500
Incremental C/E Ratio	N/A	Cost-saving	N/A	Cost-saving

	3—Moderate Disease Stage/ Slow Progressors		4—Advanced Disease Stage/ Average Progressors	
	3 mos. to 1 month	6 mos. to 3 mos.	3 mos. to 1 month	6 mos. to 3 mos.
Change in QALYs	No Change	Very Small Increase	Very Small Increase	Small Increase
Change in Cost (per 100 patients/yr)	Increase of $13,000	Decrease of $47,900	Increase of $7,100	Decrease of $51,500
Incremental C/E Ratio	N/A	Cost-saving	$23,400/ QALY	Cost-saving

Treatment of solid cancer tumors remains a serious unmet medical need. Broad acting agents could provide significant treatment advances. Current chemotherapies can provide palliation and increased survival time. However, they are highly toxic and are only effective in specific patient populations and/or for specific tumor types. The accumulation of new genetic and biological information about cancer is creating the possibility of developing new drugs with broad activity and less toxicity than current chemotherapies.

In this case study, the objective was to produce a basic, flexible computer model that incorporates the treatment paths, costs, and outcomes associated with the management and treatment of solid cancer tumors. This model incorporated data on available treatments, their efficacy, and associated adverse events. This permits a comparison between existing treatments and potential new medications, and allows the model to be used by a variety of decision makers such as managed care companies or pharmaceutical companies themselves who want to know what levels of safety, efficacy, and cost would need to be seen in potential new treatments to make them a valuable treatment option in comparison to existing options.

More specifically, the model examines the potential effects of new anticancer compounds on the health benefits (mortality, disease-free survival, etc.) and total treatment costs of solid cancer tumors. The basic structure allows the model to be quickly adapted to different solid tumor cancers such as breast or colorectal cancer. The treatment costs included in the model are the costs of surgery, chemotherapy, radiotherapy, hormone therapies and procedures, and diagnostic therapies and procedures. Into this mix of treatments, the new medication can be added as a replacement therapy or as an adjuvant therapy. The model also calculates the impact of the new compounds on patients' quality of life.

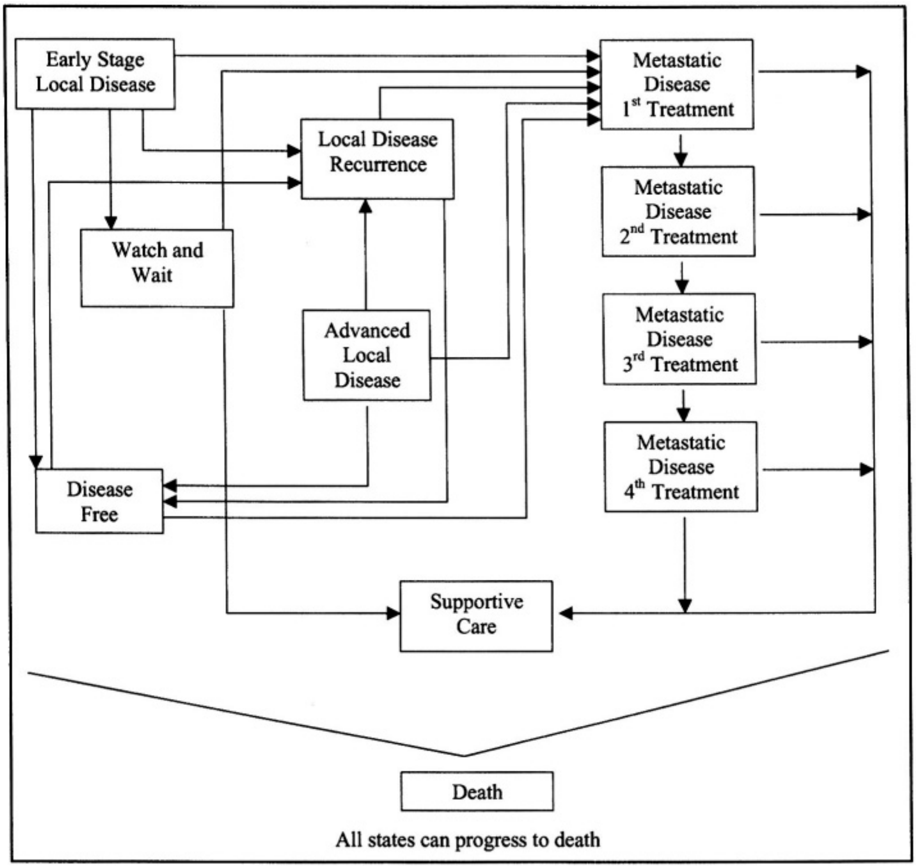
11.4.2 Methodology

The base model was constructed using a Markov framework. Figure 11.3 is a diagram of treatment pathways. Patients enter into the model in one of these disease states and then follow the pathways, marked in arrows, based on a set of transition probabilities. The transition probabilities account for disease progression over time, which incorporates both the natural disease progression as well as any impact of the selected treatments on slowing disease progression. As a patient passes through each state, costs and quality of life values are accrued.

Health states for the solid tumor cancers used current information from several sources, including: the American College of Surgeons (ACS) [13], the American Cancer Society [14], the National Cancer Institute (NCI) [15], and the National Comprehensive Cancer Network (NCCN) [16].

Transitions between these health states are determined by the rate of progression of the cancer and the therapy provided at each health state. To capture the natural cancer disease progression rates, a cycle time of three months was chosen. The natural, untreated, cancer progression between health states is summarized in a probability matrix that quantifies the

Figure 11.3 Schematic for Markov Model of solid tumor cancers


Early Stage Local Disease
Metastatic Disease 1st Treatment
Local Disease Recurrence
Metastatic Disease 2nd Treatment
Watch and Wait
Advanced Local Disease
Metastatic Disease 3rd Treatment
Metastatic Disease 4th Treatment
Disease Free
Supportive Care
Death
All states can progress to death

likelihood that a patient will progress to another health state during a given three-month period. The transition probabilities required by this model were not directly available from the published literature. Therefore, it was necessary to calculate the transition probabilities from available data on mortality, morbidity, progression to metastatic disease, and disease-free survival. The transition probabilities were heuristically developed by iteratively solving the transition matrix to provide the mortality, progression to metastatic disease, and disease-free survival from varying starting stages of cancer. All information used to calculate the health state transition probabilities came from DeVita's Cancer Anthology [17].

The impact of treatments were incorporated in one of two ways, depending on the available data. When possible, typically when there was data from a comparative clinical trial, the untreated cancer progression rates were modified by the observed changes in relative risk of disease progression. When this information was not available, the same heuristic approach used to create the base case transition matrix was used to recalculate the entire transition matrix from the mortality, morbidity, and disease progression observed in longitudinal studies of the specific treatment.

Health state utilities came from the published literature for each cancer, a good starting source being Teng and Wallace [18]. Since cancer treatments themselves affect a patient's quality of life, the utilities were dependent both on the health state itself as well as the treatment selected.

The main challenge with this model is finding sufficient baseline data. Each of the states shown in Figure 11.3, with the exception of "disease free", "supportive care" and "watch and wait", may have four classes of treatment options (surgery, chemotherapy, hormone, and radiation therapies) available. Since these classes of therapy may be given in combination with each other (e.g. [surgery and chemotherapy] or [chemotherapy and radiation therapy]), there are 14 combinations of treatment classes that may be provided. In this discussion, a combination of treatment classes is called a "treatment category". Within each of these treatment classes there may be several different medications and/or procedures that could be used. Each potential treatment option within treatment category is associated with a unique probability of transitioning to the other states, costs, health state utilities, and likelihood of adverse events associated with therapy. As a rough estimate, there are seven states, with fourteen treatment categories per state, so even if there were only 5 treatment options per treatment category, there would be about 490 different treatments for which to find data. The most feasible approach given the immense amount of data required by the model and the scarcity of suitable data sources is to limit the number of options that the decision maker can compare.

To determine the "best" (most commonly used and most relevant comparators for the novel medication) treatment options to present to the decision maker, a clinical oncologist reviewed practice patterns at his large oncology program. This clinical oncologist chose and verified the top three treatments, procedures, and/or diagnostics used in the management of the specific cancers for each health state. The clinical consultant also provided estimates of the percentage of use for each treatment in each health state. Default values for the percentage utilization rates for each treatment category were obtained using data from the American College of Surgeons National Cancer Database (NCDB) [13], which details current treatment methods for each type of cancer.

To calculate the expected cost by treatment category, the per patient costs associated with each treatment option and the per patient likelihood of the option being used were estimated, and then used to calculate the expected three-month cost of each treatment category, per patient, by health state. The costs of adverse events are also included in the total costs for each treatment. For example, the baseline cost of a specific chemotherapy includes the three-month cost of the chemotherapy in addition to the cost of treating the adverse events associated with that chemotherapy over the three months. Three month costs are calculated since this is the cycle time of the Markov model.

The types of costs that can be accrued in each health state are separated into seven categories. These are costs for: diagnostic tests and procedures; surgery; chemotherapy; radiotherapy; hormone therapies and procedures; other treatments; and new treatments. The category "other treatments" includes any cancer treatment that does not fall into the above categories; examples include biotechnology products (e.g., Herceptin (a monoclonal antibody)) and pain medications. The new treatment category is where the decision maker includes the new cancer treatment in development about which the comparisons are to be made. The new treatment can function either as a stand-alone treatment category or as new part of an existing treatment category.

The model can then be run comparing different treatment options at different stages of cancer, most specifically, those containing the new treatment as compared to those that do not.

11.4.3 Results

The model outputs include the expected time patients spend in each health state, the costs and benefits accrued in each health state, and the total costs and total benefits in terms of survival and quality-adjusted life years. In addition, if the decision maker sets a monetary value for a life year or a quality-adjusted life year, the model calculates the net benefit of the treatment (total monetary value of the benefits minus the total costs). All results are based a time horizon of 25 years, so that lifetime data is collected for most, if not all, patients, depending on cancer progression rates. If two therapies are considered, the model provides the following comparative analyses:

- Incremental total expected costs by health state;
- Incremental total expected quality-adjusted life years by health state;
- Incremental expected time spent in a given health state;

- Incremental total expected survival time; and
- Incremental (monetary) net benefit.

In addition, the model user may use the model to measure other factors, such as time to progression and survival times.

This was an interesting project because it demonstrated the information that could be obtained very early in a new medications development that could be useful in deciding on the "must have" features in terms of safety and efficacy, as well as cost, in order for a new medication to be a valuable addition to the current treatment options for cancer. Despite rough data, a base model could be constructed that passed top-line medical scrutiny. The model provided information to the internal development team responsible for the development of the new cancer medications and could be used to gather decision makers' impressions about various new medications in early development. As the results of clinical trials for new drugs become available, the model can be updated to reflect the new information. The new costs and benefits can then be shown to decision makers to gauge their level of interest in the new drugs. This model is very versatile and can provide useful information to a variety of potential end-users.

11.5 CONCLUSIONS AND AVENUES FOR FURTHER RESEARCH

The combination of OR and health economics/health outcomes research has a great deal of potential for providing useful, practical information for decision makers. The challenge will be to bring the scientific rigor and standards of OR to these fields to ensure that the best possible models and analyses are provided when using common modeling methodologies, such as decision trees, Markov models, Monte Carlo simulations, and other mathematical simulations.

It is important to remember that while models provide good estimates about the potential long-term impacts of medications, they are only estimates. As the results from long-term studies of new drugs become available, they should be used to update the models and form a basis for reevaluating medications and their role in fighting any given disease. The models are duct-tape for the decision maker: they are designed to make use of the best currently available information to help current decisions, thereby bridging the gap until better information becomes available.

Acknowledgments

These projects were conducted at RTI-Health Solutions in collaboration with teams of colleagues, to whom I am grateful for their excellent support.

References

[1] World Health Organization (1999). 1999 World Health Organization – International Society of Hypertension Guidelines for the Management of Hypertension. *Journal of Hypertension,* 17, 151-83.

[2] Hoerger, T.J., M.V. Bala, J.L. Eggleston, D.E. Hilleman, J.M. Neutel, and P.A. Tomondy (1998). A comparative cost-effectiveness study of three drugs for the treatment of mild-to-moderate hypertension. *Pharmacy and Therapeutics,* 23, 245-67.

[3] Richter, A., C. Ostrowski, M.P. Dombeck, K. Gondek, and J.L. Hutchinson (2001). Delphi panel on current hypertension treatment patterns. *Clinical Therapeutics,* 23, 160-7.

[4] Richter, A., K. Gondek, C. Ostrowski, M.P. Dombeck, and S. Lamb (2001). Mild-to-moderate uncomplicated hypertension, further analysis of a cost-effectiveness study of five drugs. *Managed Care Interface,* 14, 61-69.

[5] Penna, P., E. Cox, T. Joseph, L. Lehman, T. Morrow, A. Richter, J. Sowers, and D. Tepper (2000). Roundtable Discussion, Part I – Epidemiologic, demographic, and treatment challenges in hypertension. *Managed Care Interface,* Supplement C, 10-16.

[6] Penna, P., E. Cox, T. Joseph, L. Lehman, T. Morrow, A. Richter, J. Sowers, and D. Tepper (2000). Roundtable Discussion, Part II – Development of a Pharmacoeconomic Model in Hypertension. *Managed Care Interface,* Supplement C, 17-23.

[7] Penna, P., E. Cox, T. Joseph, L. Lehman, T. Morrow, A. Richter, J. Sowers, and D. Tepper (2000). Roundtable Discussion, Part III – Hypertension Management in Health Plans. *Managed Care Interface,* Supplement C, 24-31.

[8] U.S. Department of Health and Human Services (2002). *Guidelines for the use of antiretroviral agents in HIV-infected adults and adolescents,* http://www.hivatis.org/guidelines/adult/May23_02/AAMay23.pdf, Accessed February 4, 2002.

[9] Vandamme, A.M., F. Houyez, D. Banhegyi, B. Clotet, G. De Schrijver, K.A. De Smet, W.W. Hall, R. Harrigan, N. Hellmann, K. Hertogs, C. Holtzer, B. Larder, D. Pillay, E. Race, J.C. Schmit, R.

Schuurman, E. Schulse, A. Sonnerborg, and V. Miller (2001). Laboratory guidelines for the practical use of HIV drug resistance tests in patient follow-up. *Antiviral Therapy,* 6, 21-39.

[10] Yeni, P.G., S.M. Hammer, C.C. Carpenter, D.A. Cooper, M.A. Fischl, J.M. Gatell, B.G. Gazzard, M.S. Hirsch, D.M. Jacobsen, D.A. Katzenstein, J.S. Montaner, D.D. Richman, M.S. Saag, M. Schechter, R.T. Schooley, M.A. Thompson, S. Vella, and P.A. Volberding (2002). Antiretroviral treatment for adult HIV infection in 2002, updated recommendations of the International AIDS Society-USA Panel. *Journal of the American Medical Association,* 288, 222-35.

[11] Mellors, J.W., A. Munoz, J.V. Giorgi, J.B. Margolick, C.J. Tassoni, P. Gupta, L.A. Kingsley, J.A. Todd, A.J. Saah, R. Detels, J.P. Phair, and C.R. Rinaldo (1997). Plasma viral load and CD4+ lymphocytes as prognostic markers of HIV-1 infection. *Annals of Internal Medicine,* 126, 946-954.

[12] Red Book™ for Windows®, Version 4.0 (2001). Medical Economics Company.

[13] American College of Surgeons (2003). *Journal of the American College of Surgeons.* http://www.facs.org, Accessed May 2003.

[14] American Cancer Society (2003). http://www.cancer.org, Accessed May 2003.

[15] National Comprehensive Cancer Network (2003). *Clinical Practice Guidelines,* http://www.nccn.org, Accessed May 2003.

[16] National Cancer Institute (2003). *Cancer Treatment PDQ® (Physician Data Query), – Health Professionals,* http://www.nci.nih.gov/cancerinfo/pdq, Accessed May 2003.

[17] DeVita, V.T., S. Hellman, and S.A. Rosenberg (2001). *Cancer, Principles and Practice of Oncology. 6th Edition.* Lippincott Williams and Wilkins, Philadelphia, PA.

[18] Tengs T.O. and A. Wallace (2000). One thousand health-related quality-of-life estimates. *Medical Care,* 38, 583-637.

12 DRUG POLICY: INSIGHTS FROM MATHEMATICAL ANALYSIS

Jonathan P. Caulkins

Heinz School
Carnegie Mellon University
Pittsburgh, PA 15213
and
RAND Drug Policy Research Center
Pittsburgh, PA 15213

SUMMARY

Illicit drugs create serious health problems whose management is complicated by illegality, poor data, and market dynamics. Quantitative analysis can and does play a key role in clarifying implications of strategic choices concerning collective response to these problems. This chapter summarizes key arguments and findings concerning the effectiveness of various prevention and treatment strategies, including supply control measures. Among them are that conventional prevention programs are not very effective in an absolute sense, but they are so cheap that they are cost-effective. Likewise, treatment programs can be cost-effective despite very high relapse rates, in part because periods of heavy use impose such enormous costs on society. Enforcement can play a key role in diffusing the positive feedback loop created by contagious spread of initiation during the early phase of new drug epidemics because of its unique ability among diverse drug control interventions to focus its impact on the present.

KEY WORDS

Drug control, Optimal control, Resource allocation, Epidemic control

12.1 INTRODUCTION

Illicit drug use is an important health problem. Some 600,000 emergency department episodes in the US every year are related to illicit drugs [1]. National mortality estimates are not available, but probably on the order of 20,000 drug-induced deaths occur each year [2], with many more indirectly related to drug use. Some 5 million Americans are in need of drug treatment, but fewer than 40% receive it [3, 4]. Injection drug use is a leading cause of the spread of infectious diseases such as HIV/AIDS and Hepatitis C [5]. The social costs of illicit drug use approach those of alcohol and tobacco [6-8]. No one has estimated how many quality-adjusted life years are lost due to illicit drug use, but the number is no doubt substantial, particularly since those who die from illicit drug use are younger than those who die from most other causes.

Not surprisingly, there is an energetic debate concerning how best to control drug use and related consequences. Operations research and management science have made important contributions to this debate. However, drug policy is unlike other health policy domains in important ways. This chapter begins with a review of some important differences. The following sections then highlight key insights that quantitative models have generated concerning the relative effectiveness of different interventions, including how that effectiveness varies over the course of a drug epidemic.

12.1.1 How is drug policy different?

Drug policy differs from other health policy domains in a number of respects. First, we care as much about outcomes for other people as we do about outcomes for the person with the "condition". Cancer generates health consequences for people other than the patient, such as stress and depression among family members. Nevertheless, the focus of cancer treatment and policy is clearly and appropriately on the people who have cancer.

The consequences of drug use are more diffuse. Fear that addicts or addicts' suppliers will hurt non-users is an important source of public concern about drug use. One can argue that such fears are exaggerated. However, drug use has other health consequences for non-users that are under-appreciated. For example, addiction of all kinds, including to illicit drugs, is an important contributor to child abuse and neglect.

In this respect, drug policy is more like a public health problem than a medical problem, and the behavioral component invites comparisons to second-hand smoke and drunk driving accidents rather than malaria or cholera. However, drug use is very much a "contagious" phenomenon that can usefully be studied by epidemic models, as will be discussed below. So, analyzing drug policy merges important strands from behavioral health and the contagious disease aspects of public health.

Another fundamental difference between drug policy and other health policy problems is that the underlying activity is illegal. This has myriad ramifications, ranging from making data collection difficult to the fact that law enforcement plays an important role in controlling the prevalence and consequences of this “health problem.”

An important consequence of drugs’ illegality is the existence of black markets, which are the proximate source of many drug-related harms [9]. Such markets would not exist were it not for the drug use, and reducing drug use (e.g., through treatment) shrinks the markets. Thus, a systems analysis should consider market outcomes. The need to do so distinguishes drug policy from other health policy domains. There is no market for heart disease, and with some exceptions, such as so-called “nuisance bars” [10], the markets for tobacco and alcohol are not themselves a major problem.

A subtle consequence of the illegality of drug use is that it encourages the lumping together of all types of use because they are all the same in the eyes of the law. Not only does this blur distinctions between substances with very different health risks (the Drug Enforcement Administration places both marijuana and heroin in its most restricted – “Schedule I” – category of drugs), but it also blurs distinctions between dependent and non-dependent use.

Drug dependence is a well-defined medical condition that can be diagnosed and treated. Recreational use by non-dependent persons is not a well-defined condition, and more often than not it does not lead to dependent use. Thus the vast majority of people who use an illicit drug never have a drug-related medical condition (even though they help support a drug market that generates adverse health outcomes for others).

To complicate matters further, many of those with this medical condition deny they have it and/or are ambivalent about getting rid of it. This attitude contributes to very low compliance with treatment. It has become common to point out that compliance rates (e.g., rates of testing negative for drugs) are not so different from rates of compliance with medical regimens for conditions such as hypertension or diabetes (e.g., admonitions to alter one’s diet). However, few diabetics want to have diabetes, whereas quite a few drug users are not sure they want to stop using drugs. Furthermore, many dependent users do not have a health insurance company or personal physician vested in addressing their dependence.

More such differences could be noted, but these suffice to make the basic point that drug policy is a part of health care policy that necessarily must draw its "system boundary" quite broadly. One could examine just "drug treatment policy," focusing on issues such as queue management and matching treatment modalities to patients (e.g., [11]), but that is a different topic.

12.1.2 Scope and methods of analysis

This chapter focuses on insights from quantitative analysis of "strategic" drug policy choices. At the highest level, this paradigm views drug policy as a resource allocation problem. Some governmental entity decides how many resources to allocate to drug control and how to divide those resources across broad programmatic areas in order to achieve the greatest impact. Such analysis is helpful because a variety of drug control strategies exist, and the drug "system" is complex, so it is not intuitively obvious what the best combination of strategies is.

Analyses of this sort began to appear in the 1970s in response to the heroin epidemic, (e.g., [12-14]), and became more common after the spread of cocaine. Early contributors to this second wave included groups at the RAND Corporation's Drug Policy Research Center [15-24], UCLA (e.g., [25-27]), Carnegie Mellon University's Heinz School [28-33], and later the Technical University of Vienna (e.g., [34-39]) as well as individuals elsewhere (e.g., [40, 41]), with growing communities of analysts elsewhere in Europe (e.g., [42-44]) and notably in Australia [45-48]. There is also an extensive and fascinating literature on modeling HIV/AIDS (e.g., [49, 50]) that intersects with injection drug use, but for reasons of space these issues are dealt with very briefly here.

The method employed in this chapter is simply to skim from this literature insights that can be communicated effectively without detailed technical exposition of the underlying models and analysis, with some bias toward results that the author has observed to be compelling to policy makers and non-academics.

The methods employed in the underlying literature are diverse, but mainly involve construction of some nonlinear descriptive model of the behavior of drug users and sometimes sellers, with inputs corresponding to various policy alternatives. Depending on the sophistication of the model and associated analysis, the models are then used to reproduce past and present behavior and/or to make recommendations for the future, either in "what if" policy simulation mode or through some formal optimization.

Collectively, the greatest weakness of the literature is the inability to truly validate these models given the paucity of reliable data and the inherently small sample size when the unit of analysis is at the national level. A consequence is that specific numerical results are not very precise; the models are more reliable for general structural insights of the sort offered below. The models' greatest contribution stems from precision of a different sort, the precision and rigor that comes from translating less quantitative scholars' mental models into equations, from which powerful insights often emerge from relatively simple analysis.

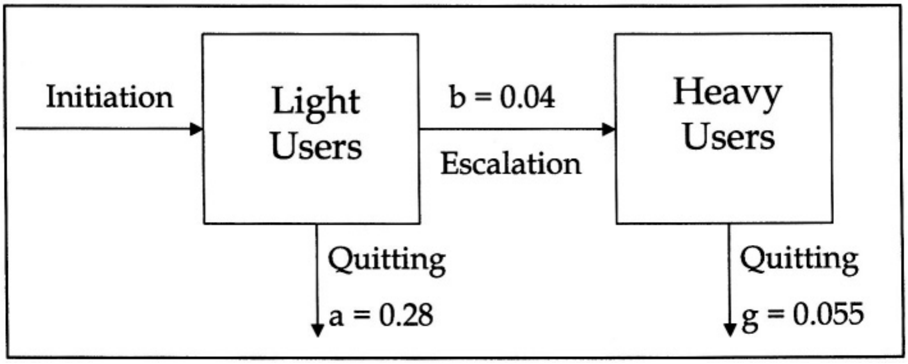
12.2 RESULTS

A number of insights emerge directly from models of use, without explicit consideration of specific control measures. We begin with a few such insights before considering results from models of prevention, treatment, enforcement, and drug epidemics.

12.2.1 Models of use

Everingham and Rydell [23] made a pioneering contribution to understanding of drug policy by developing a simple two-state Markov model of cocaine demand that distinguishes between so-called "light" and "heavy" users. Figure 12.1 illustrates a modified version of the model with flow rates recently updated by Knoll and Zuba [51].

Figure 12.1 Everingham and Rydell's light and heavy user model [23] with flow rates updated by Knoll and Zuba [51]



Several insights emerge from this simple model. For example, most (roughly 5/6th) of those who try cocaine do not escalate to heavy use, but those who do persist in the heavy use state for many years (1/g = 18 years). Per capita consumption rates for heavy users are much higher than for light users.[1] As a result, expected lifetime consumption per initiation is on the order of 225 – 475 grams [52] – but that figure is dominated by a very large expected consumption given escalation to heavy use multiplied by a small probability of escalating. Given the high social cost per gram consumed,[2] this implies that the expected social cost per cocaine use career is very large (on the order of $50,000 - $100,000 per initiate). The median cost is at least an order of magnitude lower, if not two. Indeed, given that many users appear never to proceed beyond the "very light" stage [53], the median cost per cocaine use career could be close to 0.[3]

Highly skewed consumption and social cost distributions are not unique to cocaine. The average marijuana use career involves 375 – 875 grams of consumption, or 1,000 – 2,000 joints, whereas median lifetime days of marijuana use are less than 100 [54]. For heroin in the US, per capita consumption rates are an order of magnitude lower than for cocaine but the social costs per gram are an order of magnitude higher [54] and exit rates from dependent use no higher [55]. Thus, the expected social cost for someone who escalates to dependent heroin use is at least as high as is the corresponding figure for cocaine. Indeed, inasmuch as light use of heroin is more likely to involve smoking and heavy use to involve injecting, the skew in social cost could be greater than the skew in social cost for cocaine.[4]

The sharply different exit rates for light and heavy users (factor of 5 difference in Figure 12.1) and the lag between initiation and escalation to heavy use means that the character of drug use can vary sharply over the course of a drug epidemic, as illustrated in Figure 12.2 [51]. Demand for cocaine[5] rose sharply with cocaine initiation in the mid- to late 1970s, but did not fall when initiation fell in the 1980s. Rather, sharp declines in the number of light users were offset by numerically smaller increases in the

[1] Everingham and Rydell suggested a ratio of 7.25 to 1. More recently, Abt analysts estimated that "chronic" users spend about 6 times as much per capita as do "occasional" users [94].

[2] Rydell and Everingham [24] originally estimated average social costs of about $100 per gram, but Caulkins et al. [54] use newer evidence to develop a figure of $215 per gram.

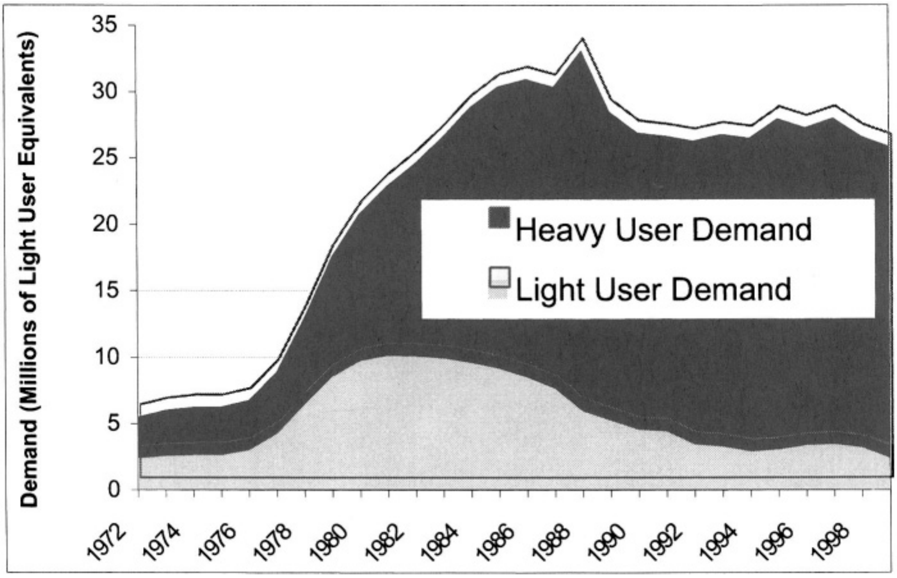
[3] Kaya et al.'s [48] analysis of Australian heroin use data shows that the number of people quitting heroin use over time is highly correlated with the number initiating, presumably because the modal career of use is very short.

[4] A tendency for light users to smoke and heavy users to inject heroin would reduce the skew in terms of grams of heroin consumed because injection is the more "efficient" route of administration, although dependent users are also more likely to have developed tolerance and to take larger effective doses.

[5] Demand is proxied by the sum of light and heavy users weighted by their relative propensities to consume.

number of heavy users, whose per capita consumption rates are much higher, leaving overall demand substantially stable for close to two decades. The proportion of that demand attributable to light vs. heavy users changed dramatically, however. In 1980 much of the demand came from light users who are not badly affected by their drug use. By 1990 most of the demand came from heavy users, many of whose lives are dominated by the drug. This picture follows fairly directly from the simple Markov model of use, yet it is so compelling that the Office of National Drug Control Policy incorporated [23] an earlier version of it in several of its National Drug Strategy Reports (e.g., [56]).

Figure 12.2 Evolution of cocaine demand in the US, expressed in millions of light users or their equivalent, assuming heavy users consume seven times as much per capita as light users



These figures also provide a convenient way of thinking about various drug control interventions. Primary prevention programs seek to reduce initiation. Treatment seeks to increase quitting from heavy use. Enforcement programs seek to do both and also reduce per capita consumption by current users by deterring users and constraining supply. Epidemic models analyze how the relative effectiveness of these interventions varies over the course of drug epidemics such as the one depicted in Figure 12.2. We turn next to a discussion of key insights relating to these programs.

12.2.2 Reducing demand through prevention

There has been great confidence that drug prevention is effective and cost effective. For example, the 1999 national drug strategy [57] stated unequivocally that, “The simplest and most cost-effective way to lower the human and societal costs of drug abuse is to prevent it in the first place.”

However, there is enormous heterogeneity in programs, ranging from adventure camps to mass media campaigns. Some are more effective than others [58]. Experimental trials have shown some school-based programs to decrease illicit drug use [59-61], yet the most popular school-based program, the Drug Abuse Resistance Education or DARE program, has not been shown to have any material effect on marijuana use [62].

Furthermore, the experimental evidence pertains only to self-reported use of indicator substances (e.g., marijuana), through final followup data collection (typically 9th or 12th grade) by people in the program. However, from a policy perspective one is interested in the impact on actual (not self-reported) use of the more damaging illicit drugs (e.g., cocaine) over the lifetime of all people affected, including those not in the program.

Caulkins et al. [52, 54] developed mathematical models for projecting total impact of school-based prevention programs based on available evidence concerning “best-practice” programs. There is considerable uncertainty concerning the projections, but the bottom-line finding is that these programs are cost effective, though not very effective.

Drug prevention is not very effective if one compares it to conventional childhood vaccination. If one gives the very best prevention program to a group of youths who would have used drugs, most will go ahead and use drugs anyhow. Even cutting edge school-based programs only reduce marijuana use by 5-15%, and for almost all programs those effects decay by the end of high school. Even recognizing that delayed initiation is associated with lower lifetime use, this translates into reductions in the present value of lifetime consumption in the single digits, and most likely just a few percent, because reductions in lifetime use are only one-fifth to one-third as great as the reductions observed immediately following program completion [54].

Thus prevention cannot be “the” solution to the drug problem. Indeed, the notion that enforcement merely needs to “hold the line” until prevention can “cut the legs out from under the epidemic” does not seem realistic given that the problem is now more endemic than epidemic. It is similarly unrealistic to hope, as some drug legalization advocates suggest, that funding drug prevention with the money saved by not having to enforce the prohibition would offset any legalization-induced increase in use.

On the other hand, prevention is cheap, even if one recognizes that the dominant societal cost of running school-based prevention programs is the opportunity cost of not using class time to teach other academic subjects. The outright budgetary costs for programs delivered by regular teachers who are already on the payroll is tiny. Since preventing drug use is so valuable and prevention so inexpensive, prevention is cost effective even though it is not very effective.

One "paradox" of prevention that an OR/MS analysis reveals is that only about one-quarter of a prevention program's impact on cocaine use comes from preventing program participants from initiating use [52]. Some impact comes from reduced use by program participants who do initiate and use at some level. Still only about one-third of the reduction in consumption is in the form of reduced consumption by program participants. Two-fifths comes from positive spillover to friends and associates of those in the program, and one-fifth comes about because reduced use by all these people shrinks the market, making enforcement against those who remain in the market more effective. Thus, conventional evaluations of prevention that focus on abstinence for program participants miss some two-thirds to three-quarters of the effects that would be assessed by a systems analysis.

Another interesting insight is that "school-based prevention should be done 15 years before one knows we need to do prevention" [52]. The average age of initiation of "hard" drugs is about eight years after the age targeted by school-based prevention programs. National recognition of a drug epidemic may occur five or more years after the peak in initiation. (US cocaine initiation peaked in 1979; it was recognized as a national crisis around 1984.). Since it takes time to appropriate funds, adapt and scale up prevention programs, and so forth, this implies that school-based prevention must be started about 15 years before it is widely appreciated that prevention is needed! Given the contagious nature of drug epidemics, prevention programs implemented before the beginning of an epidemic are likely to be many times more effective than programs implemented after the epidemic has matured.[6] Since ability to forecast drug trends is exceedingly limited, the practical implication is that prevention programs should be funded on an

[6] Precise statements are difficult because effectiveness ratios are sensitive to the number of heavy users at the beginning of the epidemic, a parameter for which data are particularly weak. See Winkler et al. [110] for more on how the relative effectiveness of different types of drug prevention varies over the course of an epidemic.

ongoing basis, not in response to current crises. Decisions about prevention should not be made only with an eye toward ameliorating the current epidemic. Instead, prevention should be seen as "lending a hand" in reducing the current drug epidemic and possibly other undesirable social trends, while also serving as a form of inexpensive insurance against possible future epidemics.

Likewise, drug prevention should be – and usually is – generic, not drug-specific. Indeed, less than half of the social benefits of school-based drug prevention stem from reduced use of illicit drugs. The majority stems from reductions in smoking and heavy drinking [54].

12.2.3 Reducing demand through treatment

Treatment is the most thoroughly evaluated drug control intervention. Indeed, the literature is so large there is even a bibliography just of other literature reviews of drug abuse treatment effectiveness [63]. Most observers conclude that drug abuse treatment is cost effective (e.g., [24, 64]). The Institute of Medicine [65] summarized the literature by saying, "Research has shown that drug abuse treatment is both effective and cost effective in reducing not only drug consumption but also the associated health and social consequences." On the other hand, a National Research Council Report [66] subsequently attacked the existing data on treatment as vulnerable to various methodological biases, concluding that, "There is little firm basis for estimating the benefit-cost ratio or relative cost effectiveness of treatment." The principal complaint was that few true randomized controlled trials had been conducted.

What is clear to a systems analyst, though not necessarily a social scientist, is that decision-relevant insight can be gleaned even if it is not possible to produce a bottom line benefit-cost ratio. For example, one can work backwards to ask how effective treatment must be to be cost-justified. If the resulting breakeven effectiveness seems implausibly high, one would be skeptical that treatment is a good investment. If it seems attainable, one might be more optimistic.

Rydell and Everingham [24] in fact performed such exercises. One of their striking findings was that even if every treatment client relapsed immediately after completing treatment, treatment could still be cost effective! The full model is too involved to explain here. It tracks cocaine as it is produced and passed through multiple distribution layers, and explicitly models user flows, prices, and market dynamics over a 15-year planning horizon.

Back-of-the-envelope calculations are sufficient to convey the same basic insight, as we now demonstrate. In terms of the drug use model above, treatment can be thought of as doing three things: it can suppress use while the person is in treatment, it can reduce use between exit from treatment and relapse, and occasionally it may encourage some users to quit permanently. Rydell and Everingham's startling insight is that the first mechanism alone can be enough to make treatment a good investment.

Rydell and Everingham [24] estimated that the average admission to treatment costs about $2,000, the average time in treatment is 3 months, and use is suppressed by about 80% during treatment. If heavy users consume at a rate of 120 grams per year, the average admission averts about 120 * (3/12) * 0.8 = 24 grams of consumption through this "incapacitation" effect alone. Harwood et al.'s [8] cost of illness study estimated that the total social cost of illicit drugs in the US in 1992 (excluding impaired productivity) was $83.5 billion. Apportioning this by substance, dividing by Rydell and Everingham's [24] estimate of 291 metric tons of cocaine consumption in 1992, and adjusting for inflation, Caulkins et al. [54] roughly estimate an average social cost of $215 per gram of cocaine consumed in the US. So Rydell and Everingham's implied social benefit per treatment admission (24 grams * $215/gram = ~$5,000) exceeds the roughly $2,000 cost.[7] Indeed, the benefit-cost ratio would be greater than one even if every user relapsed immediately after leaving treatment and treatment only suppressed use by one-third during treatment (120 * (3/12) * (1/3) * $215 > $2,000).

One can do a similar breakeven calculation with respect to treatment's impact on exit rates. Suppose the present value of the residual career length of the average treatment entrant is 8 years. (In Figure 12.1 's Markov model the undiscounted residual career length would be 1/g = 1/0.055 = 18 years, but one should discount back to the present and truncate to recognize that people – especially chronic drug users – do not live forever.) If the social cost per year of use is approximately 120 grams/year times $215/grams = $25,000 per year, then the discounted social value of averting a present value of 8 years of such use by getting a heavy user to quit is about $200,000. Hence, if even 1% of treatment admissions led to permanent cessation, the present value of treatment's benefits would equal its costs.

[7] Of course the social cost per gram of consumption averted by treatment could in theory be below average cost, but more likely it is higher. The biggest danger from light use of cocaine is the possibility of escalation to dependent use, and since many of those in treatment are "referred" by the criminal justice system, consequences of their use may be costly even relative to those of other heavy users.

Similarly, if one in 12 people entering treatment ceases use for a year (and no one quits permanently and no one else reduces use during treatment) the benefits would exceed the costs. Any linear combination of these three effects would also lead to a breakeven benefit-cost ratio. For example if one client in 20 did not relapse for a year and one in 250 quit, treatment's benefits would exceed its costs even if the treatment had no impact whatsoever on 95% of clients.

Pollack [67] has taken this insight a step further, noting that methadone maintenance (a treatment for heroin users, who often use by injection) can have benefits that exceed its costs even if it gets no credit at all for reducing drug use – simply because it can reduce the rate at which users spread HIV by sharing syringes.

More generally, interventions can reduce drug-related harm and have positive social benefit-cost ratios even if they do not reduce drug use. Indeed, treatment is sometimes described as a "hook" for getting needy people in contact with health and social service agencies. Such a "harm reduction" approach to drug control is common outside the US [68], although it has not been the subject of much formal systems analysis.

As Manski et al. [69] argue, in the absence of rigorous randomized controlled trials it is not possible to conclude with certainty that treatment is cost effective, but what is clear from Rydell and Everingham and other's work is that the breakeven effectiveness values are not very high and that relapse rates are not an adequate metric for evaluating the value of treatment. Hence, Manski et al.'s complaint that, "When complete and permanent abstinence is used as a criterion of success, between 60 and 90 percent of clients relapse to drug use within 12 months of treatment," [69] does not seem altogether damning.

The work of Rydell and Everingham also provides a cautionary note. If most people relapse, then unless those individuals can be re-enrolled rapidly, there is a limit to how quickly treatment can ameliorate the drug problem. In Rydell and Everingham's model (which assumed that 13.2% of treatment entrants left heavy use because of that treatment, with two-thirds merely de-escalating to light use), even if every heavy cocaine user received treatment once a year, cocaine use would still only be cut in half over 15 years. Furthermore, Rydell and Everingham did not consider the possibility that such an expansion in treatment might have an adverse feedback effect on initiation, as do Behrens et al. [70, 71]; such an effect would make programs less effective. Highly imperfect treatment programs, no matter how cost effective, cannot quickly eliminate an endemic drug problem. Everingham and Rydell [24] and Caulkins et al. [52] make similar points concerning prevention.

12.2.4 Reducing supply

Interventions can affect supply in two ways. Unanticipated interventions can disrupt the market equilibrium. Ideally the disruption takes the form of physical shortage, and the market does not regenerate, but that is not the norm. Usually suppliers adapt, although prices may spike and use decline in the interim [72]. At one time or another over the last 30 years, four different regions have been the principal supplier of heroin to the US (Mexico, South America, Southwest Asia, and Southeast Asia). Similarly, Colombia quickly replaced Mexico as the principal supplier of marijuana to the US in response to paraquat spraying and fears of adverse health-effects of using sprayed marijuana [73].

Enforcement can also affect supply even if the intervention is fully anticipated. For example, if smugglers knew that one-quarter of all shipments would be seized, they would ship more than if they thought none would be seized. Indeed one of the early lessons that drug policy analysis gave policy makers was that quantity seized is not a direct measure of enforcement's impact on consumption [72]. However, presumably smugglers would charge more per kilogram landed to make up for their losses. The higher prices represent a shift in supply that affects retail prices and, hence, consumption.

At one time demand was thought to be insensitive to price, but the price elasticity of demand for illicit drugs turns out to be rather high, much higher than for cigarettes. (For a review of the literature, see Chaloupka and Pacula [74]). Nor does this price-responsiveness seem to be confined to light use reported in surveys. Crane et al. [75] estimate that the elasticity of cocaine emergency department mentions with respect to price is –0.63, and Caulkins [76] notes that a simple constant elasticity model predicts emergency department mentions for both cocaine (elasticity –1.3) and heroin (elasticity –0.8).

These disequilibrium and equilibrium aspects of enforcement's effect on supply are quite distinct, and great confusion can arise if one tries to compare analyses or conclusions concerning one with those concerning another. Supply-side interventions are most likely to have disequilibrium effects if they quickly affect a large proportion of supply. For most drugs, the industry within US borders is populated by many vertically disaggregated "firms," so it is difficult for enforcement to remove a large proportion of the national domestic distribution network's capacity at any one time [77]. Furthermore, the network is robust because of its many lateral linkages, independent paths, and ability to expand quickly the capacity of individual arcs [78].

Interventions in source countries can have greater potential for market disruption because there is greater market concentration there. Perhaps the greatest success occurred when the Turkish Opium ban, the breaking of the "French Connection" case, and Mexican opium eradication substantially drove up purity-adjusted heroin prices during the mid- to late 1970's, before Asian heroin filled the gap [79]. The greatest success in disrupting the cocaine supply was the result of a combination of US efforts and the "war" between the Colombian government and the Medellin-based traffickers in 1989 which led to a sharp (50-100% at its peak) but short-lived (about 18 months) increase in cocaine prices [80]. In 1995, Peruvian interdiction of the "air bridge" to Colombia led to a smaller but identifiable increase in cocaine prices [75].

There is reason to believe that transient price increases can have meaningful effects. The heroin scarcity in the 1970s coincided with the ebbing of the heroin epidemic [81]. Emergency room and medical examiner mentions declined in parallel with higher cocaine prices in 1989-1990 [82], and there was a one-period (three month) decline in emergency mentions in late 1995 [83]. Some, however, argue that market disruptions can increase harms through unsafe use (e.g., more needle sharing) and greater market violence [84, 85].

There have been only a few analyses of the consequences of short-term disruptions (e.g., [75, 86]) and no serious estimates of the cost of generating disruptions. Hence, few real cost-effectiveness insights exist. This is clearly an area worthy of further research.

There have been far more studies of how enforcement might affect the long-run market equilibrium. Such analyses use so-called "risks and prices" calculations of the sort pioneered by Reuter and Kleiman [87]. The "risks and prices" paradigm recognizes that increasing enforcement risks for dealers raises their cost of doing business. Dealers could simply absorb those costs, but presumably prefer to pass them along to users in the form of higher retail prices, which in turn reduce consumption [88].

The literature on risks and prices calculations generates a number of insights. For example, when efficiency is defined as kilograms seized per million taxpayer dollars spent, enforcement is more efficient at seizing drugs in source countries and while drugs are being smuggled into the US than within the US. However, suppliers are also more "efficient" at replacing drugs that are seized before they enter the US because the drugs are so much less expensive in the source and transshipment countries. Unfortunately, when moving upstream in the supply chain, the "efficiency" gain for the suppliers trumps the efficiency gain for the interdictors. Hence, the effective cost to suppliers of replacing the drugs seized per million taxpayer dollars spent is lower, not higher, outside the US. Replacing seized drugs is just one of many components of the "tax" that enforcement imposes on equilibrium operations. For example, Rydell and Everingham's [24] model also considers seizure of assets, arrest, imprisonment, incarceration of sellers who are also users, and indirect effects on production costs. However, even when considering all components of the tax, the same basic pattern persists. The cost imposed on suppliers per million taxpayer dollars spent on enforcement is lower outside the US than it is within the US.

Hence, the only way international operations can be a more cost-effective "tax" on suppliers is if the tax is "multiplied" as the drugs move down the distribution chain. Boyum [89] and Caulkins [90] suggest reasons why there might be such multiplicative price transmission. Caulkins [80] finds some evidence for this proposition, but DeSimone [91] suggests variation by drug. It may be easier to create transitory disruptions through international operations, but unless a multiplicative price transmission model holds, it is harder for such enforcement to drive up equilibrium prices [79].

Within US borders, the risks and prices model has something of the feel of an arm's race. If enforcement can impose enough cost on the suppliers per taxpayer dollar spent, it could be cost effective. Most analyses find that it is costly to fight this arms race in a mature market, as an excerpt from a simple static portion of Caulkins et al.'s [92] model suggests.

Assume that the demand curve can be locally linearized with a known elasticity η, that the market is in equilibrium in the sense that suppliers' revenues just cover costs, including normal profits, and that the industry supply curve stems from the following cost structure. "Normal" business costs per unit increase linearly in volume (i.e., they follow a textbook upwardly sloping linear supply curve), but there are two additional costs: (1) costs imposed by enforcement, including compensation for the risks of arrest and imprisonment, and (2) costs that are linear in the dollar value of the drugs distributed, not their weight. The last term is important because drug distribution is almost pure brokerage activity, requiring minimal processing, and the drugs weigh next to nothing per unit value. (Cocaine and heroin sell at retail for about ten and one hundred times their weight in gold, respectively.) Thus the suppliers' costs of delivering drugs can be written as

$$\text{Total cost} = (c_0 + c_1\,Q)\,Q + E + \gamma\,(P\,Q),$$

where P and Q are the market clearing price and quantity, E is the enforcement "tax", and c_0, c_1, and γ are positive constants. With a little algebra [53] it is easy to show that shifts in demand and the enforcement tax have the following effects on the market equilibrium:

	If there is a:	
Resulting Percentage Change in	1% Increase in Costs that Enforcement Imposes on Dealers	1% Increase in Demand
Consumption	$\frac{\beta\eta}{1-\gamma+(\beta-\alpha_1)\eta}$	$\frac{1-\gamma}{1-\gamma+(\beta-\alpha_1)\eta}$
Price	$\frac{\beta}{1-\gamma+(\beta-\alpha_1)\eta}$	$\frac{\alpha_1-\beta}{1-\gamma+(\beta-\alpha_1)\eta}$
Spending	$\frac{\beta(1+\eta)}{1-\gamma+(\beta-\alpha_1)\eta}$	$\frac{1-\gamma-(\beta-\alpha_1)}{1-\gamma+(\beta-\alpha_1)\eta}$

where $\alpha_0 = c_0 \;/\; P_0$, $\alpha_1 = c_1\, Q_0 \,/\, P_0$, and $\beta = E_0 \,/\, P_0\, Q_0$ are the fractions of dealers' costs in the current equilibrium that are attributable to the linear part of the cost term above, the quadratic part of the cost term, and enforcement, respectively. Since γ is the remaining fraction, $\alpha_0 + \alpha_1 + \beta + \gamma = 1$, so one of these parameters (α_0) can be eliminated in the expressions above. Caulkins et al. [93] estimate that for the US cocaine industry in 1992, $\eta = -1$, $\alpha_0 = 0.55$, $\alpha_1 = 0$, $\gamma = 0.25$, and $\beta = 0.2$. Suppose these parameters still applied in 2000, when the Office of National Drug Control Policy [94] estimated that there were 3.035 million occasional and 2.707 chronic cocaine users who collectively spent \$35.3 billion while consuming 259 metric tons of cocaine.

Reducing equilibrium consumption by 1% would require imposing costs of $(\alpha_1 - \beta - (1 - \gamma)/\eta) * 1\% * \$35.3$ billion = \$194 million on suppliers. The cost to taxpayers to "purchase" this cost-imposition depends on how efficient enforcement is. Consider a policy of giving longer sentences to people who already would have been convicted and incarcerated at least briefly. (Thus we can ignore details of arrest, adjudication, seizures, and so forth.) Suppose drug suppliers have to increase workers' wages by \$50,000 to compensate them for the risk of each additional expected year of incarceration. It costs taxpayers about $25,000 to incarcerate someone for a year [95], so the efficiency ratio is 2:1 and taxpayers could buy that 1% reduction in cocaine consumption for $97 million per year.

Alternately, one could cut consumption by 1% by reducing demand by $1 + \eta (\beta - \alpha_l) / (1 - \gamma) = 0.73\%$. Assuming heavy users consume seven times as much per capita as do light users, that would require eliminating 0.73% * (2.707 + 3.035/7) = 23,000 heavy users. At first this might seem to be the more expensive route: $97 million would only pay for about two treatment admissions per person for 23,000 heavy users. However, the supply reduction strategy requires spending $97 million per year indefinitely. If the 23,000 heavy users were somehow removed by treating each twice, consumption would be reduced by 1% indefinitely (ignoring indirect effects on initiation, which may be a second-order effect in a mature market). At a 4% discount rate, the present value of $97 million per year forever is $2.4 billion, or about $100,000 for each of those 23,000 users, enough for some two-dozen rounds of treatment per person.

There is a sharp distinction between the timing of the costs and benefits of treatment, conventional enforcement, and extending time served for convicted traffickers with mandatory minimum sentences [93]. Raising prices by threatening sanctions brings immediate benefits, since suppliers have to adjust their cost structure in the short run. Secondary, long-lasting benefits also accrue: raising prices today suppresses initiation and increases quitting thereby reducing future demand. So supply-side enforcement's benefits are predominantly upfront. The costs of enforcement with conventional sentences also occur mainly in the first year or two, but the longer the sentence, the longer the period over which costs to taxpayers are spread. Furthermore, if the policy change is one that extends the sentence of someone who would have been incarcerated anyhow, the incremental costs do not begin to be felt until after the end of the baseline sentence. The time profile of treatment costs and benefits is very different. Treatment costs essentially all come in the first year, as do the "incapacitation" benefits of reduced use during treatment. However, the benefits of convincing someone to quit continue to accrue throughout the entire period during which they would otherwise have continued to consume. Informally, conventional enforcement is like paying cash, mandatory minimum sentences are like buying with a credit card, and treatment is like an investment.

Rydell and Everingham [24] and Caulkins et al. [93] examine in detail this issue of the timing of the benefits and costs of various interventions. Roughly speaking, the result is as follows. Suppose a treatment intervention and an enforcement intervention each have the same impact on consumption over the next 15 years, discounting future outcomes at 4% per year. (More specifically, imagine the enforcement operation is one whose effect stems from raising suppliers' cost of operations over the next year.) The treatment intervention would have about double the impact on consumption in the first year as it would in each succeeding year, whereas the ratio for enforcement is about nine to one. The enforcement intervention would have about 2.7 times as great an impact on consumption in the first year as does treatment, whereas in every succeeding year the treatment program would have 1.65 times as much impact as the enforcement program. Hence, although treatment may be the more effective way to reduce use in the long run, enforcement has greater capacity to focus its benefits in the present, a capability that may be invaluable when trying to interrupt the contagious spread of initiation early in a drug epidemic.

One concern with price-raising enforcement is that it might increase crime even if it reduces use. Most drug-related crime is "economic-compulsive" (committed to obtain money to buy drugs) or "systemic" (arising from drug selling, e.g., punishment for non-payment) and so is driven by drug dollars, not by intoxication or use per se (also called "psychopharmacological" drug-related crime). Depending on the elasticity of demand, driving up prices could actually increase, not decrease, drug-related crime. A very simple model of this conveys the basic intuition. Suppose that drug-related crime is proportional to a weighted sum of drug use and spending on drugs, with the latter accounting for 100x% of the total. So drug-related crime C equals

$$C = k_1 P Q + k_2 Q,$$

for some positive constants k_1 and k_2 such that $k_1 P Q = x (k_1 P Q + k_2 Q)$, i.e., $k_1 P / k_2 = x/(1-x)$. Taking the derivative of crime with respect to price gives

$$\begin{aligned}\frac{dC}{dP} &= k_1 Q + (k_1 P + k_2)\frac{dQ}{dP} \\ &= \frac{k_2 Q}{P}\left(\frac{x}{1-x} + \left(\frac{x}{1-x} + 1\right)\frac{dQ}{dP}\frac{P}{Q}\right) \\ &= \frac{k_2 Q}{P(1-x)}(x+\eta).\end{aligned}$$

Hence, driving up prices reduces drug-related crime if the absolute value of the elasticity of demand ($|\eta|$) is greater than x, the proportion of drug-related crime that is driven by drug spending rather than drug use.

Figure 12.3 Relative effectiveness of demand reduction and price-raising enforcement depends on the elasticity of demand

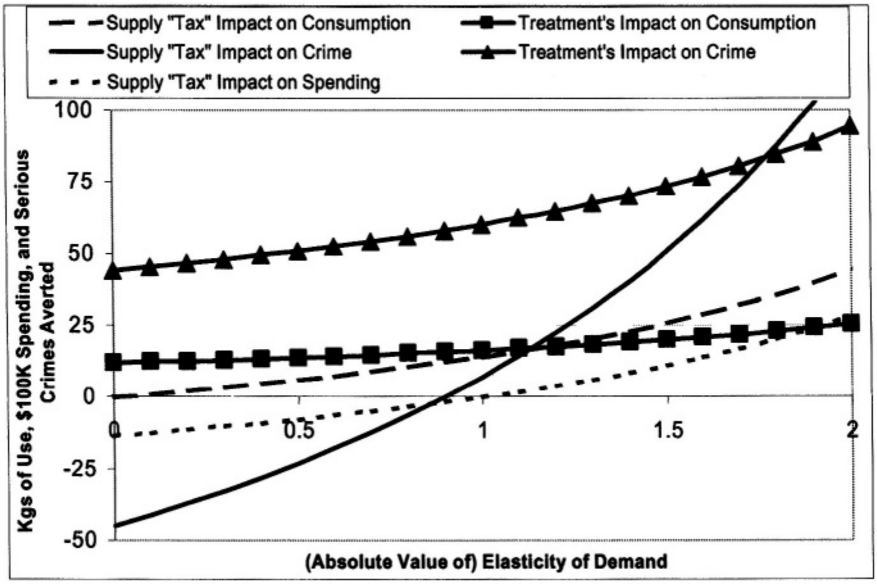


Figure 12.3 uses the market equilibrium model above to illustrate in more detail how the effects of price-raising enforcement and demand reduction on drug use, spending, and crime depend on the elasticity of demand.[8]

12.2.5 Dynamic/epidemic modeling results

Drug use varies dramatically over time, driven in no small part by endogenous nonlinear dynamics, not just in response to changes in policy or exogenous factors such as the poverty rate. Hence, one would expect the effectiveness of interventions to likewise vary with the state of the epidemic, and a growing literature investigates this possibility. According to this school of thought, it is rarely sensible to make statements such as "treatment is better than enforcement" or vice versa without qualifying the statement (e.g., "treatment is better than enforcement for controlling cocaine use in the US now that the epidemic has plateaued").

[8] Parameters from Caulkins et al. [92] and assuming 5/6th of drug-related crime is driven by spending. The enforcement intervention is imposing $1 million in costs on suppliers. The demand reduction intervention is eliminating 100 heavy users.

Perhaps the most important endogenous dynamic is the “contagious” character of drug initiation. Unlike infectious diseases, drug use has no pathogen, but drug use is contagious in the sense that drug use spreads when non-users are introduced to the drug by current users. (Contrary to once popular myth, most initiation does not stem from dealers “pushing” the drugs on potential users; rather, new users are initiated by current users.) In formal terms, there is a positive feedback from current use to initiation. Systems with such a feedback can grow explosively.

There are several models of how that explosive growth ends, depending in part on what country and drug is being modeled. In a line of modeling pioneered by Tragler [95-98], a steady state emerges when quitting at a constant per capita rate balances initiation, which is an increasing but concave function of use.

In a line of models pioneered by Behrens [70, 71, 99] the key negative feedback pertains to the drug’s reputation. As some early initiates progress from light to heavy use, the drug’s dangers become apparent and initiation declines. That decline, coupled with the high quit rates for light users, increases the ratio of heavy to light users, further enhancing the drug’s negative reputation and cutting initiation. These models can, for some parameter values, generate recurrent cycles of drug epidemics. Almeder [100] examined a related family of age-distributed models in which the nature and intensity of this feedback depends on the relative and absolute ages of the users and potential users.

In a line of models associated with Rossi and colleagues (e.g., [42, 43]), the limiting factor is the number of susceptibles. To over-simplify, essentially everyone who might try the drug ends up trying it. Most use only briefly, but some get hooked, so after the explosive growth stage there is a decline to an endemic problem characterized by a high proportion of heavy users.

The overall policy prescription from these models is to rely on enforcement early in a drug epidemic and rely on treatment later in the epidemic. Prevention can be extraordinarily cost effective if done before and at the beginning of an epidemic; later it is much less effective, but is still worth doing. In particular, keeping prices high initially is a useful way to slow the explosive spread of drug use, but later on high prices are costly to maintain and may exacerbate drug-related crime. More generally, one should initially fight very aggressively to contain a drug epidemic. Ideally the epidemic would be eradicated or stabilized at low levels, but if the intervention is too late or the epidemic growth too great, then one should accommodate the growth in drug use by gradually shifting to strategies that remove heavy users and/or ameliorate the social cost per heavy user.

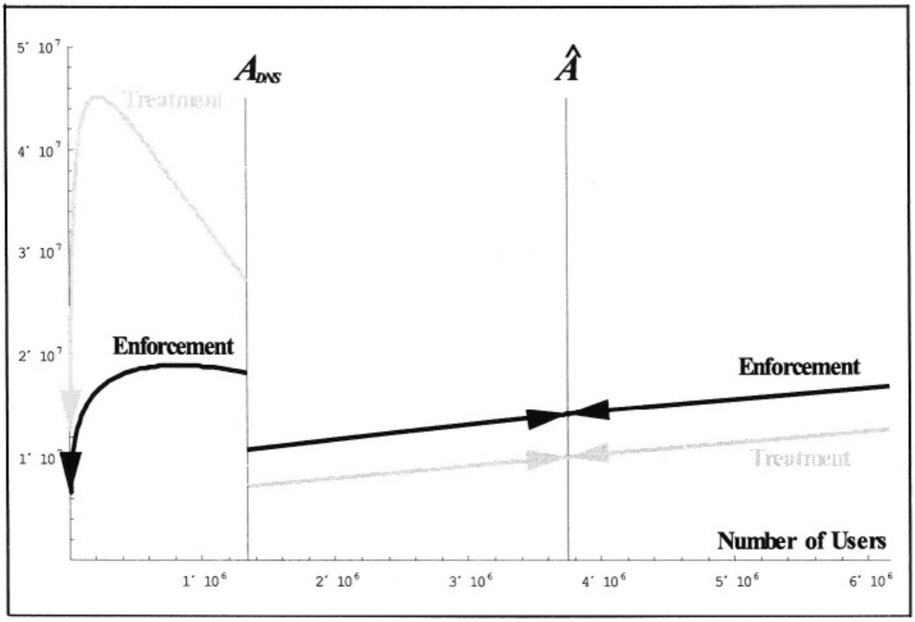
Tragler et al. [96] offer an example of such a finding. Their model seeks the optimal dynamic levels of price-raising enforcement and drug treatment that together minimize the present value of the sum of control spending plus the quantity of drugs consumed, weighted by the social cost per unit of consumption, subject to drug use evolving according to the following nonlinear model of drug use (modeled by a set of differential equations): Initiation is concave in use. "Natural" quitting is at a constant rate per capita, which can be augmented by treatment, although with diminishing efficiency as the proportion of users in treatment increases. Prices affect all flow rates and are in turn a function of the intensity of price-raising enforcement as above.

Figure 12.4 updates a figure from Tragler et al. [96], using a slightly larger exponent on endogenous initiation in light of Grosslicht's [101] findings. The horizontal axis depicts the number of users and the vertical axis gives the optimal annual control spending (in thousands of dollars). A so-called Dechert-Nishimura-Skiba threshold (labeled A_{DNS}) occurs when the number of users is about 1.3 million. If the initial number of users is less than this threshold value (i.e., control begins before the epidemic has passed this point), the optimal strategy is to use massive levels of enforcement and treatment to reduce use to some minimal level. Otherwise, it is optimal to let use grow toward a positive equilibrium. In that case, enforcement and treatment spending should increase with use, but with the proportion of control spending allocated to treatment increasing over time. This finding of a sharp choice between eradication and accommodation at the aggregate level is consistent with others' analyses of the impact of enforcement on local drug markets (e.g., [36, 40, 41, 102]).

A key driver of this dynamic is "enforcement swamping" [103]. The deterrent or price-raising potential depends on enforcement's intensity – i.e., the amount of enforcement per kilogram or per person in the market – not the absolute level of enforcement. Early in an epidemic, when the market is small, it is not so hard to achieve high enforcement intensity. When the market doubles in size, the intensity generated by a given enforcement level is halved because that enforcement is spread over a larger target. Since drug use can much more than double over an epidemic, overcoming this dilution for an established mass-market drug is very expensive.

One of the more interesting insights to emerge from these optimal control models comes from Behrens et al.'s [70, 99, 104] complementary analysis of prevention and treatment. It extends Everingham and Rydell's [23] model of cocaine use in Figure 12.1 to make initiation increasing in the number of light users and decreasing in the number of heavy users. Insights derived from this model include the following: (1) Prevention is most valuable when

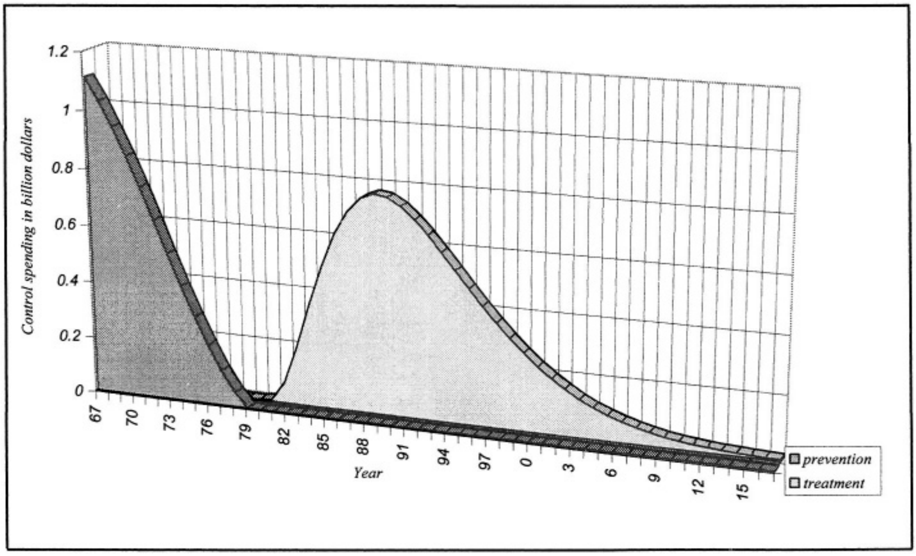
Figure 12.4 Optimal control spending as a function of the number of users, illustrating Tragler et al.'s finding that if control catches the epidemic early it should seek to "eradicate" the epidemic; otherwise accommodation is the optimal strategy



there are relatively few heavy users, such as in the beginning of an epidemic. Treatment is more effective later. (2) The transition period when it is optimal to use both prevention and treatment is very brief. (3) Total social costs increase dramatically if control is delayed.

The second insight is particularly interesting because many people describe the strategic drug policy choice as concerning supply-side vs. demand-side interventions. Behrens et al. show that it is misleading to lump together treatment and prevention even though they both affect demand. At any given point in an epidemic, prevention might be very valuable but not treatment or vice versa. Indeed, when Behrens et al.'s model is parameterized for the US cocaine epidemic and school-based prevention (which has a roughly 8-year lag between program spending and effect on initiation), it is literally never optimal to spend money on both prevention and treatment! This is illustrated in Figure 12.5, which is adapted from Behrens et al. [70]. A complete absence of overlap is not robust with respect to parameter variation, and as discussed above, prevention is probably justified on an ongoing basis because of its impact on the use of other drugs. Nevertheless, the general message is robust: It is simplistic to argue for or against "demand-side" (or "supply-side") strategies without knowing more about the specific mix of strategies *and* the current state of the epidemic in question.

Figure 12.5 Optimal cocaine control spending levels over time for school-based prevention and treatment for the past US cocaine epidemic



12.3 OPPORTUNITIES FOR FURTHER RESEARCH

Drug policy is an important domain. It has enough nonlinearities from epidemic feedback and market dynamics to challenge unguided intuition, so formal mathematical models such as those reviewed here can be a very important aid to strategic planning. There remains, however, far more that is not known than is known, so possibilities for further research are great. Many of the present generation of models are highly stylized. It is important to discover what current findings are robust and what new findings emerge as the models are expanded to consider more factors and interactions.

For example, most of the current models consider a single drug or an undefined amalgam of all drugs. However, drugs interact with one another in many ways. At the individual level, drugs interact in users' bodies so that drugs taken in combination can lead to overdose even when larger doses of each drug singly would not. At the level of a drug use career, use of one drug can affect use of others, both in the narrow economic sense of being consumption substitutes or complements, and in the broader social sense, e.g., when use of one substance brings an individual into contact with users and sellers of other drugs. Interactions also occur at the market level: for example, the presence of established distribution networks for one drug (e.g., Colombian cocaine) can facilitate the spread of another drug (e.g., Colombian heroin), and control resources devoted to one drug may not be available for another.

Dependent users often have multiple medical conditions. Many are "dually diagnosed" with mental health and substance abuse disorders. Many are infected with HIV or Hepatitis C. Complicated interactions can occur in treatment regimens (how successful will dependent users be in complying with complicated HIV control regimens?; see Turner et al. [105]) and treatment financing (cost containment pressures may encourage restrictions on drug treatment, but resumption of drug use can increase other health care costs in the long run; see Sturm et al. [106] and Sturm and Pacula [107]). In some ways it makes more sense to think about the cost effectiveness of drug treatment relative to the cost effectiveness of other medical interventions than it does to compare drug treatment to criminal justice or prevention interventions.

Drug policy intersects not only with health and crime, but also social policy more generally [108]. For example, the issues of the dually diagnosed are particularly problematic for those who are also homeless [109]. Models that disaggregate types of users (e.g., homeless vs. other) and evaluate interventions tailored to one subpopulation or another would refine current understanding of broad strategic themes.

Perhaps the greatest need, though, is for more fundamental understanding of how drug epidemics evolve. This is perhaps best gained by modeling more epidemics, both at lower levels of geographic aggregation (e.g., in individual cities within the US) and in other countries. Comparative studies across drugs, cultures, political structures, and market conditions would help clarify what aspects of epidemic dynamics are fundamental and which are idiosyncratic to a particular context. A defining characteristic of nonlinear systems is that the magnitude of the response to a given intervention is nonlinear. Sometimes the response is less than proportionate; sometimes it is much more. Historically, drug control interventions have often produced less than hoped for effects. It may be that all these interventions are inherently ineffective or have been poorly conceived or executed. An alternative explanation, however, is that they simply have not been "timed" or "tuned" appropriately because the nonlinear character of the underlying epidemics has not been fully appreciated. In this alternate, more optimistic view, advances in understanding of drug epidemics will not only help us to choose the best among a range of interventions which may all have mediocre performance, but also to enhance the effectiveness of all interventions.

References

[1] Substance Abuse and Mental Health Services Administration, Office of Applied Studies (SAMHSA) (2002). *Emergency department trends from the Drug Abuse Warning Network, preliminary estimates January – June 2001 with revised estimates 1994 – 2000,* U.S. Department of Health and Human Services, Washington, DC.

[2] Substance Abuse and Mental Health Services Administration, Office of Applied Studies (SAMHSA) (2002). *Mortality Data from the Drug Abuse Warning Network, 2000,* DAWN Series D-19, DHHS Publication No. (SMA) 02-3633, Rockville, MD.

[3] Epstein, J.F. and J.C. Gfroerer (1998). Changes affecting NHSDA estimates of treatment need for 1994-1996. In Substance Abuse and Mental Health Services Administration, *Analyses of Substance Abuse and Treatment Need Issues,* Analytical Series Document A-7, Rockville, MD

[4] Woodward, A., et al. (1997). The drug abuse treatment gap: Recent estimates. *Health Care Financing Review,* 18, 5-17.

[5] Centers for Disease Control and Prevention (2001). *HIV/AIDS Surveillance Report.* CDC, Atlanta, GA.

[6] Rice, D.P., S. Kelman, L.S. Miller, and S. Dunmeyer (1990). *The Economic Costs of Alcohol and Drug Abuse and Mental Illness: 1985.* Institute for Health and Aging, University of California, San Francisco, CA.

[7] Bartlett, J.C., L.S. Miller, P. Rice, and W.B. Max (1994). Medical care expenditures attributable to cigarette smoking – United States 1994. *Morbidity and Mortality Weekly Report,* 43, 469-472.

[8] Harwood, H., D. Fountain, and G. Livermore (1998). *The Economic Costs of Alcohol and Drug Abuse in the United States, 1992.* US Department of Health and Human Services, Washington, DC.

[9] MacCoun, R.J. and P. Reuter (2001). *Drug War Heresies: Learning from Other Vices, Times, and Places.* Cambridge University Press, New York.

[10] Cohen, J., W. Gorr, and P. Singh (forthcoming). Estimating intervention effects in varying risk settings: Do police raids reduce illegal drug dealing at nuisance bars? *Criminology.*

[11] Kaplan, E.H. and M. Johri (2000). Treatment on demand: An operational model. *Health Care Management Science,* 3, 171-183.

[12] Schlenger, W.E. (1973). A systems approach to drug user services. *Behavioral Science,* 18, 137-147.

[13] Levin, G. E., B. Roberts, and G.B. Hirsch (1975). *The Persistent Poppy: A Computer-Aided Search for Heroin Policy.* Ballinger Publishing Company, Cambridge, MA.

[14] Gardiner, L.K. and R.C. Shreckengost (1987). A system dynamics model for estimating heroin imports into the United States. *System Dynamics Review,* 3, 8-27.

[15] Crawford, G.B., P. Reuter, K. Isaacson, and P. Murphy (1988). *Simulation of Adaptive Response: A Model of Drug Interdiction.* RAND Corporation, Santa Monica, CA.

[16] Reuter, P., R. MacCoun, and P. Murphy (1990). *Money from Crime: A Study of the Economics of Drug Dealing in Washington, DC.* RAND Corporation, Santa Monica, CA.

[17] Kahan, J.P., J. Setear, M.M. Bitzinger, S.B. Coleman, and J. Feinleib (1992). *Developing Games of Local Drug Policy.* RAND Corporation, Santa Monica, CA.

[18] Kahan, J.P., C.P. Rydell, and J. Setear (1995). A game of urban drug policy, peace and conflict. *Journal of Peace Psychology,* 1, 275-290.

[19] Kennedy, M., P. Reuter, and K.J. Riley (1993). A simple economic model of cocaine production. *Mathematical and Computer Modeling,* 17, 19-36.

[20] Childress, M. (1994). *A Systems Description of the Heroin Trade.* RAND Corporation, Santa Monica, CA.

[21] Childress, M. (1994). *A Systems Description of the Marijuana Trade.* RAND Corporation, Santa Monica, CA.

[22] Dombey-Moore, B., S. Resetar, and M. Childress (1994). *A Systems Description of the Cocaine Trade.* RAND Corporation, Santa Monica, CA.

[23] Everingham, S.S. and C.P. Rydell (1994). *Modeling the Demand for Cocaine.* RAND Corporation, Santa Monica, CA.

[24] Rydell, C.P. and S.S. Everingham (1994). *Controlling Cocaine. Supply Versus Demand Programs.* RAND Corporation, Santa Monica, CA.

[25] Hser, Y.I., M.D. Anglin, T.D. Wickens, M. Brecht, and J. Homer (1992). *Techniques for the Estimation of Illicit Drug-Use Prevalence: An Overview of Relevant Issues.* National Institute of Justice, Washington, DC.

[26] Homer, J.B. (1993). A system dynamics model for cocaine prevalence estimation and trend projection. *Journal of Drug Issues,* 23, 251-279.

[27] Homer, J.B. (1993). Projecting the impact of law enforcement on cocaine prevalence: A system dynamics approach. *Journal of Drug Issues,* 23, 281-295.

[28] Caulkins, J. and R. Padman (1993). Quantity discounts and quality premia for illicit drugs. *Journal of the American Statistical Association,* 88, 748-757.

[29] Caulkins, J. and R. Padman (1993). Interdiction's impact on the structure and behavior of the export-import sector for illicit drugs. *Zeitschrift fur Operations Research,* 37, 207-224.

[30] Caulkins, J.P. (1993). Zero-tolerance policies: Do they inhibit or stimulate illicit drug consumption? *Management Science,* 39, 458-476.

[31] Caulkins, J.P. (1993). Local drug markets' response to focused police enforcement. *Operations Research,* 41, 848-863.

[32] Gorr, W.L. and A. Olligschlaeger (1994). Weighted spatial adaptive filtering: Monte Carlo studies and application to illicit drug market modeling. *Geographical Analysis,* 26, 67-87.

[33] Blumstein, A. and D. Cork (1996). Linking gun availability to youth gun violence. *Law and Contemporary Problems,* 95, 5-24.

[34] Dawid, H. and G. Feichtinger (1996). Optimal allocation of drug control efforts: A differential game analysis. *Journal of Optimization Theory and Applications,* 91, 379-297.

[35] Gragnani, A.S., S. Rinaldi, and G. Feichtinger (1997). Dynamics of drug consumption: A theoretical model. *Socio-Economic Planning Sciences,* 31, 127-137.

[36] Kort, P.M., G. Feichtinger, R.F. Hartl, and J.L. Haunschmied (1998). Optimal enforcement policies (crackdowns) on an illicit drug market. *Optimal Control Applications & Methods,* 19, 169-184.

[37] Dworak, M. (1999). *A Dynamic Model of Drug Enforcement and Property Crime.* Masters Thesis at the Technical University of Vienna, Vienna, Austria.

[38] Feichtinger, G., W. Grienauer, and G. Tragler (2002). Optimal dynamic law enforcement. *European Journal of Operations Research,* 141, 58-69.

[39] Fent, T., G. Feichtinger, and G. Tragler (2002). A dynamic game of offending and law enforcement. *International Game Theory Review,* 4, 71-89.

[40] Baveja, A., R. Batta, J.P. Caulkins, and M.H. Karwan (1993). Modeling the response of illicit drug markets to local enforcement. *Socio-Economic Planning Sciences,* 27, 73-89.

[41] Baveja, A., J.P. Caulkins, W. Liu, R. Batta, and M.H. Karwan (1997). When haste makes sense: Cracking down on street markets for illicit drugs. *Socio-Economic Planning Sciences,* 31, 293-306.

[42] Rossi, C. (1999). Estimating the prevalence of injection drug users on the basis of Markov models of the HIV/AIDS epidemic: Applications to Italian data. *Health Care Management Science,* 2, 173-179.

[43] Rossi, C. (2001). A Mover-Stayer type model for epidemics of problematic drug use. *Bulletin on Narcotics,* 53, 39-64.

[44] EMCDDA (2001). *Modeling Drug Use: Methods to Quantify and Understand Hidden Processes.* Office for Official Publications of the European Communities, Luxembourg.

[45] Weatherburn, D. and B. Lind (1997). On the epidemiology of offender populations. *Australian Journal of Psychology,* 49, 169-175.

[46] Weatherburn, D., C. Jones, K. Freeman, and T. Makkai (2001). *The Australian Heroin Drought and Its Implications for Drug Policy.*

New South Wales Bureau of Crime Statistics and Research Publication B59, Sydney, Australia.

[47] Law, M.G., et al. (In Submission). Modeling hepatitis C virus incidence, prevalence, and long-term sequelae in Australia.

[48] Kaya, C.Y., Y. Tugai, and J.A. Filar (2001). Heroin use in Australia: Population trends. Working Paper, University of South Australia.

[49] Kaplan, E.H. and M.L. Brandeau, Eds. (1994). *Modeling the AIDS Epidemic: Planning, Policy and Prediction.* Raven Press, New York.

[50] Kaplan, E.H. and R. Brookmeyer, Eds. (2002). *Quantitative Evaluation of HIV Prevention Programs.* Yale University Press, New Haven, CT.

[51] Knoll, C. and D. Zuba (2002). *Modeling the US cocaine epidemic: Dynamic trajectories of initiation and demand.* Masters Thesis at the Technical University of Vienna, Vienna, Austria.

[52] Caulkins, J.P., C.P. Rydell, S. Everingham, J. Chiesa, and S. Bushway (1999). *An Ounce of Prevention, a Pound of Uncertainty: The Cost-Effectiveness of School-Based Drug Prevention Programs.* RAND Corporation, Santa Monica, CA.

[53] Caulkins, J.P. (1997). How prevalent are 'very light' drug users? *Federation of American Scientists' Drug Policy Analysis Bulletin,* 3, 3-5.

[54] Caulkins, J.P., S. Paddock, R. Pacula, and J. Chiesa (2002). *School-Based Drug Prevention: What Kind of Drug Use Does It Prevent?* RAND Corporation, Santa Monica, CA.

[55] Hser, Y.I., M.D. Anglin, and K. Powers (1993). A 24-year follow-up of California narcotics addicts. *Archives of General Psychiatry,* 50, 577-584.

[56] Office of National Drug Control Policy (1995). *The National Drug Control Strategy.* The White House, Washington, DC.

[57] Office of National Drug Control Policy (1999). *The National Drug Control Strategy.* The White House, Washington, DC.

[58] Sherman, L.W., et al. (1997). *Preventing Crime: What Works, What Doesn't, What's Promising.* National Institute of Justice, Washington, DC.

[59] Ellickson, P.L. and R.M. Bell (1990). Drug prevention in junior high: A multi-site longitudinal test. *Science,* 247,1299-1305.

[60] Botvin, G.J., E. Baker, L. Dusenbury, E.M. Botvin, and T. Diaz (1995). Long-term follow-up results of a randomized drug abuse prevention trial in a white middle-class population. *Journal of the American MedicalAssociation,* 273, 1106-1112.

[61] Pentz, M.A. (1998). Cost, benefits, and cost-effectiveness of comprehensive drug abuse prevention. In *Cost-Benefit/Cost-Effectiveness Research ofDrug Abuse Prevention: Implications for Programming and Policy.* NIDA Research Monograph #176, US Department of Health and Human Services, Washington, DC.

[62] Tobler, N.S. (1997). Meta-analysis of adolescent drug prevention programs: Results of the 1993 meta-analysis. In W.J. Bukowski, Ed., *NIDA Research Monograph 170,* US Department of Health and Human Services, Washington, DC.

[63] Prendergast, M., D. Podus, and K. McCormack (1998). Bibliography of literature reviews on drug abuse treatment effectiveness. *Journal of Substance Abuse Treatment,* 15, 267-270.

[64] Gerstein, D.R., et al. (1994). *Evaluation Recovery Services: The California Drug and Alcohol Treatment Assessment.* National Opinion Research Center and Fairfax: Lewin-VHI, Chicago.

[65] Institute of Medicine (1996). *Pathways of Addiction: Opportunities in Drug Abuse Research.* National Academy Press, Washington, DC.

[66] Manski, C.F., J.V. Pepper, and C.V. Petrie, Eds. (1999). *Assessment of Two Cost-Effectiveness Studies on Cocaine Control Policy.* National Academy Press, Washington, DC.

[67] Pollack, H. (2002). Methadone maintenance as HIV prevention: Cost-effectiveness analysis. In E.H. Kaplan and R. Brookmeyer, Eds., *Quantitative Evaluation of HIV Prevention Programs.* Yale University Press, New Haven, CT, 118-142.

[68] MacCoun, R.J., P. Reuter, and T. Schelling (1996). Assessing alternative drug control regimes. *Journal of Policy Analysis and Management,* 15, 330-352.

[69] Manski, C.F., J.V. Pepper, and C.V. Petrie, Eds. (2001). *Informing America's Policy on Illegal Drugs: What We Don't Know Keeps Hurting Us.* National Academy Press, Washington, DC.

[70] Behrens, D.A., J.P. Caulkins, G. Tragler, J. Haunschmied, and G. Feichtinger (2000). Optimal control of drug epidemics: Prevent and treat– but not at the same time. *Management Science,* 46, 333-347.

[71] Behrens, D. et al. (2002). Why present-oriented societies undergo cycles of drug epidemics. *Journal of Economic Dynamics and Control,* 26, 919-936.

[72] Reuter, P. (1988). Quantity illusions and paradoxes of drug interdiction: Federal intervention into vice policy. *Law and ContemporaryProblems,* 51,233-252.

[73] Kleiman, M.A.R. (1992). *Against Excess: Drug Policy for Results.* Basic Books, New York.

[74] Chaloupka, J. and R.L. Pacula (2000). Economics and anti-health behavior: The economic analysis of substance use and abuse. In W. Bickel and R. Vuchinich, Eds., *Reframing Health Behavior Change with Behavioral Economics.* Lawrence Earlbaum Associates, Hillsdale, NJ.

[75] Crane, B.D., A.R. Rivolo, and G.C. Comfort (1997). *An Empirical Examination of Counter-Drug Interdiction Program Effectiveness.* Institute for Defense Analysis, Alexandria, VA.

[76] Caulkins, J.P. (2001). The relationship between prices and emergency department mentions for cocaine and heroin. *American Journal of Public Health,* 91,1446-1448.

[77] Reuter, P. (1983). *Disorganized Crime: The Economics of the Visible Hand.* Massachusetts Institute of Technology Press, Cambridge, MA.

[78] Caulkins, J. (1997). Modeling the domestic distribution network for illicit drugs. *Management Science,* 43, 1364-1371.

[79] Reuter, P. (1985). *Eternal Hope: America's International Narcotics Efforts.* RAND Corporation, Santa Monica, CA.

[80] Caulkins, J.P. (1994). *Developing Price Seriesfor Cocaine.* RAND Corporation, Santa Monica, CA.

[81] DuPont, R.L. and M.H. Greene (1973). The dynamics of a heroin addiction epidemic. *Science,* 181, 716-722.

[82] Office of National Drug Control Policy (1992). *The National Drug Control Strategy.* The White House, Washington, D.C.

[83] Substance Abuse and Mental Health Services Administration, Office of Applied Studies (1998). US Department of Health and Human Services, Washington, DC.

[84] Maher, L. and D. Dixon (1999). Policing and public health. *British Journal of Criminology,* 39, 488-511.

[85] Maher, L. and D. Dixon (2001). The cost of crackdowns: Policing Cabramatta's heroin market. *Current Issues in Criminal Justice,* 13, 5-22.

[86] Weatherburn, D., et al. (2002) Supply control and harm reduction: Lessons from the Australian heroin 'drought'. *Addiction,* 98, 83-91.

[87] Reuter, P. and M.A.R. Kleiman (1986). Risks and prices: An economic analysis of drug enforcement. In M. Tonry and N. Morris, Eds., *Crime and Justice: An Annual Review of Research,* Vol. 7, University of Chicago Press, Chicago, IL.

[88] Caulkins, J.P. and P. Reuter (1998). What price data tell us about drug markets. *Journal of Drug Issues,* 28, 593-612.

[89] Boyum, D. (1992) *Reflections on economic theory and drug enforcement.* Doctoral Dissertation, Harvard University, Cambridge, MA.

[90] Caulkins, J.P. (1990). *The distribution and consumption of illicit drugs: Some mathematical models and their policy implications.* Doctoral Dissertation, Massachusetts Institute of Technology, Cambridge, MA.

[91] DeSimone, J. (In submission). The relationship between illegal drug prices at different market levels.

[92] Caulkins, J.P., C.P. Rydell, W.L. Schwabe, and J. Chiesa (1997). *Mandatory Minimum Drug Sentences: Throwing Away the Key or the Taxpayers' Money?* RAND Corporation, Santa Monica, CA.

[93] Caulkins, J.P. and P. Reuter (1997). Setting goals for drug policy: Harm reduction or use reduction. *Addiction,* 92, 1143-1150.

[94] Office of National Drug Control Policy (2001). *The National Drug Control Strategy.* The White House, Washington, DC.

[95] Greenwood, P.W., et al. (1994). *Three Strikes and You're Out: Estimated Benefits and Costs of California's New Mandatory Sentencing Law.* RAND Corporation, Santa Monica, CA.

[96] Tragler, G., J.P. Caulkins, and G. Feichtinger (2001). Optimal dynamic allocation of treatment and enforcement in illicit drug control. *Operations Research,* 49, 352-362.

[97] Caulkins, J.P., M. Dworak, G. Feichtinger, and G. Tragler (2000). Drug enforcement and property crime: A dynamic model. *Journal of Economics,* 71, 227-253.

[98] Mautner, K. (2002). *A dynamic one-state two-control optimization model of the current Australian heroin problem.* Masters Thesis at the Technical University of Vienna, Vienna, Austria.

[99] Behrens, D.A., J.P. Caulkins, G. Tragler, J. Haunschmied, and G. Feichtinger (1999). A dynamic model of drug initiation: Implications for treatment and drug control. *Mathematical Biosciences,* 159, 1-20.

[100] Almeder, C., J.P. Caulkins, G. Feichtinger, and G. Tragler (2001). An age-specific multi-state initiation models: Insights from considering heterogeneity. *Bulletin on Narcotics,* 53, 105-118.

[101] Grosslicht, F. (2002). *Optimal dynamic allocation of prevention and treatment in a model of the Australian heroin epidemic.* Masters Thesis at the Technical University of Vienna, Vienna, Austria.

[102] Naik, A.V., A. Baveja, R. Batta and J.P. Caulkins (1996). Scheduling crackdowns on illicit drug markets. *European Journal of Operational Research,* 88, 231-250.

[103] Kleiman, M.A.R. (1993). Enforcement swamping: A positive-feedback mechanism in rates of illicit activity. *Mathematical and Computer Modeling,* 17, 65-75.

[104] Behrens, D.A., J.P. Caulkins, G. Tragler, and G. Feichtinger (1997). *Controlling the U.S. cocaine epidemic: Prevention from light vs. treatment of heavy use.* Working Paper #214, Vienna University of Technology, Vienna, Austria.

[105] Turner, B.J., et al. (2001). Effects of drug abuse and mental disorders on use and type of antiretroviral therapy in HIV-infected *persons. Journal of General Internal Medicine,* 16, 625-633,

[106] Sturm, R., W. Zhang, and M. Schoenbaum (1999). How expensive are unlimited substance abuse benefits under managed care? *Journal of Behavioral Health Services Research,* 26, 203-210.

[107] Sturm, R. and R. Liccardo Pacula (2000). State mental health parity laws: Cause or consequence of differences in use? *Health Affairs,* 18, 182-192.

[108] Boyum, D. and P. Reuter (2001). Reflections on drug policy and social policy. In P.B. Heymann and W.N. Brownsberger, Eds., *Drug Addiction and Drug Policy,* Harvard University Press, Cambridge, MA.

[109] Stecher, B.M., et al. (1995). Implementation of residential and nonresidential treatment for the dually diagnosed homeless. *Evaluation Review,* 18, 690-718.

[110] Winkler, D., J.P. Caulkins, D. Behrens, and G. Tragler. (2004). Estimating the relative efficiency of various forms of prevention at different stages of a drug epidemic. *Socio-Economic Planning Sciences,* 38, 43-56.

13 MODELING THE COSTS AND EFFECTS OF MAINTENANCE TREATMENT FOR OPIATE ADDICTION

Gregory S. Zaric

Richard Ivey School of Business
University of Western Ontario
London, Ontario, Canada N6A 3K7

SUMMARY

In this chapter we discuss recent operations research advances in modeling drug treatment programs for injection drug users, in particular maintenance treatment programs for opioid addicts. We focus on four main questions for which operations research techniques have proven beneficial: How effective are opioid maintenance programs? Do the benefits of methadone maintenance treatment justify its costs? Are alternative forms of maintenance treatment cost effective? If opioid maintenance treatment programs are expanded, how many new treatment slots are needed? We discuss a number of methodological issues and highlight directions for future research.

KEY WORDS

HIV, Methadone, Cost-effectiveness analysis, Cost-benefit analysis

13.1 INTRODUCTION

Injection drug use is a significant public health problem in the United States. Between 750,000 [1] and 1.2 million [2] individuals in the U.S. are injection drug users (IDUs). Injection drug use is a major risk factor for human immunodeficiency virus (HIV). The prevalence of HIV among IDUs exceeds 40% in some cities [3]. Approximately 20-25% of all new HIV cases and 35% of all acquired immune deficiency syndrome (AIDS) cases in the United States have injection drug use as a risk factor [4]. IDUs may serve as a "core group" [5] in the HIV epidemic and spread HIV to non-IDUs through sexual contact. In addition to HIV, IDUs are subject to a number of other comorbidities including hepatitis, tuberculosis [6], overdose and accidental death [7], and have mortality rates that are up to 60 times greater than those of other members of their age group [8]. IDUs may make greater use of emergency health services and less use of regular health services, and have annual health care expenditures that are three to four times greater than those of other members of their age group [9, 10]. Injection drug use is also associated with increased criminal activity and increased costs to the criminal justice and welfare systems [11, 12].

Methadone was developed in Germany during World War II as a substitute for morphine and has been used in the treatment of heroin addiction for more than 30 years [13]. Methadone has a slow onset and a long delay, with effects lasting up to 24 hours, and can be taken once a day to curb heroin withdrawal symptoms. Methadone maintenance treatment (MMT) is associated with reduced illicit drug use, reduced HIV risk behavior, and reduced drug and property-related criminal activity [14, 15]. Non-HIV health care expenditures are lower for IDUs in MMT than for IDUs not in MMT [16]. A meta-analysis of studies comparing IDUs in MMT versus IDUs not in MMT found a relative risk of death of 0.24 – 0.43 associated with MMT [17].

Methadone is a highly regulated substance in the U.S. [18] and is classified as a Schedule II narcotic (meaning that it has a high potential for abuse) by the U.S. Drug Enforcement Administration [19]. There are only approximately 115,000 methadone treatment slots in the U.S., or roughly enough for 10-20% of all IDUs nationwide [18]. Some states do not have any methadone programs [20]. Methadone is typically administered daily under supervised settings to prevent potential abuse of the drug. This and other regulations contribute to the high cost of MMT. Estimates of the annual cost of one methadone treatment slot range from $4,300 [21] to $5,250 [22] (in 1996 dollars). However, methadone is a generic drug that costs less than $1 per day [23]. One study found that drug costs accounted for only 5-6% of the total cost of a methadone treatment slot [21].

MMT is controversial in the United States. The former U.S. "Drug Czar" (Director of the Office of National Drug Control Policy) publicly supported increased funding and use of methadone as part of a drug abuse reduction strategy [24]. A report from the National Institutes on Drug Abuse has advocated expanded methadone capacity [25], and drug abuse prevention has recently been included among the principles for HIV prevention among IDUs [26]. However, support for MMT is not universal. In 1999, a bill was introduced in the U.S. Congress that called for limiting methadone funding and access [27]. In 1996, the mayor of New York City declared that methadone was immoral and represented the substitution of one drug (methadone) for another (heroin) [28]. He subsequently reversed his position and devoted $5 million to increased city funding of methadone programs [29].

This chapter reviews modeling work that evaluates programs to treat opioid dependence. We focus on four major questions where operations research techniques have provided insight into the value of such programs:

1. How effective are opioid maintenance programs?

2. Do the benefits of MMT justify its costs?

3. Are alternative forms of maintenance treatment cost effective?

4. If opioid maintenance treatment programs are expanded, how many new slots are needed?

Following the discussion of the questions we highlight some methodological issues and describe a number of promising areas for future research.

13.2 MODELS OF OPIOID MAINTENANCE PROGRAMS

13.2.1 How effective are opioid maintenance programs?

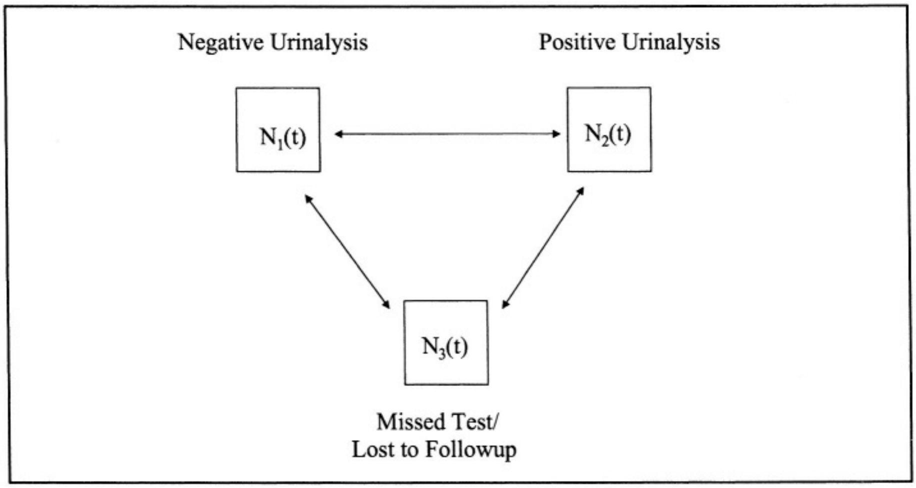
MMT programs may not lead to a complete cessation of drug use but, rather, a reduction in usage; similarly, they may not lead to complete cessation of needle sharing. Additionally, many IDUs lead unstructured lives, leading to substantial difficulties in fulfilling the follow-up and monitoring requirements of many studies. Statistical techniques that only record "success" or "failure", as well as those that do not handle large amounts of missing or censored data, may not be suited to the assessment of opioid maintenance programs. Thus the motivation to develop new techniques.

Lee [30] and Weng [31] developed models to assess the effectiveness of methadone and buprenorphine maintenance programs in the presence of missing observations. Buprenorphine is an alternative to methadone that has only recently been approved for maintenance treatment in the U.S. Both models were fit using data from a 17-week randomized clinical trial to evaluate the effectiveness of buprenorphine [32]. Patients in the trial were randomized into three groups: Group 1 received 8 mg of buprenorphine daily; Group 2 received 20 mg of methadone daily; and Group 3 received 60 mg of methadone daily. Patients in each group were asked to provide urine samples three times per week to assess their drug consumption while in treatment. Between 60% and 80% of the members of each group were lost to followup, and approximately 18% of urine samples among those not lost to followup were missed in each group.

Weng [31] developed a stochastic compartmental model with three compartments representing negative urinalysis $(N_1(t))$, positive urinalysis $(N_2(t))$, and missed test $(N_3(t))$. [31] The model was formulated as a continuous-time stochastic process, as depicted in Figure 13.1. Clinical trials data was used to estimate flow rates between states for the three study groups. Transitions between any two states were allowed, and the population was assumed to be closed. The 17-week period was partitioned into four or five segments for each group. The steady state probability of being in the negative state was found to be between .403 and .566 for Group 1 (buprenorphine); between .183 and .353 for Group 2 (20 mg methadone); and between .271 and .465 for Group 3 (60 mg methadone).

Lee [30] used a two-state discrete-time Markov chain to examine the effectiveness of methadone and buprenorphine programs. The discrete time steps corresponded to urine sample collection points. The two states, denoted by 0 and 1, represented negative and positive urinalysis results, respectively. Estimation procedures were developed to estimate the transition probabilities between the two states given a sequence of urinalysis results that may contain missing observations. Maximum likelihood estimates for P1, the probability of opiate use during the 17-week period, were calculated. In one set of calculations, it was found that P1 = 0.4734 for Group 1 (buprenorphine), P1 = 0.6288 for Group 2 (20 mg methadone), and P1 = 0.4970 for Group 3 (60 mg methadone). In a second set of calculations it was found that P1 was 0.3664, 0.6260, and 0.4854 for the three groups, respectively.

Figure 13.1 Compartmental model to investigate the effectiveness of methadone and buprenorphine treatment



The methods developed by Lee and Weng [30, 31] address some of the difficulties in assessing the effectiveness of opioid maintenance programs. However, there are still opportunities for new methods. There may be many definitions of "success" related not only to drug use but also to the frequency of engaging in risky behavior and frequency of use of drugs other than heroin. Also, IDUs may simultaneously use several drug treatment services (i.e. MMT, counseling, support groups). Future methods may seek to address multiple definitions of success as well as the incremental impact of each service.

13.2.2 Do the benefits of MMT justify its costs?

Drug treatment programs, including MMT, are often seen as primarily benefiting one group (IDUs) while being paid for by another group (taxpayers). This discordance has generated much interest in understanding whether the benefits of MMT justify its costs. Much of the analysis of this question has utilized cost-benefit analysis (CBA) or cost-effectiveness analysis (CEA). Recent debate has questioned whether CBA and CEA can be considered equivalent [33-37]. Applications of these techniques in the evaluation of opiate treatment programs clearly are not equivalent: researchers performing CBA tend to focus on the social impact of drug treatment (such as crime, judicial costs, social welfare, etc), whereas researchers performing CEA tend to focus on the health impacts of drug treatment (including mortality, comorbidities, and HIV infection).

Both techniques often make use of the quality-adjusted life year of survival (QALY) as a way of characterizing the health benefits of a program [38]. QALYs represent utilities for health states and are scaled between 0, representing death, and 1, representing perfect health.

Cost benefit analysis of methadone programs In cost-benefit analysis, all costs and benefits of a proposed program are converted into monetary units. For a health care program, this requires conversion of health outcomes into monetary units. The requirement that health outcomes be explicitly valued is often cited as a criticism of CBA in the evaluation of health care programs. Results of a CBA are typically expressed in several different formats, including the net benefits approach (net benefits = total benefits – total costs) and the benefit-to-cost ratio. It has been argued that CBA may be preferable to CEA for drug abuse interventions since many of the benefits (e.g., improved employment, reduced criminal activity) are not health related [39].

An early cost benefit analysis of methadone programs was provided by Hannan [40]. The analysis focused on four direct impacts of methadone treatment: decreased criminal justice expenditures, decreased health care costs for heroin-related conditions, decreased expenditures on heroin, and increased legal earnings among those treated. The monetary value of property theft crimes was not included in the analysis since it is a transfer of wealth with no net impact. The analysis was based on data from a New York MMT program in 1965. For a six-year time horizon, benefit-to-cost ratios of 1.47 to 4.40 were found, depending on which benefits were included in the analysis. For a projected 33-year time horizon, benefit-to-cost ratios of 1.86 to 5.09 were found. In all cases considered, the benefits were substantially greater than the costs. A limitation of this study was that health care costs were included, but the health benefits of MMT were not.

French and colleagues described a methodology for conducting benefit-cost analysis of methadone programs [39, 41]. The methodology involves converting scores from a disease severity index into QALYs, and then multiplying QALY estimates by the societal willingness to pay (WTP) for a QALY to yield the monetary value of health outcomes. Health benefits are converted into monetary outcomes using the following formula:

$$\text{health benefits} = \frac{1}{N}\sum_{i=1}^{N}(1-QA_i)\times(\$/QALD) \tag{1}$$

where N = 19 (corresponding to 19 comorbidities that are common among IDUs), QA_i is the quality-of-life adjustment for condition i, and QALD is the societal WTP for one quality-adjusted life day [41]. The value of QALD was

$173.08, derived from an estimate of the value of life [42]. The net cost is computed by adding the other monetary costs to the health costs. This CBA methodology is illustrated with sample calculations [39] and with a full CBA based on data from the Philadelphia Target Cities Project [41].

A similar methodology was used to conduct a benefit-cost analysis of two levels of intensity of addiction service, denoted by "partial continuum" (PC) and "full continuum" (FC), in Washington State [43]. Costs included those related to health care, psychiatric status, employment status, drug and alcohol use, and legal status. Health benefits were derived in part by converting changes in the Addiction Severity Index [44] for treated patients into monetary units. FC only, PC only, and FC and PC together had average net benefits of $17,833, $11,173, and $15,305, respectively (expressed in 1997 dollars) and respective benefit-cost ratios of 9.70, 23.33, and 14.87.

The conversion of health benefits to monetary units is a necessary part of any CBA, but formula (1) has some shortcomings. The formula does not address the possibility that some comorbidities are more common among IDUs than others. Dividing by N (N=19 in the example given) implicitly assumes a prevalence of 1/N (5.3% when N = 19) for all comorbidities, and that there is no correlation between different comorbidities. The method also does not quantify the impact that drug treatment has on reducing the probability of developing one of the comorbid conditions. Also, QALYs represent utilities, and it is unclear if scores from a disease severity index can be converted to utilities.

The cost effectiveness of methadone Cost-effectiveness analysis involves calculation of the incremental cost-effectiveness ratio, which is defined as the incremental costs of an intervention divided by the incremental health benefits of the intervention [45]. The incremental costs of an intervention include the cost of the intervention itself plus the costs associated with all future changes in health caused by the intervention. The incremental cost term may include non-health care costs if a societal perspective is taken. Costs and benefits are typically discounted to reflect the principle that costs in the future are preferred to costs today, and benefits today are preferred to benefits in the future.

If the health benefits of the intervention are expressed in terms of QALYs, then a CEA may be referred to as a cost-utility analysis (CUA). Conversion of the health benefits of interventions into QALYs allows interventions that yield very different benefits to be compared. By expressing results as a ratio, CEA avoids having to explicitly assign a monetary value to health outcomes.

Barnett constructed a life table to examine the costs, benefits, and cost effectiveness of MMT [46]. Age-specific mortality rates for non-IDUs (i.e., individuals who do not inject drugs) were obtained from U.S. life tables. Numerous studies have compared mortality rates for IDUs in and out of MMT versus those of the general population. For instance, a study in Sweden found that IDUs in MMT had 12 times the annual mortality rate of individuals in their age group, and IDUs not enrolled in MMT had 63 times the mortality rate of individuals in their age group [8]. Age-specific mortality rates for IDUs in and out of MMT were obtained by multiplying the age-specific rates for the general population by these relative risk rates applicable to IDUs.

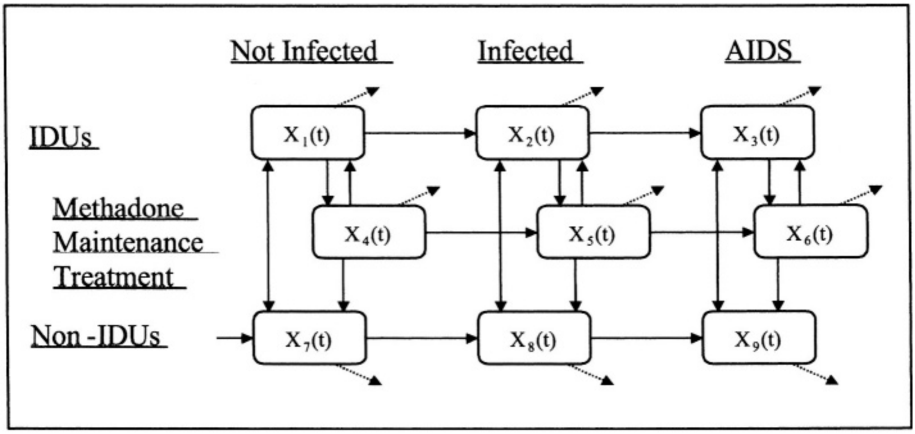
Survival until age 65 for an initial cohort of 1,000 25-year old IDUs was calculated using the estimated age-specific mortality rates. This was compared to survival of a similar cohort that was assumed to have access to MMT. It was assumed that 57.5% of patients in the MMT group received methadone and hence receive the survival advantage associated with methadone. Total discounted life years of survival attained and the total costs associated with MMT for each group were determined. The cohort that received methadone experienced 8,704 additional discounted life years of survival at an incremental cost of $51,486,000, resulting in a cost-effectiveness ratio of $5,915 per life year gained. Extensive one-way sensitivity analyses revealed cost-effectiveness ratios between $3,300 and $9,100 per life year gained.

Kahn and colleagues analyzed a number of HIV prevention programs including methadone maintenance [47]. In their analysis of MMT they constructed two scenarios representing cities with different drug and HIV epidemics. They assessed the impact over five years of a one-year expansion in MMT capacity. They found that the extra MMT capacity had a cost-effectiveness ratio of $48,000 to $60,000 per undiscounted life year gained. The analysis considered only the impact of MMT on the spread of HIV and did not consider other health care costs.

A compartmental model of methadone maintenance Zaric et al. developed a compartmental epidemic model to evaluate the cost effectiveness of expanded methadone treatment capacity on a population of IDUs and non-IDUs [48, 49]. The work was motivated by the difficulty that other evaluation techniques have had in quantifying the impact of expanded methadone on the spread of HIV. Characterizing the impact of new treatment programs on the spread of HIV is important because HIV has a significant impact on total health care costs and mortality.

A model was developed in which the population was divided into nine compartments based on behavior (IDU, IDU in methadone, or non-IDU) and disease status (not infected with HIV, HIV-infected, and AIDS). The model is illustrated in Figure 13.2. The arrows in Figure 13.2 represent transitions between behavior and risk classes; these transitions were modeled using a system of differential equations. All individuals enter the model as uninfected 18-year-old non-IDUs. They remain in that state until they die or age out of the model, or they become HIV infected, or they become IDUs. IDUs can become infected through sexual or needle-sharing contacts with infected individuals, and non-IDUs can only become infected through sexual contact. Non-IDUs were assumed to become IDUs at a fixed rate. IDUs could remain as IDUs, they can re-enter the non-IDU population, or they can enter MMT slots as space became available. IDUs in MMT can leave MMT at any time and enter either the IDU or non-IDU population.

Figure 13.2 Compartmental model to examine the impact of MMT on HIV



The model was used to dynamically calculate new infections and rates of entry into treatment. Let $X_i(t)$ be the number of individuals in compartment i at time t, i = 1,...,9, and let N be the total number of MMT slots available. One constraint ensured that the MMT slots were always filled and a second constraint ensured that new entrants to MMT were drawn from each disease state according to prevalence in the population.

A challenging aspect of the model formulation was defining a mixing model for a population with two types of risk behavior (sexual mixing and needle sharing), two levels of sexual risk (with and without condoms), differing rates of participation in the two risk activities, and like-with-like sexual preferences [50, 51]. The mixing model was defined by first specifying a term for the number of new infections. The number of new infections among members of compartment i is given by:

$$NI_i(t) = X_i(t)\sum_{j=1}^{9}\left(\gamma_{ij}(t) + \beta_{ij}^{H}(t) + \beta_{ij}^{L}(t)\right) \quad i = 1, 4, 7 \tag{2}$$

where $\gamma_{ij}(t)$, $\beta_{ij}^{H}(t)$, and $\beta_{ij}^{L}(t)$ are the sufficient contact rates between compartments i and j for injection, high-risk sexual contact (sexual contact in which a condom is not used), and less-risky sexual contact (sexual contact in which a condom is used), respectively.

For risky injections, the probability of a contact between members of compartments i and j was assumed to be proportional to the total number of injections by members of those two compartments. That is, the probability that an individual has a contact with a member of compartment j is the total number of injections by all members of compartment j divided by the total number of injections by all members of all compartments. This probability changes over time since the number of people in each compartment changes over time. Thus, the sufficient contact rate for injections, $\gamma_{ij}(t)$, was given by the number of injections per person in compartment i multiplied by the probability of a contact between compartments i and j multiplied by the probability of disease transmission for a contact between compartments i and j.

The formulas for the sufficient contact rates for sexual mixing (β_{ij}^{H} and β_{ij}^{L}) were similar to those for shared injections but modified somewhat to account for the presence of two types of risky contacts, different rates of condom use between IDUs and non-IDUs, and preferential mixing. Let G be the proportion of sexual contacts that IDUs have with other IDUs. Let P_i^S be the average annual number of new sexual partners, and let P_i^R be the average number of new sexual partners of risk R, R = L, H, among members of compartment i. Then $P_i^L = d_i \times P_i^S$, and $P_i^H = (1-d_i) \times P_i^S$, where d_i is the probability that an individual in compartment i uses a condom. Let ce be the risk reduction achieved by using a condom [52], and let τ_{ij}^{L} and τ_{ij}^{H} be the probabilities of HIV transmission through sexual contacts of type L and H, respectively, where $\tau_{ij}^{L} = ce \times \tau_{ij}^{H}$. Let $M_{ij}^{R}(t) = [m_{ij}^{R}(t)]$ be the mixing matrix for sexual contacts of type R, R = L, H. Then $m_{ij}^{R}(t)$ is given as follows:

$$m_{ij}^{R}(t)=\begin{cases} G\dfrac{X_j(t)P_j^R}{\sum_{k=1}^{6}X_k(t)P_k^R} & i=1,\dots,6, j=1,\dots,6 \\ (1-G)\dfrac{X_j(t)P_j^R}{\sum_{k=7}^{9}X_k(t)P_k^R} & i=1,\dots,6, j=7,\dots,9 \\ \dfrac{(1-G)\sum_{k=1}^{6}X_k(t)P_k^R}{\sum_{k=7}^{9}X_k(t)P_k^R}\times\dfrac{X_j(t)P_j^R}{\sum_{k=1}^{6}X_k(t)P_k^R} & i=7,\dots,9, j=1,\dots,6 \\ \left[1-\dfrac{(1-G)\sum_{k=1}^{6}X_k(t)P_k^R}{\sum_{k=7}^{9}X_k(t)P_k^R}\right]\times\dfrac{X_j(t)P_j^R}{\sum_{k=7}^{9}X_k(t)P_k^R} & i=7,\dots,9, j=7,\dots,9 \end{cases} \tag{3}$$

Each expression in (3) has two terms. The first is the probability that an individual from compartment i has a contact with an individual from the group of compartments indexed by j. The second is the conditional probability that a contact is with someone from compartment j given that there is a contact with someone from the specified group of compartments. We explain these values for the first and third expressions in $m_{ij}^R(t)$ below.

For i = 1,…,6, j = 1,…,6, the expression for $m_{ij}^R(t)$ is the probability that an IDU has a contact with another IDU multiplied by the probability that the contact is with someone from compartment j given that it is with someone from compartments 1,...,6. For i = 7,...,9, j = 1,...,6, the expression for $m_{ij}^R(t)$ contains two terms. The first is the probability that a non-IDU has a contact with an IDU. Since IDUs have $(1-G)\sum_{i=1}^{6}X_i(t)P_i^R$ total contacts with non-IDUs, non-IDUs must have the same total number of contacts with IDUs in order for the total number of sexual contacts to balance. Thus, the first term is the proportion of total contacts by non-IDUs that are with IDUs, and is interpreted as the probability that a non-IDU has a contact with an IDU. The second term is the probability that the contact is with a member of compartment j given that the contact is with an IDU. Following from the above discussion, the sufficient contact rates for sexual mixing at time t are thus given by

$$\beta_{ij}^R(t)=P_i^R m_{ij}^R(t)\tau_{ij}^R \quad \mathrm{R = H, L, i = 1,\dots,9, j = 1,\dots,9} \tag{4}$$

The number of new infections among members of compartment i, given by (2), is found by multiplying the number of individuals in compartment i by their rate of sufficient contacts with member of compartment j for each type of contact.

Scenarios representing regions with HIV prevalence among IDUs of 5%, 10%, 20%, and 40% were simulated. Total costs, QALYs, and new infections, as well as the incremental cost effectiveness ratio (ICER), were calculated for a 10 year time horizon. An expansion of MMT by 10% of current capacity (i.e., increasing the proportion of IDUs enrolled in MMT from 15% to 16.5%) was analyzed in each scenario. In the 5% scenario, the expansion of methadone capacity would result in cost-effectiveness ratio of $10,900 per QALY gained. In the 40% scenario, the expansion of methadone capacity would result in a cost-effectiveness ratio of $8,200 per QALY gained. These cost-effectiveness ratios compare favorably to a number of HIV prevention and treatment programs [48, 53]. In the 5% scenario, approximately 36% of HIV infections averted and 71% of QALYs gained accrued to non-IDUs. In the 40% scenario, approximately 28% of infections averted and 58% of QALYs gained accrued to non-IDUs. Thus, substantial health benefits of MMT programs accrue to non-IDUs.

A number of sensitivity analyses were performed to consider the cost effectiveness of increased methadone capacity if the newly created slots were less effective and/or more costly than the existing slots. New MMT slots may be less effective than existing slots if new recruits are less motivated to change their behavior than those already in MMT. New MMT slots may be more expensive than existing slots if there are additional costs associated with outreach to fill the new slots. If all new slots are half as effective and twice as costly as existing slots, expanded MMT capacity had a cost-effectiveness ratio of $36,100 in the 5% scenario and $38,300 in the 40% scenario.

This study found MMT to be cost effective based on commonly accepted standards, under a wide range of assumptions. An important conclusion was that MMT could be cost effective even if it did not lead to a complete cessation of risky injections. Some factors, such as crime and changes in employment among IDUs, were omitted from the analysis. Inclusion of these factors would likely lead to more favorable conclusions regarding the cost effectiveness of methadone.

13.2.3 Are alternative forms of maintenance treatment cost effective?

Buprenorphine is a potential alternative to methadone and may be useful for expanding treatment capacity. Buprenorphine is subject to a different regulatory environment than methadone. Methadone is listed as a Schedule II drug by the U.S. Drug Enforcement Administration, while buprenorphine is a Schedule V drug (having low potential for abuse and accepted medical uses) [19]. Compared to methadone, buprenorphine is safer in overdose, has lower abuse potential, and fewer withdrawal symptoms when discontinued [54]. To curtail abuse through injection , buprenorphine can be taken orally and mixed with naltrexone or naloxone, both of which have unpleasant effects if injected but are relatively harmless when taken orally [54, 55]. The ability to mix buprenorphine with naloxone makes buprenorphine potentially attractive in the development of take-home or prescription maintenance formulations.

Barnett et al. [56] modified the model of Zaric et al [48, 49] to evaluate the cost effectiveness of buprenorphine maintenance treatment. The model of MMT cost effectiveness was modified to account for observed differences in the effectiveness of methadone versus buprenorphine as well as likely cost differences between the two products.

A meta-analysis of trials comparing buprenorphine to methadone found that patients maintained on buprenorphine had 8.3% more positive urinalyses and a 26% higher dropout rate than patients in methadone [57]. One analysis of the economic impact of a potential take-home formulation of buprenorphine with naloxone concluded that the buprenorphine formula would cost between 81% and 113% as much as methadone when patient travel time was not included, and 44% to 76% as much as methadone when patient travel time was included [58]. Barnett et al. estimated that a take-home formulation of buprenorphine and naloxone would cost between $5 and $30 per day, corresponding to annual costs of $5733 to $14,858 [56].

Buprenorphine may be preferred to methadone by some IDUs. Thus, some newly created buprenorphine treatment slots may be filled by patients formerly in MMT. Let f_M and f_B be the efficacy of methadone and buprenorphine slots in reducing risky behavior and let f_{AVE} be the average efficacy of all treatment slots. Let N_M be the initial number of methadone slots, and let N_B be the number of newly created buprenorphine slots. Let p be the proportion of new slots filled by individuals who switch from MMT. Adding N_B buprenorphine slots results in a net expansion of capacity of $(1-p)N_B$ slots. The average efficacy of all slots is defined as:

$$f_{AVE} = \frac{f_B N_B + f_M \left(N_M - pN_B\right)}{N_B + N_M - pN_B} \qquad (5)$$

The reduction in sharing, change in mortality rates, and dropout rates for the treatment compartments, consisting of both MMT and buprenorphine patients, reflected weighted averages given by (5).

For the case where there is no switching (p = 0), additional buprenorphine slots equal to 10% of current MMT capacity would have an incremental cost-effectiveness ratio of $14,000 per QALY gained if buprenorphine cost $5 per dose, ranging up to $44,200 per QALY gained if buprenorphine cost $30 per dose. If half of all new slots are occupied by individuals switching from MMT (p = 1/2), then buprenorphine costs $17,700 per QALY gained at $5 per dose, and $84,700 per QALY gained at $30 per dose. Extensive sensitivity analysis was done on quality of life and the benefits of treatment on quality of life.

In all cases buprenorphine was found to be less cost effective than methadone. However, buprenorphine is still cost effective compared to a number of other medical interventions. Additionally, buprenorphine has fewer regulatory impediments and may represent an option for expansion of drug treatment programs where methadone is not an option.

A number of issues have been raised regarding the study by Barnett et al. [59-61]. Reductions in crime are often cited as the major benefit of drug treatment programs, but crime was not considered in the model. The analysis was done from the perspective of a health payer who may not be concerned with reductions in crime. However, government policy makers may be concerned about such costs [60]. The use of QALYs as an outcome measure has also been questioned for an intervention that is not seen exclusively as a health care program and that has substantial non-health benefits [59, 61].

Wall and Pollack [62] adapted the model of Zaric et al. [48, 49] to evaluate several drug treatment expansion strategies involving buprenorphine. They assumed that the effectiveness of existing and new treatment slots was a function of the size of the daily dose of methadone or buprenorphine, consistent with evidence that dose size and treatment efficacy may be related [63]. They considered a number of strategies involving increasing the methadone dosage of existing slots, converting existing slots to buprenorphine, and expansion with methadone and buprenorphine at varying dosage levels.

Increasing methadone dosage for existing slots was found to be cost saving, while switching all existing slots to buprenorphine was a dominated strategy (i.e., more expensive and less effective than another strategy). Expanding capacity with methadone was found to be very cost effective. The methadone-only strategies all had cost-effectiveness ratios of less than $4,000 per QALY gained. Expanding with a mix of methadone and buprenorphine was found to have a cost-effectiveness ratio of less than $30,000 per QALY gained, and expanding capacity with buprenorphine only was a dominated strategy.

Additional studies of buprenorphine are warranted given its recent approval for use in the U.S [64]. A number of alternatives to methadone and buprenorphine treatment exist and have not been subject to the kind of modeling described in this chapter. For instance, L-alpha-acetylmethadol (LAAM) may be used to control opiate addiction. Detoxification programs combined with intensive social services may also be valuable. Rapid detoxification (so-called "rapid detox") may represent an alternative to treatment MMT [65-67]. Different treatment modalities have emerged, including prescription or take-home formulations of methadone and buprenorphine. Prescription buprenorphine is available in France [68] and prescription methadone is available in the United Kingdom [69]. Prescription heroin has also been proposed by some [70]. All of these options merit consideration in future investigations.

13.2.4 If opioid maintenance treatment programs are expanded, how many new slots are needed?

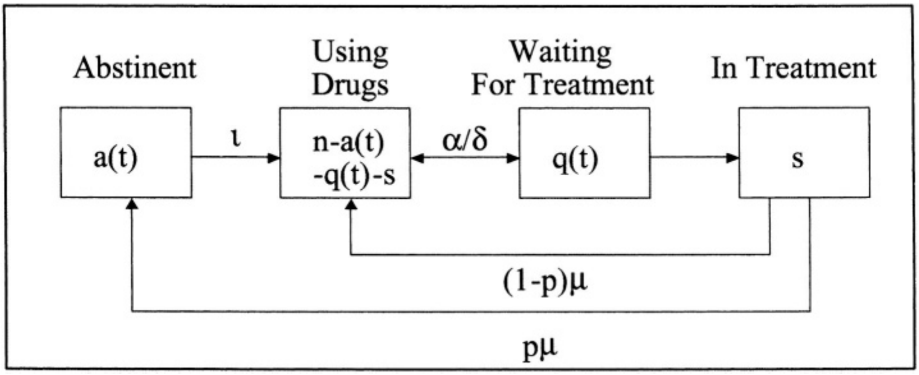
Lengthy waiting lists for drug treatment have been documented in many places [71]. Some jurisdictions require new entrants to MMT to be HIV infected or to have tried another drug treatment program (e.g., detoxification) unsuccessfully. An important question is how many treatment slots would be needed to eliminate or reduce treatment queues. Another important issue is the impact of extra capacity on waiting list performance measures.

Ideas from queuing theory have been used to predict capacity requirements for treatment programs such that individuals can receive "treatment on demand" [72]. "Treatment on demand" was defined as having no wait for treatment once treatment was requested. In the model, N customers arrive seeking treatment, a proportion R1 remain on the list and enter treatment, and 1-R1 enter the waiting list but do not wait for treatment.

Kaplan and Johri investigated the impact on drug treatment waiting lists of providing additional drug treatment capacity [73]. The model was not intended to represent any particular drug treatment program but rather to represent a general drug treatment model. (Experience with treatment on demand in San Francisco has been described elsewhere [74].) Kaplan and Johri examined operational outcomes including queue lengths, waiting times to enter treatment, and service levels. The service level was defined as the proportion of those requesting treatment who remained on the waiting list long enough to be admitted into treatment. Numerous factors contribute to the inability of drug users to remain on waiting lists until a treatment slot becomes available, such as a loss of interest in treatment, arrest, or inability of the treatment facility to place the person at the front of the queue.

Kaplan and Johri [73] developed a model in which drug users can be in one of four states at any time: abstinent (not in treatment and not using drugs); not in treatment and using drugs but not waiting for treatment; waiting for treatment and using drugs; and in treatment and not using drugs. Let a(t) be the number of drug users who are abstinent at time t, and let q(t) be the number who are waiting for treatment at time t. They modeled a closed population (i.e., no new entrants and no departures) of size n, with a constant number of treatment slots, s. The model is depicted in Figure 13.3.

Figure 13.3 Treatment-on-demand model



Let δ be the tolerance to wait for treatment; this is the rate at which individuals waiting for treatment leave the waiting list. Let μ be the rate at which treatment is completed. Let ι be rate at which abstinent users resume drug use. Let α be the rate at which those not in treatment request treatment. Let p be the probability of success per treatment episode. The population was modeled using the following system of differential equations:

$$\frac{da(t)}{dt} = s\mu p - \iota a(t) \tag{6}$$

$$\frac{dq(t)}{dt} = \alpha\left(n - s - q(t) - a(t)\right) - \delta q(t) - \mu s \tag{7}$$

The first equation says that the rate of change of the size of the abstinent population is equal to the number of successful completions of drug treatment minus the number of abstinent users who resume active drug use. The second equation says that the rate of change of the queue length is equal to the number of users who request treatment minus the number who drop out of the queue minus the number on the queue who enter treatment.

Closed-form solutions for a(t) and q(t) can be obtained and used to estimate the service level. The number of new slots needed to eliminate queues in the long run is given by:

$$s^{*} = n\frac{\alpha\iota}{\alpha\iota + \alpha\mu p + \iota\mu} \qquad (8)$$

The value of s is the number of users currently in the queue, in general. Thus, the naïve approach of adding as many treatment slots as there are patients currently in the queue would not be the correct way to eliminate the treatment queue, in general.

The model was illustrated with data from San Francisco. There were n = 45,000 drug users in San Francisco, s = 6,300 treatment slots, q(t) = 1,400 currently waiting to enter treatment, and a(t) = 17,600 abstinent drug users. Four values for the tolerance to wait for treatment (δ) among drug users were considered.

The San Francisco data showed that small increases in waiting times could lead to large reductions in service levels. Although s^{*} is independent of δ in (8), numerical analysis using the San Francisco data showed that the number of slots needed to eliminate queues for treatment in the long run was highly dependent on the tolerance to wait for treatment. This is due to the relationship between α and δ in equilibrium and the methods used to estimate α. If the tolerance to wait is one year, then 6,710 treatment slots (less than 6,300 + 1,400) would be sufficient to eliminate treatment queues. If the tolerance to wait is only one day, then 11,500 slots would be required to eliminate treatment queues. However, it could take 22 years or more to eliminate the queues using these long-run estimates. For very short tolerance to wait, the number of slots needed to immediately eliminate queues could be as high as 18,000.

13.3 METHODOLOGICAL ISSUES AND FUTURE WORK

Quality-of-life estimates are available for injection drug use and for HIV, but currently there is no specific estimate for "IDUs with HIV". Thus, the separate quality-of-life estimates must be combined in some way. It is not clear if the aggregate quality-of-life estimate for a compartment representing many quality-of-life decrements should be derived through a multiplicative model (as in [48, 49, 56]), an additive model (as in [39, 59]), or neither. Issues related to combining QALY estimates have been discussed elsewhere [75].

Several researchers have noted that IDUs do not mix randomly, but rather their injection contact patterns form structured social networks. Studies have revealed the structure of IDU social networks in Colorado Springs [76, 77] and New York City [78]. The compartmental epidemic models described in the previous section assume random mixing, in which each person selects a new partner randomly from the entire population. While random mixing in compartmental models leads to "worst case" epidemics – that is, epidemics with the greatest possible spread of disease among all possible mixing patterns [79-81] – it is unclear whether the random mixing assumption overestimates or underestimates the incremental impact of drug maintenance programs.

Network epidemic models have been used as an alternative to compartmental epidemic models and may be useful if connectivity or network structure is important. However, network models are often significantly more complex than compartmental models. The threshold conditions for an endemic epidemic may be very different for a network model than a compartmental model [82]. Watts looked at epidemic spread in static connected networks and found that network structure had little impact on eventual epidemic outcomes [83]. Zaric directly compared random versus non-random mixing in network epidemic models and found that random mixing led to small increases in the number of new infections [84]. However, the observed difference may be smaller than the range in uncertainty in the parameters of the statistical distributions. To our knowledge, no research has yet directly addressed the question of whether an intervention would appear more or less cost effective when evaluated using a network model with nonrandom mixing versus a model with random mixing.

A compartmental model forces all individuals into a finite number of discrete compartments, with members of each compartment assumed to be homogeneous. In some cases there may be large variations in characteristics of members of various groups. Estimates of injection frequency vary from 1-3 injections per month [85] to more than 100 per month [86]; estimates of the number of new sexual partners also vary over a wide range [87]. Ignoring population heterogeneity by using average or representative values may lead to systematically biased estimates of outcomes when Markov models are used to generate cost-effectiveness ratios [88, 89]. Similar biases may exist in compartmental models.

Pollack noted that the choice of time horizon may be important when compartmental epidemic models are used to evaluate the costs and benefits of medical interventions [90]. For modest interventions (defined as those for which the reduction in the sufficient contact rate is very small) short-term incidence analysis would underestimate long-term effectiveness when the equilibrium prevalence is below 50%, and overstate the long-term benefits when the equilibrium prevalence is above 50%. These findings may have implications for the choice of time horizon for evaluating programs directed to IDUs, where prevalence of HIV and hepatitis C may be very high.

Numerous studies have shown that IDUs who inject cocaine or speedballs (cocaine and heroin mixed together) inject far more often than those who primarily inject heroin. MMT provides relief from opioid dependence but may not have the same impact on cocaine injectors. Some have argued that methadone use may actually lead to an increase in cocaine use [91], or that cocaine users should not be allowed to enter MMT [92]. Future empirical research could look at the impact that MMT has on cocaine injection frequency. Future modeling efforts may involve construction of a model with separate compartments for IDUs who primarily inject heroin and IDU who inject cocaine.

13.4 CONCLUSIONS

Much of the debate around drug treatment is concerned with political and philosophical issues such as whether MMT is a “moral” way to treat opioid dependence. These considerations cannot be ignored in policy formulation [61]. Operations research models cannot address such issues. However, OR models can be used to identify good policies and to distinguish good policies from poor ones. They can also provide methods to facilitate cost-effectiveness analysis and to examine the health and economic tradeoffs associated with drug abuse treatment programs. Analysis of drug abuse treatment programs represents a valuable research area for operations researchers in the future, one where OR models can provide value input to important public policy questions.

Acknowledgments

This work was supported by the Natural Sciences and Engineering Research Council of Canada (NSERC) and the National Institute on Drug Abuse (Grant Number R-01-DA-15612). The author is the M.B.A. ’80 Faculty Fellow at the Ivey School of Business, University of Western Ontario.

References

[1] Spencer, B.D. (1989). On the accuracy of estimates of numbers of intravenous drug users. In *AIDS: Sexual Behavior and Intravenous Drug Use,* C.F. Turner, H.G. Miller, and L.E. Moses, Eds., National Academy Press, Washington, DC.

[2] Hahn, R.A., I.M. Onorato, T.S. Jones, and J. Dougherty (1989). Prevalence of HIV infection among intravenous drug users in the United States. *Journal of the American Medical Association,* 261, 2677-2684.

[3] Nicolosi, A., et al. (1992). Incidence and prevalence trends of HIV infection in intravenous drug users attending treatment centers in Milan and northern Italy, 1986-1990. *Journal of Acquired Immune Deficiency Syndromes,* 5, 365-373.

[4] Centers for Disease Control and Prevention (2001). HIV/AIDS Surveillance Report. *Morbidity and Mortality Weekly Report,* 13.

[5] Hadeler, K.P. and C. Castillo-Chavez (1995). A core group model for disease transmission. *Mathematical Biosciences,* 128, 41-55.

[6] Haverkos, H.W. and W.R. Lange (1990). Serious infections other than human immunodeficiency virus among intravenous drug abusers. *Journal of Infectious Diseases,* 161, 894-902.

[7] Watterson, O., D.D. Simpson, and S.B. Sells (1975). Death rates and causes of death among opioid addicts in community drug treatment programs during 1970-1973. *American Journal of Drug and Alcohol Abuse,* 2, 99-111.

[8] Gronbladh, L., L.S. Ohlund, and L.M. Gunne (1990). Mortality in heroin addiction: impact of methadone treatment. *Acta Psychiatrica Scandinavica,* 82, 223-227.

[9] Gerstein, D.R., et al. (1994). *Evaluating Recovery Services: The California Drug and Alcohol Treatment Assessment (CALDATA),* California Department of Alcohol and Drug Programs, Sacramento, CA.

[10] U.S. Department of Commerce (1997). *Statistical Abstract of the United States,* U.S. Department of Commerce, Washington, DC.

[11] Bell, J. and D. Zador (2000). A risk-benefit analysis of methadone maintenance treatment. *Drug Safety,* 22, 179-190.

[12] Ralston, G.E. and P. Wilson (1996). Methadone programmes: The costs and benefits to society and the individual. *Pharmacoeconomics,* 10, 321-326.

[13] Dole, V.P. and M.E. Nyswander (1976). Methadone maintenance treatment. A ten-year perspective. *Journal of the American Medical Association,* 235, 2117-2119.

[14] Marsch, L.A. (1998). The efficacy of methadone maintenance interventions in reducing illicit opiate use, HIV risk behavior and criminality: a meta- analysis. *Addiction,* 93, 515-532.

[15] Kwiatkowski, C.F. and R.E. Booth (2001). Methadone maintenance as HIV risk reduction with street- recruited injecting drug users. *Journal of Acquired Immune Deficiency Syndromes,* 26, 483-489.

[16] Sambamoorthi, U., L.A. Warner, S. Crystal, and J. Walkup (2000). Drug abuse, methadone treatment, and health services use among injection drug users with AIDS. *Drug and Alcohol Dependence,* 60, 77-89.

[17] Desmond, D.P. and J.F. Maddux (2000). Deaths among heroin users in an out of methadone treatment. *Journal of Maintenance in the Addictions,* 1, 45-61.

[18] Rettig, R.A. and A. Yarmolinsky, Eds. (1995). *Federal Regulation of Methadone Treatment.* National Academy Press, Washington, DC.

[19] U.S. Drug Enforcement Administration (2002). *DEA Briefs: Background, Drug Policy, Drug Scheduling,* http://www.usdoj.gov/dea/pubs/scheduling.html, Accessed August 12, 2002.

[20] Stone, E. (2002). Family torn by addiction. Without methadone clinic, son is forced out of state. *Burlington Free Press,* Burlington, VT, January 20.

[21] Bradley, C.J., M.T. French, and J.V. Rachal (1994). Financing and cost of standard and enhanced methadone treatment. *Journal of Substance Abuse Treatment,* 11, 433-442.

[22] Barnett, P.G. and J.H. Rodgers (1998). *The Cost of Substance Abuse Treatment: A Multivariate Cost Function Using the NDATUS,* VA Health Care System, Palo Alto, CA.

[23] Barnett, P.G. and S.S. Hui (2000). The cost-effectiveness of methadone maintenance. *Mount Sinai Journal of Medicine,* 67, 365-374.

[24] Riza, M. (1998). *Close to Home Online - Policy: The Politics of Methadone,* http://www.pbs.org/wnet/closetohome/policy/html/methadone.html, Accessed August 12, 2002.

[25] Mathias, R. (1997). NIH panel calls for expanded methadone treatment for heroin addiction. *NIDA Notes,* 12.

[26] Lambert, E.Y., H.K. Cesari, R.H. Needle, and J.B. Stein (2002). *Principles of HIV Prevention in Drug-Using Populations: A Research-Based Guide,* National Institute on Drug Abuse, Washington, DC.

[27] U.S. Senate (1999). *Senator John McCain - Press Releases,* http://www.senate.gov/~mccain/methhero.htm, Accessed August 12, 2002.

[28] Swarns, R.L. (1998). Giuliani orders 5 city hospitals to wean addicts off methadone. *New York Times,* New York, August 15.

[29] Rubinowitz, S. (1999). Turnaround Rudy puts $5M in methadone clinics. *New York Post,* New York, October 6, 4.

[30] Lee, M.L. (1992). A Markov model for NIDA data on treatment of opiate dependence. *NIDA Research Monographs,* 128, 160-169.

[31] Weng, T.S. (1992). Toward a dynamic analysis of disease-state transition monitored by clinical laboratory tests. *NIDA Research Monographs,* 128, 137-157.

[32] Johnson, R.E. and P.J. Fudala (1992). Background and design of a controlled clinical trial (ARC 090) for the treatment of opioid dependence. *NIDA Research Monographs,* 128, 14-24.

[33] Phelps, C.E. and A.I. Mushlin (1991). On the (near) equivalence of cost-effectiveness and cost-benefit analyses. *International Journal of Technology Assessment in Health Care,* 7, 12-21.

[34] Donaldson, C. (1998). The (near) equivalence of cost-effectiveness and cost-benefit analyses. Fact or fallacy? *Pharmacoeconomics,* 13, 389-396.

[35] Laska, E.M., M. Meisner, C. Siegel, and A.A. Stinnett (1999). Ratio-based and net benefit-based approaches to health care resource allocation: proofs of optimality and equivalence. *Health Economics,* 8, 171-174.

[36] Bala, M.V., G.A. Zarkin, and J.A. Mauskopf (2002). Conditions for the near equivalence of cost-effectiveness and cost-benefit analyses. *Value in Health,* 5, 338-346.

[37] Dolan, P. and R. Edlin (2002). Is it really possible to build a bridge between cost-benefit analysis and cost-effectiveness analysis? *Journal of Health Economics,* 21, 827-843.

[38] Kaplan, R.M. (1995). Utility assessment for estimating quality-adjusted life years. In *Valuing Health Care: Costs, Benefits, and Effectiveness of Pharmaceuticals and Other Medical Technologies,* F.A. Sloan, Ed., Cambridge University Press, Cambridge, UK.

[39] French, M.T., J.A. Mauskopf, J.L. Teague, and E.J. Roland (1996). Estimating the dollar value of health outcomes from drug-abuse interventions. *Medical Care,* 34, 890-910.

[40] Hannan, T.H. (1976). The benefits and costs of methadone maintenance. *Public Policy,* 24, 197-226.

[41] French, M.T., H.J. Salome, J.L. Sindelar, and A.T. McLellan (2002). Benefit-cost analysis of addiction treatment: methodological guidelines and empirical application using the DATCAP and ASI. *Health Services Research,* 37, 433-455.

[42] Viscusi, W.K. (1993). The value of risks to life and health. *Journal of Economic Literature,* 32, 1912-1946.

[43] French, M.T., *et al.* (2000). Benefit-cost analysis of residential and outpatient addiction treatment in the State of Washington. *Evaluation Review,* 24, 609-634.

[44] McLellan, A.T., *et al.* (1992). The fifth edition of the Addiction Severity Index. *Journal of Substance Abuse Treatment,* 9, 199-213.

[45] Gold, M.R., J.E. Siegel, L.B. Russell, and M.C. Weinstein (1996). *Cost-Effectiveness in Health and Medicine.* Oxford University Press, New York.

[46] Barnett, P.G. (1999). The cost-effectiveness of methadone maintenance as a health care intervention. *Addiction,* 94, 479-488.

[47] Kahn, J.G., et al. (1992). *Updated estimates ofthe impact and cost of HIVprevention in injection drug users. Report preparedfor Centers for Disease Control and Prevention.* University of California, San Francisco, CA.

[48] Zaric, G.S., P.G. Barnett, and M.L. Brandeau (2000). HIV transmission and the cost effectiveness of methadone maintenance. *American Journal of Public Health,* 90, 1100-1111.

[49] Zaric, G.S., M.L. Brandeau, and P.G. Barnett (2000). Methadone maintenance treatment and HIV prevention: A cost effectiveness analysis. *Management Science,* 25, 1013-1031.

[50] Booth, R.E. (1995). Gender differences in high-risk sex behaviors among heterosexual drug injectors and crack smokers. *American Journal of Drug and Alcohol Abuse,* 21, 419-432.

[51] Battjes, R.J., R.W. Pickens, and Z. Amsel (1991). HIV infection and AIDS risk behaviors among intravenous drug users entering methadone treatment in selected U.S. cities. *Journal of Acquired Immune Deficiency Syndromes,* 4, 1148-1154.

[52] Davis, K.R. and S.C. Weller (1999). The effectiveness of condoms in reducing heterosexual transmission of HIV. *Family Planning Perspectives,* 31, 272-279.

[53] Pinkerton, S.D., A.P. Johnson-Masotti, D.R. Holtgrave, and P.G. Farnham (2001). Using cost-effectiveness league tables to compare interventions to prevent sexual transmission of HIV. *AIDS,* 15, 917-928.

[54] Raisch, D.W., C.L. Fye, K.D. Boardman, and M.R. Sather (2002). Opioid dependence treatment, including buprenorphine/naloxone. *Annals of Pharmacotherapy,* 36, 312-321.

[55] Lewis, J.L. and D. Walter (1992). Buprenorphine - background to its development as a treatment for opiate dependence. In *Buprenorphine:*

An Alternative Treatment for Opioid Dependence, J.D. Blaine, Ed., U.S. Department of Health and Human Services, Rockville, MD.

[56] Barnett, P.G., G.S. Zaric, and M.L. Brandeau (2001). The cost-effectiveness of buprenorphine maintenance therapy for opiate addiction in the United States. *Addiction,* 96, 1267-1278.

[57] Barnett, P.G., J.H. Rodgers, and D.A. Bloch (2001). A meta-analysis comparing buprenorphine to methadone for treatment of opiate dependence. *Addiction,* 96, 683-690.

[58] Rosenheck, R. and T. Kosten (2001). Buprenorphine for opiate addiction: potential economic impact. *Drug and Alcohol Dependence,* 63, 253-262.

[59] French, M.T. (2001). Cost-effectiveness of buprenorphine maintenance versus methadone maintenance: a comment on Barnett et al. (2001). *Addiction,* 96, 1515-1517.

[60] Reuter, P. (2001). Cost-effectiveness estimates for buprenorphine should factor in crime. *Addiction,* 96, 1515.

[61] Sindelar, J.L. (2001). Opioid maintenance: the politics matter. *Addiction,* 96, 1517-1518.

[62] Wall, M.J. and H.A. Pollack (2002). Cost-utility of selective buprenorphine substitution for methadone maintenance in preventing HIV transmission. *Value in Health,* 5, 260.

[63] Schottenfeld, R.S., J.R. Pakes, A. Oliveto, D. Ziedonis, and T.R. Kosten (1997). Buprenorphine vs methadone maintenance treatment for concurrent opioid dependence and cocaine abuse. *Archives of General Psychiatry,* 54, 713-720.

[64] U.S. Food and Drug Administration (2002). *Subutex and Suboxone approved to treat opiate dependence,* http://www.fda.gov/bbs/topics/ANSWERS/2002/ANS01165.html, Accessed April 25, 2002.

[65] O'Connor, P.G. and T.R. Kosten (1998). Rapid and ultrarapid opioid detoxification techniques. *Journal of the American Medical Association,* 279, 229-234.

[66] Hamilton, R.J., et al. (2002). Complications of ultrarapid opioid detoxification with subcutaneous naltrexone pellets. *Academic Emergency Medicine,* 9, 63-68.

[67] Kutz, I. and V. Reznik (2001). Rapid heroin detoxification using a single high dose of buprenorphine. *Journal of Psychoactive Drugs,* 33, 191-193.

[68] Thirion, X., et al. (2002). Buprenorphine prescription by general practitioners in a French region. *Drug and Alcohol Dependence,* 65, 197-204.

[69] Strang, J., J. Sheridan, and N. Barber (1996). Prescribing injectable and oral methadone to opiate addicts: results from the 1995 national postal survey of community pharmacies in England and Wales. *British Medical Journal,* 313, 270-272.

[70] Wodak, A. (2002). Methadone and heroin prescription: babies and bath water. *Substance Use and Misuse,* 37, 523-531.

[71] Wenger, L.D. and M. Rosenbaum (1994). Drug treatment on demand – not. *Journal of Psychoactive Drugs,* 26, 1-11.

[72] Simeone, R.S. (1993). A note on waiting lists and demand estimation. *International Journal of the Addictions,* 28, 1033-8.

[73] Kaplan, E.H. and M. Johri (2000). Treatment on demand: An operational model. *Health Care Management Science,* 3, 171-183.

[74] Guydish, J., *et al.* (2000). Drug abuse treatment on demand in San Francisco: preliminary findings. *Journal of Psychoactive Drugs,* 32, 363-370.

[75] Bonds, D.E. and K.A. Freedberg (2001). Combining utility measurements – Exploring different approaches. *Disease Management and Health Outcomes,* 9, 507-516.

[76] Darrow, W.W., et al. (1999). Using knowledge of social networks to prevent human immunodeficiency virus infections: The Colorado Springs study. *Sociological Focus,* 32, 143-158.

[77] Klovdahl, A.S., et al. (1994). Social networks and infectious disease: The Colorado Springs study. *Social Science and Medicine,* 38, 79-88.

[78] Neaigus, A., et al. (1996). High-risk personal networks and syringe sharing as risk factors for HIV infection among new drug injectors. *Journal of Acquired Immune Deficiency Syndromes,* 11, 499-509.

[79] Kaplan, E.H., P.C. Crampton, and A.D. Paltiel (1989). Nonrandom mixing models of HIV transmission. In *Mathematical and Statistical Approaches to AIDS Epidemiology,* C. Castillo-Chavez, Ed., Springer-Verlag, Berlin.

[80] Kaplan, E.H. and Y.S. Lee (1990). How bad can it get? Bounding worst case endemic heterogeneous mixing models of HIV/AIDS. *Mathematical Biosciences,* 99, 157-180.

[81] Kaplan, E.H. (1992). Asymptotic worst-case mixing in simple demographic models of HIV/AIDS. *Mathematical Biosciences,* 108, 141-156.

[82] Pastor-Satorras, R. and A. Vespignani (2001). Epidemic spreading in scale-free networks. *Physical Review Letters,* 86, 3200-3203.

[83] Watts, D.J. (1999). *Small Worlds: The Dynamics of Networks Between Order and Randomness.* Princeton University Press, Princeton, NJ.

[84] Zaric, G.S. (2002). Random vs. nonrandom mixing in network epidemic models. *Health Care Management Science,* 5, 147-155.

[85] Koblin, B.A., J. McCusker, B.F. Lewis, and J.L. Sullivan (1990). Racial/ethnic differences in HIV-1 seroprevalence and risky behaviors among intravenous drug users in a multisite study. *American Journal of Epidemiology,* 132, 837-846.

[86] Meandzija, B., P.G. O'Connor, B. Fitzgerald, B.J. Rounsaville, and T.R. Kosten (1994). HIV infection and cocaine use in methadone maintained and untreated intravenous drug users. *Drug and Alcohol Dependence,* 36, 109-113.

[87] Laumann, E.O., J.H. Gagnon, R.T. Michael, and S. Michaels (1994). *The Social Organization of Sexuality: Sexual Practices in the United States.* University of Chicago Press, Chicago, IL.

[88] Kuntz, K.M. and S.J. Goldie (2002). Assessing the sensitivity of decision-analytic results to unobserved markers of risk: Defining the effects of heterogeneity bias. *Medical Decision Making,* 22, 218-227.

[89] Zaric, G.S. (2002). The danger of ignoring population heterogeneity when Markov models are used in cost effectiveness analysis. *Value in Health,* 5, 125.

[90] Pollack, H.A. (2001). Ignoring 'downstream infection' in the evaluation of harm reduction interventions for injection drug users. *European Journal of Epidemiology,* 17, 391-395.

[91] Weber, J.C. and J. Kopferschmitt (1998). Substitution treatments for drug abusers. *Presse Medicale,* 27, 2088-2099.

[92] Caulkins, J.P. and S.L. Satel (1999). Methadone patients should not be allowed to persist in cocaine use. *Federation of American Scientists' Drug Policy Analysis Bulletin,* 6, 1-4.

14 HARM REDUCTION IN THE CONTROL OF INFECTIOUS DISEASES AMONG INJECTION DRUG USERS

Harold Pollack

School of Public Health
University of Michigan
Ann Arbor, MI 48109

SUMMARY

Operations research has contributed to the control of blood-borne epidemics among injection drug users. The analysis of random-mixing models has led to a deeper understanding of both syringe exchange programs and substance abuse treatment in the control of HIV/AIDS and hepatitis. This chapter presents some of these results, and analyzes illustrative models to show how simplified, but empirically pertinent mathematical models can assist policymakers evaluate public health interventions.

KEY WORDS

Epidemic control, HIV, Hepatitis C

14.1 INTRODUCTION

Public policymakers have tried many approaches to address the social, economic, and medical problems associated with substance abuse. For many participants in the drug policy debate, “harm reduction” provides the touchstone in evaluating the success of these efforts, and a useful alternative to simple “use reduction” as a guide for public policy [1]. Harm reduction admits diverse meanings among policymakers, clinicians, and academic researchers in the drug policy debate. Yet it commands broad support among those who seek to balance the competing harms caused by both drug use and by public policies to hinder, deter, or punish drug use [2].

Harm reduction has proved especially important in evaluating clinical and policy responses to injection drug use. Injection drug users (IDUs) have long experienced high rates of avoidable mortality and morbidity from infectious disease [3]. The most deadly threat now arises from HIV/AIDS. Yet less visible infectious diseases, especially hepatitis B and C, endocarditis, and tuberculosis also threaten the health and survival of IDUs. Heroin overdose provides an additional source of premature mortality and morbidity among IDUs [4].

Some problems associated with drug use are intimately connected with the intensity and the duration of drug consumption among IDUs. For example, interventions to reduce property crime by IDUs may fail if they do not reduce individual expenditures on illicit drugs. Yet such use reduction is sometimes impossible or unnecessary to achieve the desired policy goal [3]. Many OECD (Organisation for Economic Cooperation and Development) countries have successfully reduced HIV incidence and prevalence among IDUs – even within populations that continue to regularly inject heroin or other illegal drugs [5].

Harm reduction provides the guiding question, though not a clear algorithm, to address these concerns. From this perspective, policy analysts, clinicians, and policymakers seek to clarify the goals of public policy, and to scrutinize the ability of specific policies to advance the well-being of the general community and of drug users themselves.

14.2 CLINICAL AND POLICY RESPONSES

This chapter focuses on two kinds of interventions often described under the rubric of harm reduction: substance abuse treatment (specifically, methadone maintenance treatment) and syringe exchange programs. However, to place these interventions in context one must consider their place in broader public policy.

Three kinds of public policy interventions seek to address such threats to life and limb among IDUs: supply-side or demand-side law enforcement, harm reduction interventions such as syringe exchange for active drug users, and substance abuse treatment. Operations researchers have performed important policy analysis of all three kinds of intervention. The impact and cost-effectiveness of law enforcement efforts are outside the scope of this chapter. Because operations researchers have contributed to policy analysis of all three kinds of interventions, and because the term "harm reduction" is sometimes applied to analyze law enforcement policies, we briefly discuss this research.

Supply- and demand-side law enforcement The most traditional drug policy interventions are supply-side law enforcement efforts to deter or punish individuals who sell or distribute heroin or other injectable drugs. Source-country enforcement activities and border interdiction efforts seek to disrupt the organizations and firms involved in drug trafficking. Supply-side enforcement also encompasses the arrest of street-level drug users, a subject of great complexity given the vagaries of low-wage labor markets for potential drug sellers and the high prevalence of substance use and dependence among street-level dealers [6, 7].

Such efforts seek to contract drug supply, thereby raising market prices of illicit substances. In economic terms, these law enforcement policies reduce the quantity of drugs supplied at any specific market price – a shift in the supply curve – raising market prices and reducing drug consumption in the resulting market equilibrium [8].

The short-term and long-term effectiveness of interdiction and source country policies is influenced by the elasticity of supply for illicit substances. If drug suppliers are price-elastic, supply-side law enforcement is likely to have a small effect on both prices and consumption. Effective interdiction and source-country policies will simply induce new entrants to the market.

The effectiveness of supply-side enforcement is also influenced by the responsiveness of IDUs to changes in market prices. If the price-elasticity of demand for heroin is less than –1.0, enforcement-linked price increases will induce accompanying reductions in both drug consumption and in overall expenditures by IDUs. If the quantity consumed is insensitive to market prices, price increases will induce only a small decline in heroin consumption and will induce an overall *increase* in overall expenditures for heroin among IDUs [9].

Demand-side enforcement measures include penalties for drug possession and purchase. Such policies raise the (non-price) costs of illicit substance use, and may thereby reduce demand for these substances [10]. Such policies are attractive if they deter substance use, and attractive as a mechanism to reduce the profits associated with illicit drug sales. Moreover, many IDUs commit larceny and other property crimes. Arresting IDUs for simple drug possession may therefore be a low-cost means of incapacitating (non-drug) criminal offenders [11].

An important drawback of enforcement policies is that they impose large costs on both individual IDUs and on the wider society. IDUs bear the short-term and often lifelong consequences of incarceration or other judicial interventions. Taxpayers must finance law enforcement and correctional interventions. Moreover, specific law enforcement strategies such as aggressive enforcement of illicit drug paraphernalia laws may encourage needle-sharing, unsterile discard of used syringes, and other high-risk behaviors [12].

If substance abuse treatment or prevention interventions can halt, reduce, or prevent injection drug use, less punitive alternatives may be preferable to criminal justice interventions. Studies by Caulkins and colleagues examine the cost-effectiveness (cost per unit of reduced drug consumption) of a wide range of prevention, treatment, and criminal justice system interventions [11, 13]. Such studies provide strong support for the cost-effectiveness of prevention and treatment interventions.

The price-elasticity of demand for illicit drugs is especially important from the perspective of crime control, since many drug users finance their consumption through property crime or other illegal activities [14]. Operations researchers play an important role in this debate through the detailed analysis of illicit drug markets and the relationship between interdiction efforts and resulting drug prices [15]. Operations researchers, including Caulkins and colleagues, have explored these issues in some depth. Using data from the Drug Enforcement Administration’s System to Retrieve Information from Drug Evidence (STRIDE) database, these researchers have explored regional variations and changing market conditions for marijuana, cocaine, heroin, and other illicit drugs [15, 16].

Three findings from this literature are noteworthy.

One striking finding speaks to the difficulties of supply-side enforcement. Purity-adjusted illicit drug prices have declined despite substantial supply-side interdiction and enforcement efforts [2, 10]. Declining prices and increasing purity of street heroin have posed complex challenges for substance abuse treatment, and have fostered non-injecting forms of heroin use such as snorting [17].

A second finding is that drug users respond to the price of illicit drugs, particularly in the long run [9]. This result – predicted by "rational addiction" dynamic optimization models of drug consumption – suggests that the long-term effect of price decreases is to significantly increase the number of illicit drug users. Optimal control theory and other operations research methods have been applied, profitably, by health economists and others seeking to understand illicit drug markets [18].

A third and related insight speaks to the dynamic character of drug markets [19]. Patterns of illicit drug consumption changed rather rapidly over the 1975-2000 period, with current prevalence responsive to past consumption, positive and negative "role modeling" by current and past drug users, and other feedback effects. Forecasting models have been developed to explore these effects.

Of particular importance are the transitions in drug-using behavior among IDUs. The propensity of light or casual users to become heavy users, and quit rates among different categories of users powerfully influence the number of future IDUs, and influence the likely social harms associated with different forms of substance use [20].

The remainder of this chapter focuses on the two remaining kinds of harm reduction interventions, substance abuse treatment and syringe exchange programs. For more information on substance abuse policy, see Kleiman [7], MacCoun and Reuter [2], and the collection of essays edited by Heymann and Brownsberger [21].

Substance abuse treatment includes a broad array of inpatient and outpatient medical, psychiatric, and social service interventions designed to halt or reduce illicit drug use [4, 22]. This chapter focuses on methadone maintenance treatment (MMT), because this is the principal modality used to treat injection drug use. Massing [23] describes the history and development of MMT. Although many challenges exist to the effectiveness of MMT, ranging from inadequate dosing to the difficulties of treating poly-drug use, the impact and cost-effectiveness of MMT is well established [24, 25].

The value of such substance abuse treatment has been underscored by randomized trials of MMT. In one study of Swedish IDUs, 2 of 17 members of the non-MMT control group died from apparent overdose. One other member of the control group suffered a leg amputation, while two others suffered severe infection. Among the remaining controls, two were incarcerated, and 9 of the remaining 10 continued illicit drug use. Over the same period, none of the MMT group suffered major health problems, and 13 of the original 17 were no longer using illicit drugs [26]. Three more members of the control group died over the following three years, in a study completed before the era of HIV/AIDS [4].

Syringe exchange programs (SEPs) are a more pure form of "harm reduction." Although the design and operation of SEPs differ, the common aim is to prevent infectious disease transmission among active IDUs through the provision of sterile injection equipment and through the safe collection of discarded syringes. To focus on the harm reduction dimension of SEPs, we assume in this chapter that such interventions have no other impact on the frequency of drug use among IDUs, and that SEP has no impact on the removal of program clients from the population of active IDUs. Because this chapter does not consider the role of SEP as a conduit into MMT or other treatment and social services, this is an important oversimplification [27]. A fuller treatment would likely indicate greater impact and cost-effectiveness of SEPs.

14.3 THE CONTRIBUTION OF OPERATIONS RESEARCH TO POLICY

Many clinicians and policy makers are skeptical about the merits of analytic modeling to scrutinize drug control policies, especially the special problems of IDUs.[1] Much of this skepticism arises because of real limitations of the data and models available to study this population. IDUs are a hidden population whose risk behavior, and even whose absolute numbers, are imperfectly known [29]. Basic parameters must be indirectly inferred from fragmentary data. Nationally representative surveys provide poor coverage of high-risk populations, including IDUs. Data from clinical services such as hospital emergency departments or drug treatment programs are based upon a self-selected group of patients and may not apply to out-of-treatment IDUs [30].

The probability of infection with HIV or hepatitis when a susceptible IDU uses an infected needle is imperfectly known. Several analyses seek to estimate this parameter based upon needle-stick accident data among hospital personnel. Other analyses indirectly estimate these probabilities based upon observed patterns of disease spread [31]. Neither method is fully satisfactory in characterizing risk exposures among IDUs.

[1] This section modifies the discussion in Pollack [28].

Observational studies suggest that methadone maintenance treatment (MMT) reduces the rate of new HIV infections (HIV incidence) among IDUs. MMT clients are less likely than out-of-treatment IDUs to share needles, inject drugs less frequently, and are less likely to practice other behavioral risks [32]. Several studies document large differences in HIV incidence between steady methadone clients and out-of-treatment IDUs [33-35].

The impact of MMT on hepatitis C (HCV) transmission is less encouraging. Like HIV, HCV is spread through sharing of infected injection equipment, including syringes, "cookers," and filters [36, 37]. However, studies of both IDUs and health care workers exposed to needle-stick injuries indicate that HCV is more efficiently transmitted [38, 39].

This high infectivity poses a basic challenge to any prevention intervention that seeks to reduce the frequency and duration of injection drug use. From an analytic perspective, differences between HIV and HCV underscore the difficulties one encounters in evaluating interventions. Individual behavior changes and other impact measures may be readily observed. Yet these measures are difficult to link with underlying patterns of infectious disease spread. Analytic models become essential to make this connection, to clarify the value of alternative data sources and measures, and to scrutinize causal assumptions that undergird prevention interventions.

Most sobering are the many IDU populations with low HIV prevalence but endemic prevalence of HCV. Pollack and Heimer [40] reviewed published literature on HCV prevalence among European IDUs. Although results vary across populations, most studies found prevalences between 65-85%. Only four of 40 examined studies found HCV prevalence below 50% [40]. Similar results are found in studies of MMT clients [41-43], including studies in the U.S., Australia, and many places in Western Europe, typically reporting HIV prevalence below 10 percent, but HCV prevalence exceeding 70 percent [44-46]. Results for young IDUs are somewhat more hopeful [45-48]. However, other studies have yielded more disappointing results [49, 50].

HCV prevalence comparisons between MMT clients and out-of-treatment IDUs have yielded mixed results. Out-of-treatment IDUs are often found to have lower HCV prevalences. However, this result may be confounded by the older age of the in-treatment population.

Epidemiological studies and analytic models of syringe exchange programs (SEPs) indicate the same contrasts between HIV and HCV prevalence. Many studies indicate that SEPs can reduce HIV incidence. Such findings undergird the long-standing support by most public health researchers for syringe exchange and similar programs [30, 51].

The demonstrated impact of SEPs in preventing HCV transmission is less favorable. Using SEP data from Seattle, Hagan and collaborators found no protective effects [52]. Theoretical analysis due to Pollack found little impact and poor cost-effectiveness of typical SEPs in HCV prevention [53, 54]. As discussed below, models based upon the short-term impact of syringe exchange may greatly overstate long-term SEP effectiveness in reducing incidence of highly infectious agents such as HCV.

Sexual and needle-sharing mixing patterns among IDUs are also poorly understood. Models in which IDUs share needles with random partners are widely used because random mixing provides a tractable worst-case analysis [55-57]. However, social network models are likely to provide a more sociologically plausible pattern of infectious disease spread [58, 59].

Equally important, rigorous evaluations of specific interventions may not be generalizable across populations and settings. Substance abuse treatment and SEPs differ greatly in both effectiveness and cost. Such diversity calls into question any analysis that draws sweeping comparisons across diverse categories of competing interventions [30].

Although one must acknowledge reasons for skepticism in applying analytic models to policy, such efforts provide important contributions to policy debates. Modeling exposes for scrutiny the implicit assumptions that policymakers are already using in addressing injection drug use. Public policies are often based upon unexamined assumptions that appear questionable or implausible when brought to light.

For example, some clinicians advocate the proliferation of difficult-to-reuse syringes to slow HIV spread. Simple but compelling epidemiological models indicate that, if the frequency of injection among IDUs is insensitive to the supply of new needles, such devices are likely to accelerate infectious disease spread [60].

As another example, Kaplan and Pollack reviewed procedures used to allocate HIV prevention resources [30, 61]. Many U.S. policy makers try to allocate resources based upon the number of individuals in each risk group. Such an approach is inappropriate when either program effectiveness or HIV incidence varies across the pertinent risk groups.

Worse, the political and organizational realities of group decision processes easily foster arbitrary policies and arbitrary allocation of resources. Altman and colleagues note (p. 81) that health planners respond to technical and political uncertainty by seeking "convenient proxies for need to be applied in allocation decisions" [62]. Wary of debating the merits of specific facilities, many health system planners are drawn to elaborate need-assessment formulas to evaluate proposed services.

Such methods provide poor guidance regarding the impact or cost-effectiveness of proposed expenditures, but find wide appeal as planners seek credible focal points to resolve internal disputes and to justify controversial policies. Such approaches are widespread in many areas of resource allocation [63]. Brandeau and colleagues and Kaplan discuss more rigorous and explicit approaches to allocating scarce resources [61, 64]. Explicit modeling helps to discipline group decision processes, and allows policymakers to explore the unintended assumptions and consequences of appealing but limited algorithms that are widely used to allocate resources.

Models also help policy makers understand the linkage between the available data and the latent causal processes one seeks to influence through public intervention. Paltiel and Stinnett describe many ways that analytic models can interrogate the premises and likely consequences of policy interventions [65].

Analytic models clarify the links between readily-measured or readily-influenced intermediate outcomes and the ultimate outcomes of direct policy concern. Many of the best evaluations of HIV prevention interventions do not directly scrutinize HIV incidence among program participants. Rather, such evaluations explore the impact of such interventions on important behavioral risks [66-68]. Analytic models help establish the linkage between these behavioral risks and actual health outcomes. For interventions such as syringe exchange that have not been (or cannot be) evaluated through prospective randomized trials, analytic models can scrutinize the findings of bservational studies of those interventions.

Analytic models can also identify the kinds of data required for resource allocation and for other public health functions. Public health reporting and data systems are largely designed to accomplish classic functions of epidemiological surveillance such as case enumeration and contact tracing. The quality and performance of such systems is traditionally scrutinized through such measures as the completeness of case finding and avoidance of duplication when the same case is reported multiple times or in multiple jurisdictions.

Although such performance measures are pertinent to the provision of medical and other services to all infected individuals, they are sometimes misleading when surveillance data are used for other purposes. When allocating a fixed pool of resources across populations and jurisdictions, the most important characteristic of HIV or other surveillance systems is their ability to provide comparable and unbiased estimates of prevalence and incidence across populations. Researchers are beginning to apply more explicit scrutiny to the process of funding allocation, and are examining how different approaches to centralized resource allocation influence resource allocation across competing jurisdictions [69].

Techniques such as sensitivity analysis can also support the reliability and robustness of even highly simplified or empirically uncertain models in providing policy guidance. For example, research on nonrandom mixing and random graph theory highlights the value of random mixing models in characterizing infectious disease transmission within high-risk populations [70].

Operations researchers also help direct the attention of policymakers and analysts to critical concerns that might be overlooked in the absence of formal analytical models. As an example, infectious disease prevention measures are typically based upon disease prevalence across the population. Prevalence is easily measured using existing clinical data systems when infected individuals reliably seek medical attention. Moreover, prevalence-based allocation is often the best strategy to allocate *treatment* resources across competing populations and interventions. However, in a changing epidemic, current prevalence may provide poor guidance about the specific risk groups that are currently experiencing the highest rate of new infection. HCV incidence analysis indicates that young and inexperienced IDUs are experiencing high rates of new infection [71]. For HIV, incidence-based resource allocation is likely to channel greater resources to nonwhites and to residents of southeastern states [30, 72].

Analytic techniques can also demonstrate how interventions that are effective for one problem are likely to be much less effective in addressing related problems in a different setting. As discussed below, simple analytic models help researchers and policymakers to establish the success of SEPs in slowing HIV spread. Short-term reduction in HIV transmission is sufficient to reduce long-run incidence and prevalence because the HIV virus, though deadly, is inefficiently transmitted in each individual act of needle sharing between infected and uninfected persons. However, infectious disease transmission models indicate that similar-quality SEP interventions are less effective in the prevention of HCV than in the prevention of HIV [53, 54, 73, 74]. This has been observed in many IDU populations, which display endemic HCV prevalence despite well-implemented prevention interventions that successfully maintain low HIV prevalence [42, 43].

14.4 AN ANALYTIC MODEL OF BLOOD-BORNE DISEASE AMONG IDUs

Basic insights can be illustrated using a simple but useful epidemiological model of infectious disease transmission among IDUs. This section presents an analytic framework that examines the long-term and short-term impact of both MMT and SEPs in reducing infectious disease spread. This model focuses on the *cost-effectiveness* of such interventions, conceptualized as the costs per averted HIV or HCV infection associated with the prevention intervention. It does not include a more complete cost-utility model. Zaric and colleagues have published several analyses from a cost-utility perspective [57, 75].

The model below, like others in the policy analysis literature, is based on a simplified depiction of injection drug use and treatment interventions. It does not consider heterogeneity among IDUs in the manner, frequency, and social context of their injection drug use, though these characteristics are known to vary among IDUs [36]. It does not consider differences in transmission risk associated with viral load and other complex characteristics of infected and uninfected persons. It uses a standard, random-mixing model of infectious disease transmission rather than a more sociologically nuanced network model of needle-sharing networks [76]. It does not consider sexual risk among IDUs.

Each of the above simplifications is costly, because each excludes something important for infectious disease spread. Despite these simplifications, the resulting model illuminates the basic trade-offs that confront policymakers, and it helps to identify critical parameters that determine likely policy success. By allowing explicit cost-effectiveness calculations, this model provides a simplified, but useful yardstick to compare MMT to other prevention efforts.

We use the model presented by Pollack and Heimer [40] to present the basic story. In particular, we consider a self-contained population of some N(t) active IDUs. This number might vary over time as a result of prevention interventions to discourage drug use. New (uninfected) IDUs enter the population at a constant rate of θ per day. IDUs leave the population at random at some constant rate of δ per person per day. This implies that the average duration of an active drug use "career" is $(1/\delta)$. In a particularly unrealistic but useful assumption, the exit rate δ is assumed to be independent of both disease status and one's previous experience as a drug user. Averaging estimates from Kaplan's "needles that kill" analysis and those reported from among Baltimore's ALIVE cohort, we set δ=1/(3994 days) [53, 77, 78].

We collapse all injecting equipment into a single entity – syringes – and we posit that IDUs freely share syringes at shooting galleries and similar venues that promote random mixing.[2] These circumstances promote rapid disease spread as susceptible IDUs encounter contaminated, potentially infectious injection paraphernalia. Although random mixing is a worst-case assumption, mathematical models indicate that it provides a good approximation to non-random models when there are high contact rates and some overlap across disparate sharing networks [56, 70, 85].

Drug users are assumed to frequent these locations with a constant arrival rate of λ per unit time. The true value of λ is difficult to directly observe. Some research has assumed that IDUs share syringes once per week [31]. More recent data suggest less frequent sharing, though IDUs may under-report the extent of sharing [86]. Infectious disease transmission can occur when an uninfected person shares a syringe first used by an infected person. When such sharing occurs, we assume a constant probability of κ that the virus is actually transmitted. Rather than pick a specific point estimate, we examine a range of values across the empirically pertinent range from a low of $\kappa = 0.5\%$ to a high value of $\kappa = 7.5\%$. The low value corresponds to published analyses HIV transmission, while the latter value is extrapolated from data from needle-stick accidents involving health care workers [57].

At any given time t, there are some I(t) infected individuals. The proportion of infected individuals is the ratio $\pi(t)=I(t)/N(t)$.

Table 14.1 summarizes the relevant parameters and simulation values.

14.4.1 Baseline epidemiological model

The basic model is most readily presented in the absence of policy intervention. On any given day, $N(t) - I(t) = N(t)[1-\pi(t)]$ uninfected IDUs remain susceptible to infection. Each IDU shares at a rate of λ per day. Given random mixing, the probability that an uninfected IDU shares a needle with an infected IDU is identical to $\pi(t)$, the proportion of infected persons within the active population of IDUs. When a susceptible shares with an infected person, she has probability κ of becoming infected.

[2] IDUs also share 'cookers' and filters, and water sources contaminated by syringe mixing. IDUs also use previously-used syringes in ways that allow for further infection [79-84]. Cookers, filters, and water may be more important for HCV transmission than for HIV given the differences in infectivity between the two agents.

Table 14.1 Parameters and values

Para-meter	Definition	Baseline Value
N(t)	IDU population	See text
I(t)	Number of infected	See text
$\pi(t)$	HIV or HCV prevalence	See text
θ	Arrival rate into IDU population	0.5/day
λ	Arrival rate into shooting galleries	1/(7days)
κ	Infectivity	0.005 – 0.075
δ	Exit rate from active IDU population	1/(4000 days)
M	Number of treatment slots	See text
C	Treatment cost/person/day	$14.00
d	Cost of SEP/person/day	$5.00
β	Reduction in injection rate during treatment	75%
γ	Proportional reduction in syringe sharing rate associated with syringe exchange	1/3
μ	Exit rate from treatment	1/(400 days)
	Analytic results	
V_0	Present discounted value of infections without any MMT slots	See text
V_M	Present discounted value of infections given M MMT slots	See text
ι_0	Steady-state infectious disease incidence without intervention	See text
N_0	Steady-state population size without intervention	See text
π_0	Steady-state prevalence without intervention	See text
N^*	Steady-state population size with MMT implemented	See text
$\Delta\iota_{\text{short-term}}$	Short-term incidence decline attributable to SEP	See text
$\Delta\iota_{\text{long-term}}$	Long-term incidence decline attributable to SEP	See text

Combining these terms, infectious disease incidence – the number of new infections per day – is

$$\iota_0(t) = \kappa\lambda\pi_0(t)[1-\pi_0(t)]N_0(t) \tag{1}$$

Here the subscript 0 is used to denote variables in the absence of intervention. The epidemic spreads most rapidly when half of the population is infected. At this prevalence, the number of sharing pairs that involve one infected and one uninfected person is maximized.

Since some individuals exit the population of active IDUs,

$$\frac{dI_0(t)}{dt} = \iota_0(t) - \delta I_0(t) \tag{2}$$

In steady state, $I_0(t)$ is no longer a function of time; that is, $dI_0(t)dt=0$. We therefore use the subscript 0 and omit references to t to indicate a steady-state value.

Thus, $I_0 = \iota_0(1/\delta)$. This implies that the number of infected IDUs equals the rate of new infections per unit time multiplied by the mean duration of infected IDUs within the population.

The same analysis indicates that the proportional steady-state prevalence, $\pi_0 = I_0/N_0$, is

$$\pi_0 = 1 - \frac{\delta}{\kappa\lambda} = 1 - \frac{1}{R_0} \tag{3}$$

The quantity $(\delta/\kappa\lambda)$ is the reciprocal of the reproduction number R_0. Absent intervention, R_0 is the expected number of individuals who would be infected by a single infected drug user introduced into an entirely susceptible population. Clinical or policy interventions that drive R_0 below 1.0 will drive steady-state prevalence to zero. Such interventions might reduce the frequency of needle sharing (λ) through health education interventions, increase the rate of exit (δ) from the IDU population, or reduce infectivity (κ) through the provision of bleach to clean potentially infected syringes. More complex models yield different values of R_0. This parameter is fundamental to many epidemiological policy models of infectious disease spread [87].

One must consider the size of the overall population of IDUs. Every day, some number θ of uninfected IDUs enter the population. If there are N(t) IDUs at time t, some $(N(t)\delta)$ will exit the IDU population every day. In steady state, there will be N_0 active drug users, where the number of new IDUs entering the population balances the number of IDUs who leave the population. These flows balance when population size is equal to the arrival rate θ of new individuals per unit time, multiplied by the mean length of time $(1/\delta)$ that an individual remains an active IDU:

$$N_0 = \frac{\theta}{\delta} \tag{4}$$

Finally, one must consider steady-state disease incidence, ι_0. Since steady-state prevalence equals incidence multiplied by duration of IDUs within the active population, we have $\iota_0 = \delta I_0 = \delta N_0 \pi_0$. This implies that

$$\iota_0 = \delta N_0 \left[1 - \frac{1}{R_0}\right] = \theta \left[1 - \frac{1}{R_0}\right] \tag{5}$$

In cost-effectiveness analysis, the important quantity is the number of averted infections associated with treatment intervention. However, the *timing* of infections also matters. An averted infection five years from now is less valuable than an averted infection today. Given the time value of money, future averted infections must be discounted by precisely the same factor as the funds expended to finance the intervention. Given a discount rate r, the present discounted value of new infections is expressed mathematically as

$$V_0 = \int_0^\infty \iota_0(t) e^{-rt}\,dt = \int_0^\infty \kappa\lambda\pi_0(t)[1 - \pi_0(t)]N_0(t)e^{-rt}\,dt \tag{6}$$

Here r is a discount rate appropriate for public policy intervention.

14.4.2 The impact of syringe exchange and methadone maintenance treatment

One can augment the basic model to consider the impact of both methadone treatment and syringe exchange. This model abstracts from a complex reality to highlight the qualitative impact of both kinds of interventions. MMT is presumed to induce a constant exit rate from the drug-using population of μ per person per unit time, over and above the "natural" exit rate δ from the drug-using population. MMT also reduces the rate of hazardous syringe sharing among clients who would otherwise use illicit drugs. Instead of going to shooting galleries at a rate of λ times per week, MMT clients frequent these places at the rate of $\lambda(1-\beta)$. Complete adherence corresponds to a value of $\beta=1.0$.

To focus on the harm reduction dimension distinctive to SEP, the intervention is presumed to have *zero* impact on the frequency of drug use, and no impact on the exit rate of IDUs from the population of active injectors. Instead of going to shooting galleries at a rate of λ times per week, SEP clients frequent these places at the rate of $\lambda(1-\gamma)$.

For both SEP and MMT, we assume that disease prevalence among treatment participants mirrors prevalence among all IDUs. On any given day, $N(t) - I(t) = N(t)[1-\pi(t)]$ uninfected drug users remain susceptible to infection. However, uninfected MMT clients who adhere to treatment do not share needles. Assuming that disease prevalence among methadone clients mirrors prevalence in the broader drug-using population, and that treatment reduces syringe sharing by the proportion β, we must subtract $M\beta[1-\pi(t)]$ from the population of those at risk, leaving $(N(t)-\beta M)[1-\pi(t)]$ susceptible drug users who are actively at risk.

Both of these factors alter infectious disease incidence to

$$\iota(t)=\kappa\lambda(1-\gamma)\pi(t)[1-\pi(t)][N(t)-\beta M] \tag{7}$$

In like fashion,

$$\frac{dI(t)}{dt}=\iota(t)-(\delta+M\mu)I(t) \tag{8}$$

Each MMT “slot” costs $C per person per day in pharmaceutical costs, labor, and other expenses. Treatment slots are always filled. This assumption matches conditions of excess demand in many U.S. and European cities that experience long waiting lists. Following previous research, we posit that $C is approximately $14/person/day. Each SEP treatment slot costs some $d per person per day. Because SEP is a less intensive intervention, we posit that d is approximately $5/person/day.

As in the baseline model, some θ uninfected IDUs enter the population every day. Only now, if M IDUs receive MMT, some number $(N(t)\delta+M\mu)$ will exit the population every day. In steady state, there will be N active drug users, where

$$N^* = \frac{\theta - M\mu}{\delta} \tag{9}$$

Thus, one benefit of MMT is to reduce the overall population of active drug users.

Given M treatment slots, the present discounted value of new infections is

$$V_M = \int_0^\infty \iota(t)e^{-rt}dt = \int_0^\infty \kappa\lambda(1-\gamma)\pi(t)[1-\pi(t)](N(t)-\beta M)e^{-rt}dt \tag{10}$$

In similar fashion, the present discounted cost of maintaining M treatment slots in perpetuity is \$Mc/r. If, considering treatment costs, the reduced lifespan and the reduced well-being of infected persons, one values an averted infection at some monetised level \$S, the optimum policy is to choose the number of slots M that minimises SV_M-Mc/r, the present monetised value of disease incidence minus the overall treatment cost.

14.5 AVERAGE COST PER AVERTED INFECTION

In comparing MMT to other prevention efforts or other competing uses of public funds, it is especially illuminating to calculate the average cost of MMT per averted infection. If there are no available treatment slots, the present discounted value of new infections is some (larger) quantity V_0. So the average cost per averted infection would be

$$\frac{Mc}{r[V_0 - V_M]} \tag{11}$$

Unfortunately, V_M is difficult to solve analytically, though it is easily computed numerically in specific cases. Pollack [88] provides further specific results.

Table 14.2 is drawn from Pollack [88]. It shows the results of one sensitivity analysis generated using these models. Compared with later analyses, including those by Zaric and colleagues [57, 75], Pollack [88] likely understates the cost-effectiveness of MMT for HIV prevention. As discussed below, costs per averted HIV infection strongly increase with underlying HIV prevalence in the absence of intervention. MMT and other harm reduction interventions are highly cost-effective when applied in conditions of relatively low prevalence. Such interventions are markedly less cost-effective in conditions of very high prevalence because feasible interventions have only a small impact on steady-state prevalence.

Table 14.2 Average cost per averted Hepatitis C infection (60% of IDUs in MMT): κ=0.01, varying rates of needle sharing and treatment adherence

30% of MMT Clients Share Needles	**50% Treatment Adherence (β=0.5)**	**75% Treatment Adherence (β=0.75)**	**Full Treatment Adherence (β=1.0)**
90% Relapse	$321,304	$278,720	$240,166
80% Relapse	$140,655	***$113,083***	$103,634
70% Relapse	$114,072	$104,695	$99,540
60% Relapse	$107,140	$101,912	$98,419
20% of MMT Clients Share Needles	**50% Treatment Adherence**	**75% Treatment Adherence**	**Full Treatment Adherence**
90% Relapse	$481,932	$418,062	$360,236
80% Relapse	$210,983	$169,625	$155,458
70% Relapse	$171,108	$157,042	$149,306
60% Relapse	$160,710	$152,868	$147,630
10% of MMT Clients Share Needles	**50% Treatment Adherence**	**75% Treatment Adherence**	**Full Treatment Adherence**
90% Relapse	$963,556	$836,151	$720,491
80% Relapse	$421,966	$339,249	$310,917
70% Relapse	$342,217	$314,085	$298,613
60% Relapse	$321,421	$305,736	$295,260

Pollack [88] assumes very high HIV prevalence (exceeding 65%) absent intervention. Such a model matches the observed prevalence among street IDUs in New Haven, Connecticut prior to implementation of syringe exchange. However, this analysis likely overstates HIV prevalence in later IDU cohorts, in which rates of needle-sharing have declined and from which the core group of IDUs at greatest risk may have exited the population through HIV infection.

The results in Table 14.2 also demonstrate the value of treatment adherence, β as a function of relapse rates from MMT. At high relapse rates, treatment adherence must also be high for cost-effective intervention. When relapse rates are low, MMT appears quite cost-effective even given low adherence to the intervention.

These results are also remarkable as an argument for the cost-effectiveness of even highly imperfect MMT interventions. In the baseline case, we posit that patients are 75% adherent to the treatment, and that fully 80% of MMT clients eventually relapse into injection drug use. None of the traditional (and large) social benefits associated with MMT – improved health status and productivity, and reduced criminal offending among MMT clients – are considered in this analysis. Yet the costs of MMT per averted infection are only $113,000.

This estimate is far below reasonable valuations of the social and individual costs of HIV infection. For example, Holtgrave and Pinkerton estimate present discounted lifetime treatment costs associated with HIV infection to be $195,000 [89]. More important is the impact of HIV prevention on individual well-being. Holtgrave and Pinkerton estimate that HIV infection is associated with a loss of 7.10 quality-adjusted life-years (QALYs). Across a wide range of public health interventions, interventions costing between $50,000 and $ 150,000 per QALY are widely regarded to be cost-effective by policymakers and the public [90]. By this cost-utility standard, MMT appears highly cost-effective in virtually all of our specifications when compared with other public health interventions.

As shown in Table 14.3, results are more discouraging for the prevention of HCV infection and other highly infectious agents. Within the same analytic framework, with all parameters identical except for a higher infectivity κ, MMT has only a small impact on HCV incidence and prevalence due to the higher probability of HCV transmission when needle sharing occurs. In most cases, costs per averted HCV infection are correspondingly much higher than those for HIV. Given modest estimates of lifetime expected treatment costs for acute and chronic HCV infection, it is difficult to justify MMT based on its role in HCV prevention [53, 91, 92].

Although these results are discouraging, they also indicate the great potential contribution of program quality to program effectiveness. Highly effective MMT programs – those with low relapse rates and high treatment adherence – can have a strong effect on HCV spread and can be cost-effective.

14.6 SHORT-TERM INCIDENCE ANALYSIS OF SEP

The full analytic framework for both SEP and MMT must be solved numerically, and is difficult to interpret from a qualitative perspective.

Table 14.3 Average cost per averted Hepatitis C infection (60% of IDUs in MMT): κ=0.03, varying rates of needle sharing and treatment adherence

30% of MMT Clients Share Needles	**50% Treatment Adherence (β=0.5)**	**75% Treatment Adherence (β=0.75)**	**Full Treatment Adherence (β=1.0)**
90% Relapse	$724,851	$580,067	$450,781
80% Relapse	$314,433	***$180,162***	$81,548
70% Relapse	$210,434	$118,877	$76,188
60% Relapse	$163,421	$95,055	$75,809
20% of MMT Clients Share Needles	**50% Treatment Adherence**	**75% Treatment Adherence**	**Full Treatment Adherence**
90% Relapse	$1,087,180	$870,020	$676,163
80% Relapse	$471,641	$270,239	$122,321
70% Relapse	$315,647	$178,314	$114,376
60% Relapse	$245,130	$142,582	$113,697
10% of MMT Clients Share Needles	**50% Treatment Adherence**	**75% Treatment Adherence**	**Full Treatment Adherence**
90% Relapse	$2,174,360	$1,740,102	$1,351,685
80% Relapse	$943,270	$540,473	$244,641
70% Relapse	$631,283	$356,626	$228,733
60% Relapse	$490,258	$285,163	$227,426

Fortunately, the short-run and steady-state implications of these models are tractable, and have been explored by several authors.

The most important set of models are due to Kaplan and collaborators, and include the noted “circulation model” of needle exchange [93]. The circulation model has been well-described elsewhere; its details will not be repeated here. Two features of that model, however, are noteworthy for this discussion.

First, the circulation model related specific data – observed HIV prevalence among needles returned to the New Haven SEP – to an underlying model of infectious disease transmission among IDUs. It therefore provided a more epidemiologically credible account of program effects than could be obtained through more traditional and less direct methodologies, such as studies that scrutinize self-reported risk behaviors among IDUs.

Second, the circulation model explores the impact of SEP on short-term HIV incidence among program clients. The model assumes that SEP has little impact on HIV *prevalence,* the number of IDUs affected by the intervention, or the exit rate of IDUs from the active drug-using population. Within this framework, SEP reduces immediate HIV incidence by removing infected needles from the population. This effectively reduces the rate of new infections by reducing the product $(\kappa\lambda)$ among active IDUs.

For simplicity, assume that there are no MMT slots: infectious disease spread can only be reduced by SEP. If SEP reduces incidence by some factor γ, the short-term incidence decline can be shown to be [53]

$$\Delta\iota_{short-term} = \gamma\theta[1 - \frac{1}{R_0}] \tag{12}$$

Using this type of model, Kaplan and Heimer estimated that the New Haven SEP reduced short-term HIV incidence by approximately one-third. If steady-state prevalence is approximately 65% and δ is approximately 1/(4,000 days), an SEP that serves a population of 300 IDUs will experience a short-term incidence decline of (1/3)(300)(1/4000)(0.65)=0.01625 infections per day, or approximately 5.9 averted HIV infections per year.

Although this appears to be a small program effect, SEP is an inexpensive intervention, costing approximately $5 per client per day. This simple short-term model therefore yields an estimate of $5*300/0.01625=$92,300 per averted infection. This is a highly cost-effective intervention.

Because a highly infectious agent such as HCV has a higher rate of new infections than HIV, this short-term incidence model yields slightly smaller estimated costs per averted infection for HCV than for HIV. Unfortunately, as shown below (Section 14.8), such findings can be misleading because they fail to account for long-term effects.

14.7 SHORT-TERM INCIDENCE ANALYSIS OF MMT

A similar short-term incidence model is readily derived for MMT. The short-term impact of MMT on infectious disease incidence can be considered to be the short-term reduction in the rate of new infections, assuming that infectious disease prevalence and the overall number of IDUs are stable. Expressed more formally, the short-term effect of a small addition to the number of MMT slots may be written as

$$\left[\frac{\partial\iota}{\partial M}\right]_{\pi,N} = \frac{-\beta\iota}{N-\beta M} \tag{13}$$

At the margin, one additional treatment slot will cost $C, so the marginal cost per averted infection is

$$\frac{C(N-\beta M)}{\beta\iota} \tag{14}$$

If one posits that M is close to 0, and applies the baseline model of SEP – steady-state HIV prevalence of 65% and $\beta=0.75$ – an MMT intervention that costs $14 per day yields an estimated cost per averted infection of $114,872.

14.8 STEADY-STATE CALCULATIONS

Explicit and tractable frameworks such as the circulation model brought new rigor to HIV prevention policy. However, the specific features, findings, and simplifying assumptions of such models, while appropriate for the HIV epidemic among IDUs, may prove misleading in other settings. HIV disease unfolds over a long period of time and is life-threatening. HIV is relatively difficult to transmit in any one exposure, such as a hospital needle-stick accident or the sharing of needles between infected and uninfected IDUs. When one alters these features, short-term incidence analysis may have important shortcomings.

One might assume that short-term incidence analysis understates the long-term value of prevention. If an intervention directly prevents 100 IDUs from being infected this year, the intervention also benefits the sexual and needle-sharing partners of these IDUs. Such “downstream” infections are not considered in short-term incidence models. This intuition is correct for prevention interventions such as polio vaccination that provide long-term protection. However, this intuition is false when prevention interventions provide imperfect or temporary protection to treated individuals. If steady-state prevalence is quite high, many of the original 100 IDUs will become infected in later periods. Because a prevention intervention merely delays infection for some treated individuals, short-term analysis of disease incidence can provide over-optimistic estimates of program effectiveness. In fact, ignoring “downstream” infections can either overstate or understate long-term program effects [54].

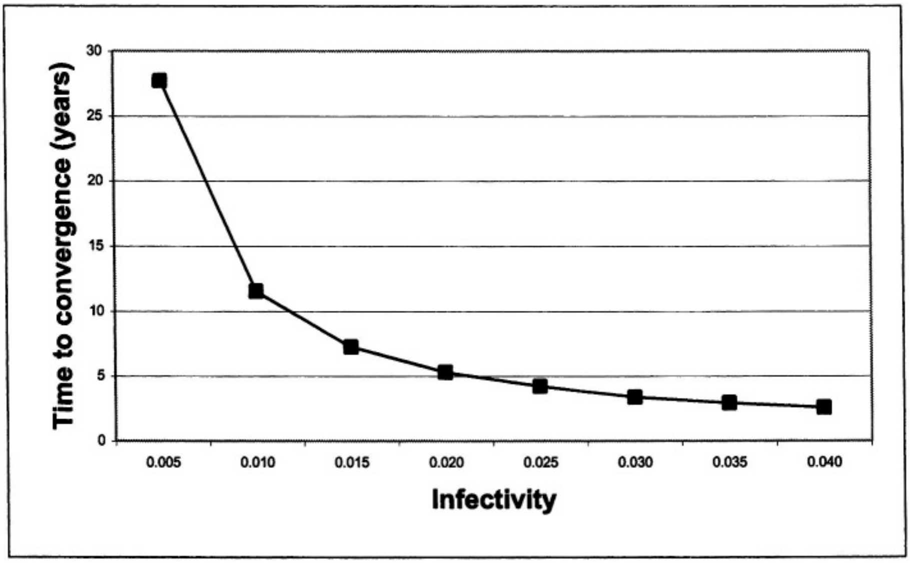
Steady-state analysis allows one to explore these claims, and to scrutinize the specific conditions under which short-term incidence analysis will overstate or understate long-term program effects [54]. The steady-state approach is especially suited to the analysis of rapid infectious disease transmission within a stable environment. As shown by Pollack [54], spread of a highly infectious agent such as HCV quickly approaches equilibrium incidence and prevalence. Such an analysis is less applicable to a less efficiently transmitted agent such as HIV, which displays much slower convergence to steady-state prevalence.

Figure 14.1, drawn from Pollack [54], provides more specific information. It is computed using the needle-sharing rates and mean drug-using careers shown in Table 14.1. The infectivity κ is allowed to vary across the empirically plausible range for both HIV and HCV. The figure displays the time required to move from 5% initial prevalence to 90% of steady-state prevalence across the empirically pertinent range of parameters. This framework overstates the time required to converge to steady state in actual policy settings, because HCV often reaches endemic levels before policy makers are able to intervene.

At low infectivities, the time required to reach steady state is substantial. For example, HIV policy analysts have used the value $\kappa=0.0036$ in published work. At this infectivity, numerical analysis indicates a convergence time of more than 30 years. Under these assumptions, steady-state analysis is less pertinent than short-term incidence analysis for public policy. Moreover, short-term analyses such as the circulation model yield results similar to those obtained through more elaborate dynamic models. Somewhat fortuitously, short-term incidence analysis for HIV provides an acceptable approximation of long-term effects.

Convergence times rapidly decline as infectivity increases. For example, if $\kappa=1.5\%$, convergence is reached in 7.25 years. When $\kappa=0.025$, convergence is reached within 4.2 years. For HCV and other highly infectious diseases, infectivity is even higher, making steady-state analysis most pertinent to evaluate medium-term and long-term effects. Such rapid convergence to steady state is also observed empirically, for example in the high rates of HCV incidence among young Baltimore IDUs [71].

Figure 14.1 Time to convergence in random-mixing models



For SEPs, one can explicitly calculate the steady-state impact. Steady-state incidence is

$$\iota^*_{SEP} = \theta\left[1-\frac{1}{R_0(1-\gamma)}\right] = \delta N_0\left[1-\frac{1}{(1-\gamma)R_0}\right] \qquad (15)$$

Manipulating equation (15) and assuming positive prevalence, the steady-state change in HCV incidence is given by

$$\Delta\iota_{long-term} = \frac{\gamma N_0 \delta}{R_0(1-\gamma)} = \frac{\gamma\theta}{R_0(1-\gamma)} \qquad (16)$$

Comparing the long-term and short-term changes in incidence, short-term analysis will overstate steady-state program effectiveness whenever $\Delta\iota_{long\text{-}term} > \Delta\iota_{short\text{-}term}$. This happens exactly when $R_0 > (2-\gamma)/(1-\gamma)$.

When $\gamma=1/3$, the break-even point occurs when $R_0=2.5$, or, equivalently, $\pi_0=0.60$. Equivalently, short-term incidence analysis will overstate steady-state program effectiveness whenever steady-state prevalence exceeds 60% in the absence of SEP. By the same logic, short-term incidence analysis will *understate* program effectiveness when steady-state prevalence is below 60% absent SEP.

If one provides SEP to all active IDUs, the average cost per averted infection is

$$AC=\frac{N_0 d}{\iota_0-\iota_{SEP}}=\frac{N_0 d}{\delta N_0[(1-\frac{1}{R_0})-(1-\frac{1}{(1-\gamma)R_0})]}=\frac{d(1-\gamma)R_0}{\delta\gamma} \qquad (17)$$

Setting d=\$5/day, γ=0.333, and δ=1/(4000 days), this implies that AC=40,000R_0. When R_0=2.5, SEP would prevent infections at an approximate cost of \$100,000 per averted infection.

Pollack compares short-term and steady-state models [53, 54]. Figure 14.2 shows these results. The y-axis indicates, in percentage terms, the amount that short-term analysis overstates (or understates) the steady-state impact of prevention interventions. At low steady-state prevalences, short-term incidence analysis understates long-term program effects. In such cases, averted secondary infections magnify the benefits of prevention interventions. At high steady-state prevalences, the opposite effects occur. Although incidence declines in the short-term, individuals who received short-term protection are likely to become infected later. Thus, the long-term impact of intervention is much smaller than one would predict based on short-term program effects.

One can conduct a similar steady-state analysis of MMT. In steady state, there will be N^* active drug users, with steady-state prevalence π^*. Every day, some θ uninfected individuals initiate drug use, while $(N^*\delta+M\mu)$ IDUs leave the population. So

$$N^*=\frac{\theta-M\mu}{\delta} \qquad (18)$$

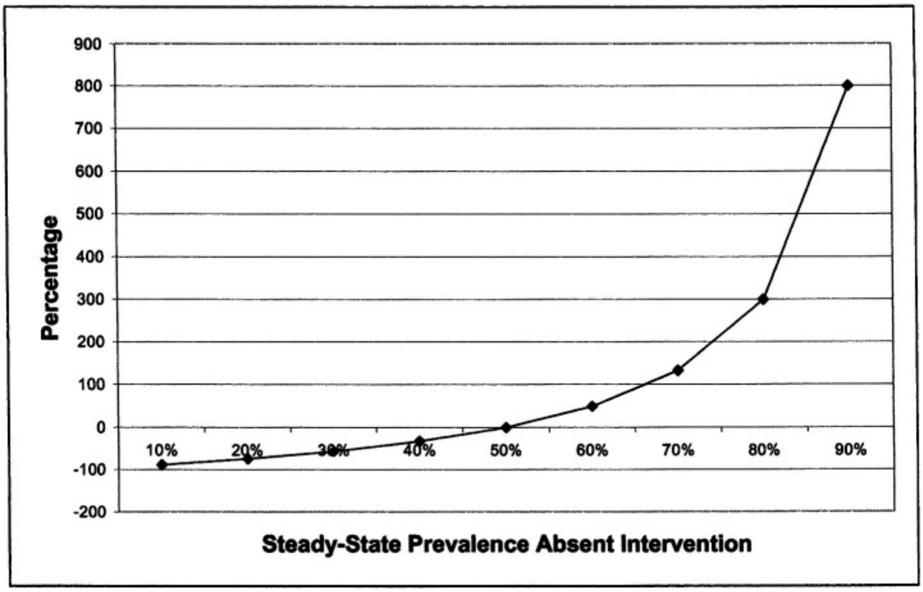
When steady-state prevalence is positive, one can show after algebra that

$$\pi^*=1-\frac{1}{R_0}\left(\frac{\theta}{\theta-M[\mu+\beta\delta]}\right) \qquad (19)$$

As before, the quantity $(\delta/\kappa\lambda)$ is the reciprocal of the reproductive rate of infection, or R_0.

The quantity $\theta/[\theta-M(\mu+\beta\delta)]$ reflects the reduction in disease prevalence attributable to treatment. The quantity $(\mu+\beta\delta)$ captures the effect of MMT on increasing exit from the drug-using population (μ), and also includes the

Figure 14.2 Bias in short-term incidence estimation for modest interventions (negative values indicate understatement of program effect)



effect of treatment on reducing needle sharing while individuals are in treatment $(\beta\delta)$.

One can show that steady-state disease incidence is given by

$$\iota^*(M) = \theta\pi^* = \theta\left[1 - \frac{1}{R_0}\left(\frac{\theta}{\theta - M[\mu+\beta\delta]}\right)\right] \tag{20}$$

Since treatment costs \$c per client per day, the total cost of drug treatment is \$Mc per day. At positive steady-state prevalence, average cost per averted infection is therefore*

$$AC = \frac{cM}{\iota_0 - \iota^*(M)} = \left(\frac{C}{\mu+\beta\delta}\right)R_0\left(1 - \frac{M[\mu+\beta\delta]}{\theta}\right) \tag{21}$$

Costs decline as exit rates of IDUs attributable to treatment intervention, $(\mu+\beta\delta)$, increase. Costs decline with the number of treatment slots (M), and

* If steady-state prevalence goes to zero, the average cost per averted infection is given by $AC=cMR_0/[\theta(R_0-1)]$.

depend on $(\mu+\beta\delta)/\theta$, the ratio of exits due to treatment over the arrival rate of uninfected people into the IDU population.

As with SEP, the cost per averted infection is proportional to R_0. Thus, measures that reduce steady-state prevalence can be significant, even if steady-state prevalence remains high. Suppose, for example, that steady-state prevalence π_0 is 90% prior to any intervention, and that the average cost associated with MMT per averted infection is $100,000. If, independent of MMT, one could reduce needle sharing rates or other risks to reduce π_0 to 85%, this small change in prevalence would reduce R_0 from a value of 10 to a value of 6.67. This apparently small prevalence decline corresponds to a one-third improvement in the cost-effectiveness of MMT.

When the number of treatment slots is extremely small compared to the population of IDUs, the average costs per averted infection is

$$AC = \left(\frac{cR_0}{\mu+\beta\delta}\right) \qquad (22)$$

If one applies the figures for HIV prevalence discussed above $(R_0=2.5)$, the average cost per HIV infection in steady-state is approximately $15,000 – a figure far below that obtained by short-term analysis. Because MMT reduces the overall size of the IDU population and reduces the steady-state prevalence of infection, short-term incidence analysis understates the value of MMT.

Note also that MMT has economies of scale. Steady-state prevalence, incidence, and the average cost per averted infection all decline as a larger fraction of active IDUs is served. Broad provision of MMT assists individual clients. It also generates a kind of herd immunity – creating beneficial spillovers to reduce prevalence among all IDUs [88].

Sometimes – but not always – broad provision of MMT can drive steady-state prevalence to zero. Given imperfect adherence, an epidemic can survive at positive steady-state prevalence even when all IDUs are enrolled in MMT. Setting $M=N^*$, it is possible to drive prevalence to zero exactly when $\mu > \delta[R_0(1-\beta)-1]$.

14.9 CONCLUSIONS AND FUTURE RESEARCH

Many insights for public policy can be drawn from the epidemics of substance abuse and HIV/AIDS. Operations researchers have provided many of these insights, and have the tools to critically scrutinize these insights when they are applied to new problems in new ways.

Operations researchers have provided data and methodologies that allow fair comparison of competing strategies to reduce illicit substance use. For HIV prevention, policy models allow policymakers to evaluate public investments in MMT and SEP by many of the same impact and cost-effectiveness standards as other public health measures. When such comparisons are made, HIV prevention interventions for IDUs compare favorably to prenatal care, car safety seats, and other widely accepted interventions [94].

Some lessons learned from HIV may not apply to other problems. Opponents of harm reduction argue that measures to make substance use safer are a foolish and ineffective response to the individual and social harms associated with injection drug use. According to this view, "use reduction" is essential to achieve lasting social benefit. The effectiveness and cost-effectiveness of SEP for HIV prevention provides a strong rebuttal of such use-reduction arguments. The need for use reduction appears more compelling when one considers more infectious agents such as HCV [28].

Public health challenges facing IDUs raise new challenges for both operations researchers and for policy.

The impact of high street purity on drug use behavior and drug treatment outcomes remains unknown. Many heroin users now consume the drug in non-injectable form. If such drug use is stable over time, non-injectable forms of heroin use may help to slow blood-borne epidemics among IDUs. Yet if non-injecting heroin users frequently transition to injection, the rise of heroin snorting and other behaviors may be a significant problem for both substance abuse policy and public health. In one study of Baltimore IDUs, only one-fourth of respondents had initiated heroin by injecting. Yet two-thirds of respondents reported some injection drug use. The most durable changes in route of heroin administration were towards high-risk behaviors [95].

The impact of improved HIV treatments raises more complex concerns for the design and operation of harm reduction and treatment interventions. Improved treatment lengthens life, may lengthen the period of high-risk behavior among IDUs, and may also reduce the probability of HIV transmission when there is needle sharing between infected and uninfected IDUs. The impact of such therapies has spawned a large literature in epidemiological policy modeling [87]. The spread of multi-drug-resistant strains has also attracted attention [96]. All of these developments heighten the importance of long-term prevention interventions for HIV-infected IDUs.

The histories of the substance abuse epidemic and the HIV epidemic are tragic in many ways. In the U.S., HIV occasioned late and inadequate policy responses to an epidemic afflicting IDUs and other stigmatized groups. This led to much avoidable mortality and morbidity among IDUs, their sexual partners, their children, and others [97, 98]. In the case of illicit drug use, policymakers continue to favor law-enforcement policies that are more punitive and less cost-effective than best-practice prevention or treatment interventions.

The most important reasons for these policy failures lay outside the immediate realm of policy analysis: they have arisen due to the quality of public management, moral and ideological choices, and interest-group politics that do not favor the groups at greatest risk. The best policy analysis is often powerless to overcome these factors. Sigmund Freud once commented that the voice of intellect is soft, but will not rest until it gains a hearing [99]. In this quiet but insistent way, operations researchers remind skeptical citizens and policymakers of the value of sound interventions that reduce premature death and avoidable suffering among IDUs.

Acknowledgements

The editors provided useful and patient guidance. Jonathan Caulkins, Edward Kaplan, Mark Kleiman, David Paltiel, and Peter Reuter provided many of the insights discussed in the chapter.

References

[1] Caulkins, J. and P. Reuter (1997). Setting goals for drug policy: harm reduction or use reduction? *Addiction,* 92, 1143-1150.

[2] MacCoun, R. and P. Reuter (2001). *Drug War Heresies: Learning from Other Vices, Times & Places.* Cambridge University Press, New York.

[3] Des Jarlais, D., S. Friedman, and T. Ward (1993). Harm reduction: a public health response to the AIDS epidemic among injecting drug users. *Annual Review of Public Health,* 14, 413-450.

[4] Gerstein, D.R. and H.J. Harwood (1990). *Treating Drug Problems.* National Academy Press, Institute of Medicine, Washington, DC.

[5] Des Jarlais, D., et al. (1995). Maintaining low HIV prevalence in populations of injecting drug users. *Journal of the American Medical Association,* 274, 1226-1231.

[6] Reuter, P., R. MacCoun, and P. Murphy (1990). *Moneyfrom Crime: A Study of the Economics of Drug Dealing in Washington, DC,* RAND, Santa Monica.

[7] Kleiman, M. (1993). *Against Excess: Drug Policyfor Results.* Basic Books, New York.

[8] Moore, M. (2001). *Toward a Balanced Drug-Prevention Strategy: A Conceptual Map,* in *Drug Addiction and Drug Policy,* P. Heyman and W. Brownsberger, Eds., Harvard University Press, Cambridge, MA.

[9] Saffer, H. and F. Chaloupka (1999). The demand for illicit drugs. *Economic Inquiry,* 37, 401-411.

[10] Reuter, P. (2001). The need for dynamic models of drug markets. *Bulletin on Narcotics,* LIII, 1-10.

[11] Caulkins, J. (1997). *Mandatory Minimum Drug Sentences: Throwing Away the Key or the Taxpayers' Money?* RAND Drug Policy Research Center., Santa Monica, CA.

[12] Burris, S., J. Welsh, M. Ng, M. Li, and A. Ditzler (2002). State syringe and drug possession laws potentially influencing safe syringe

disposal by injection drug users. *Journal of the American Pharmaceutical Association,* 42, S94-s98.

[13] Caulkins, J.P., C.P. Rydell, S.S. Everingham, J. Chiesa, and S. Bushway (forthcoming). *An Ounce of Prevention, a Pound of Uncertainty: The Cost-effectiveness of School-based Drug Prevention Programs.* RAND Drug Policy Research Center, Santa Monica, CA.

[14] Boyum, D. and M. Kleiman (1995). Alcohol and other drugs, in *Crime,* J.Q. Wilson and J. Petersilia, Eds., Institute for Contemporary Studies, San Francisco, CA.

[15] Caulkins, J.P. and P. Reuter (1998). What price data tell us about drug markets. *Journal of Drug Issues,* 28, 593-612.

[16] Caulkins, J. (1995). Domestic geographic variation in illicit drug prices. *Journal of Urban Economics,* 37, 38-56.

[17] Bach, P. and J. Lantos (1999). Methadone dosing, heroin affordability, and the severity of addiction. *American Journal of Public Health,* 89, 662-665.

[18] Becker, G. and K. Murphy (1988). A theory of rational addiction. *Journal of Political Economy,* 96, 675-700.

[19] Caulkins, J. (2001). The dynamic character of drug problems. *Bulletin on Narcotics,* LIII, 11-23.

[20] Rossi, C. (2001). A mover-stayer type model for epidemics of problematic drug use. *Bulletin on Narcotics,* LIII, 39-64.

[21] Heymann, P. and W. Brownsberger, Eds. (2001). *Drug Addiction and Drug Policy: The Struggle to Control Dependence.* Harvard University Press, Cambridge, MA.

[22] Hubbard, R.L. (1989). *Drug Abuse Treatment: A National Study of Effectiveness.* University of North Carolina Press, Chapel Hill, NC.

[23] Massing, M. (1998). *The Fix.* Simon and Schuster, New York, NY.

[24] D'Aunno, T. and H. Pollack (2002). Changes in methadone treatment practices: Results from a national panel study, 1988-2000. *Journal of the American Medical Association,* 288, 850-856.

[25] Ball, J.C. and A. Ross (1991). *The Effectiveness of Methadone Maintenance Treatment: Patients, Programs, Services, and Outcomes.* Springer-Verlag, New York, NY.

[26] Gunne, L. and L. Gronbladh (1981). The Swedish methadone maintenance program: a controlled study. *Drug and Alcohol Dependence,* 7, 249-256.

[27] Heimer, R. (1998). Can syringe exchange serve as a conduit to substance abuse treatment? *Journal of Substance Abuse Treatment,* 15, 183-191.

[28] Pollack, H. (2001). Controlling infectious diseases among injection drug users: Learning (the right) lessons from Acquired Autoimmunodeficiency Syndrome (AIDS). *Bulletin on Narcotics,* LIII, 91-104.

[29] Kaplan, E. and D. Soloschatz (1993). How many drug injectors are there in New Haven? Answers from AIDS data. *Mathematical and Computer Modelling,* 17, 109-115.

[30] Institute of Medicine (2000). *No Time to Lose: Making the Most of HIV Prevention.* National Academy Press, Washington, DC.

[31] Kaplan, E. and R. Heimer (1992). A model-based estimate of HIV infectivity via needle sharing. *Journal of AIDS,* 5, 1116-1118.

[32] Metzger, D., Navaline H, and W. GE (1998). Drug abuse treatment as AIDS prevention. *Public Health Reports,* 113, 97-106.

[33] Langendam, M., G. van Brussel, R. Coutinho, and E. van Ameijden (1999). Methadone maintenance treatment modalities in relation to incidence of HIV: results of the Amsterdam cohort study. *AIDS,* 13, 1711-1716.

[34] Langendam, M., G. van Brussel, R. Coutinho, and E. van Ameijden (2000). Methadone maintenance and cessation of injecting drug use: results from the Amsterdam Cohort Study. *Addiction,* 95, 591-600.

[35] Metzger, D., et al. (1993). Human immunodeficiency virus seroconversion among intravenous drug users in- and out-of-treatment: An 18-month prospective follow-up. *Journal of Acquired Immune Deficiency Syndromes and Human Retrovirology,* 6, 1049-1056.

[36] Battjes, R., R. Pickens, and L. Brown (1995). HIV infection and AIDS risk behaviors among injection drug users entering methadone treatment: An update. *Journal of Acquired Immune Deficiency Syndromes and Human Retrovirology,* 10, 90-96.

[37] McCoy, C., L. Metsch, D. Chitwood, P. Shapshak, and S. Comerford (1998). Parenteral transmission of HIV among injection drug users: Assessing the frequency of multiperson use of needles, syringes, cookers, cotton, and water. *Journal of Acquired Immune Deficiency Syndromes and Human Retrovirology,* 18, S25-S29.

[38] Alter, M. and L. Moyer (1998). The importance of preventing Hepatitis C virus infection among injection drug users in the United States. *Journal of Acquired Immune Deficiency Syndromes and Human Retrovirology,* 18, S6-S10.

[39] Short, L. and D. Bell (1993). Risk of occupational infection with blood-borne pathogens in operating and delivery room settings. *American Journal of Infection Control,* 21, 343-350.

[40] Pollack, H. and R. Heimer (forthcoming). Impact and cost-effectiveness of methadone maintenance treatment in preventing HIV and hepatitis C, in *Impact and Costs of Hepatitis C in Injecting Drug Users in the European Union,* EMCDDA, Ed., European Monitoring Centre for Drugs and Drug Addiction, Lisbon, Portugal.

[41] Selvey, L., M. Denton, and A. Plant (1997). Incidence and prevalence of hepatitis C among clients of a Brisbane methadone clinic: Factors influencing hepatitis C serostatus. *Australia and New Zealand Journal of Public Health,* 21, 102-104.

[42] Crofts, N., et al. (1994). Blood-borne virus infections among Australian injecting drug users: Implications for spread of HIV. *European Journal of Epidemiology,* 10, 687-694.

[43] Crofts, N., L. Nigro, K. Oman, E. Stevenson, and J. Sherman (1997). Methadone maintenance and Hepatitis C virus infection among injecting drug users. *Addiction,* 92, 999-1005.

[44] Goldberg, D., G. Allardice, J. McMenamin, and G. Codere (1998). HIV in Scotland: The challenge ahead. *Scottish Medical Journal,* 43, 168-172.

[45] Goldberg, D., S. Cameron, and J. McMenamin (2000). Hepatitis C virus antibody prevalence among injecting drug users in Glasgow has

fallen but remains high. *Community Diseases and Public Health,* 1, 95-97.

[46] Taylor, A., et al. (2000). Prevalence of hepatitis C virus infection among injecting drug users in Glasgow 1990-1996: Are current harm reduction strategies working? *Journal of Infection,* 40, 176-183.

[47] Broers, B., et al. (1998). Prevalence and incidence rate of HIV, hepatitis B and C among drug users on methadone maintenance treatment in Geneva between 1988 and 1995. *AIDS,* 12, 2059-2066.

[48] Hope, V., et al. (2001). Prevalence of hepatitis C among injection drug users in England and Wales: is harm reduction working? *American Journal of Public Health,* 91, 38-42.

[49] Van Ameijden, E., J. Van den Hoek, G. Mientjes, and R. Coutinho (1993). A longitudinal study on the incidence and transmission patterns of HIV, HBV and HCV infection among drug users in Amsterdam. *European Journal of Epidemiology,* 9, 255-262.

[50] Mansson, A., T. Moestrup, E. Nordenfelt, and A. Widell (2000). Continued transmission of hepatitis B and C viruses, but no transmission of human immunodeficiency virus among intravenous drug users participating in a syringe/needle exchange program. *Scandinavian Journal of Infectious Diseases,* 32, 253-258.

[51] Institute of Medicine (1995). *Preventing HIV Transmission: The Role of Sterile Needles and Bleach.* Institute of Medicine, National Academy Press, Washington, DC.

[52] Hagan, H., et al. (1999). Syringe exchange and risk of infection from hepatitis B and C viruses. *American Journal of Epidemiology,* 149, 203-213.

[53] Pollack, H. (2001). Cost-effectiveness of harm reduction in preventing hepatitis C among injection drug users. *Medical Decision Making,* 21, 357-367.

[54] Pollack, H. (2001). Ignoring 'downstream infection' in the evaluation of harm reduction interventions for injection drug users. *European Journal of Epidemiology,* 17, 391-395.

[55] Kaplan, E. (1995). A circulation theory of needle exchange. *Operations Research,* 43, 558-569.

[56] Kaplan, E., P. Cramton, and A. Paltiel (1989). Nonrandom mixing models of HIV transmission, in *Lecture Notes in Biomathematics,* C. Castillo-Chavez, Ed., Springer-Verlag, New York, NY.

[57] Zaric, G.S., M.L. Brandeau, and P.G. Barnett (2000). Methadone maintenance and HIV prevention: A cost-effectiveness analysis. *American Journal of Public Health,* 90, 1100-1111.

[58] Morris, M. (1997). Social networks and HIV. *AIDS,* 11, S209-s216.

[59] Kretzschmar, M. and L. Wiessing (1998). Modelling the spread of HIV in social networks of injecting drug users. *AIDS,* 12, 801-811.

[60] Caulkins, J., E. Kaplan, P. Lurie, P. O'Connor, and S.-H. Ahn (1998). Can difficult-to-reuse syringes slow the spread of HIV among injection drug users? *Interfaces,* 28, 23-33.

[61] Kaplan, E.H. and H.A. Pollack (1998). Allocating HIV prevention resources. *Socio-Economic Planning Sciences,* 32, 257-263.

[62] Altman, D., R. Greene, and H. Sapolsky (1981). *Health Planning and Regulation: The Decision-making Process.* AUPHA Press, Ann Arbor, MI.

[63] Downes, T. and T. Pogue (2002). How best to hand out money: Issues in the design and structure of intergovernmental aid formulas. *Journal of Official Statistics,* 18, 329-352.

[64] Brandeau, M. (2002). Difficult choices, urgent needs: Optimal investment in HIV prevention programs, in *Quantitative Evaluation of HIV Prevention Programs,* E. Kaplan and R. Brookmeyer, Eds., Yale University Press, New Haven, CT.

[65] Paltiel, A. and A. Stinnett (1998). Resource allocation and the funding of HIV prevention, in *Handbook of Economic Evaluation of HIV Prevention Programs,* D. Holtgrave, Ed., Plenum Press, New York, NY.

[66] Choi, K. and T. Coates (1994). Prevention of HIV infection. *AIDS,* 8, 1371-1389.

[67] Holtgrave, D.R., Ed. (1998). *Handbook of HIV Prevention Policy Analysis.* Plenum Press, New York, NY.

[68] Holtgrave, D.R. and J.A. Kelly (1996). Preventing HIV/AIDS among high-risk urban women: The cost-effectiveness of a behavioral group intervention. *American Journal of Public Health,* 86, 1442-1445.

[69] Zaslavsky, A. and A. Schirm (2002). Interactions between survey estimates and federal funding formulas. *Journal of Official Statistics,* 18, 371-393.

[70] Watts, D. (1999). *Small Worlds: The Dynamics of Networks Between Order and Randomness.* Princeton University Press, Princeton, NJ.

[71] Garfein, R., et al. (1998). Prevalence and incidence of Hepatitis C virus infection among young injection drug users. *Journal of Acquired Immune Deficiency Syndromes and Human Retrovirology,* 18, S11-S19.

[72] Holmberg, S.D. (1996). The estimated prevalence and incidence of HIV in 96 large U.S. metropolitan areas. *American Journal ofPublic Health,* 86, 642-654.

[73] Coutinho, R. (1998). HIV and Hepatitis C among injecting drug users: Success in preventing HIV has not been mirrored for Hepatitis C. *British Medical Journal,* 317, 424-425.

[74] Pollack, H. (2001). Can we protect drug users from Hepatitis C? *Journal of Policy Analysis and Management,* 20, 358-364.

[75] Zaric, G.S., P.G. Barnett, and M.L. Brandeau (2000). HIV transmission and the cost-effectiveness of methadone maintenance. *Management Science,* 46, 1013-1031.

[76] Anderson, R.M. and R.M. May (1991). *Infectious Diseases of Humans.* Oxford University Press, Oxford.

[77] Kaplan, E.H. (1989). Needles that kill: Modeling human Immunodeficiency virus transmission via shared needle injection equipment in shooting galleries. *Reviews of Infectious Diseases,* 11, 289-298.

[78] Vlahov, D., et al. (1995). Incidence and risk factors for human T-lymphotropic virus type II seroconversion among injecting drug users in Baltimore, Maryland, U.S.A. *Journal of Acquired Immune Deficiency Syndrome and Human Retrovirology,* 9, 89-96.

[79] Stark, K., R. Muller, U. Bienzle, and I. Guggenmoos-Holzmann (1996). Frontloading: a risk factor for HIV and hepatitis C virus infection among injecting drug users in Berlin. *AIDS,* 10, 311-317.

[80] Shah, S., et al. (1996). Detection of HIV-1 DNA in needle/syringes, paraphernalia, and washes from shooting galleries in Miami: A preliminary laboratory report. *Journal of Acquired Immune Deficiency Syndromes,* 11, 301-306.

[81] Koester, S., R. Booth, and W. Wiebel (1990). The risk of HIV transmission from sharing water, drug mixing containers and cotton filters among intravenous drug users. *International Journal of Drug Policy,* 1, 28-30.

[82] Koester, S., R. Booth, and E. Zhang (1996). The prevalence of additional injection-relation HIV risk behaviors among injection drug users. *Journal of Acquired Immunodeficiency Syndrome and Human Retrovirology,* 12, 202-207.

[83] Hagan, H., et al. (2001). Sharing of drug preparation equipment as a risk factor for hepatitis C. *American Journal of Public Health,* 91, 42-46.

[84] Grund, J., C. Kaplan, N. Adriaans, and P. Blanken (1991). Drug sharing and HIV transmission risks: The practice of frontloading in the Dutch injecting drug user population. *Journal of Psychoactive Drugs,*23, 1-10.

[85] Kaplan, E. and Y. Lee (1990). How bad can it get? Bounding worst case endemic heterogeneous mixing models of HIV/AIDS. *Mathematical Biosciences,* 99, 157-180.

[86] Des Jarlais, D., et al. (1999). Audio-computer interviewing to measure risk behaviour for HIV among injecting drug users: A quasi-randomised trial. *Lancet,* 353, 1657-1661.

[87] Velasco-Hernandez, J., H. Gershengorn, and S. Blower (2002). Could widespread use of combination antiretroviral therapy eradicate HIV epidemics? *Lancet Infectious Diseases,* 2, 487-493.

[88] Pollack, H. (2002). Methadone treatment as HIV prevention: Cost-effectiveness analysis, in *Quantitative Evaluation of HIV Prevention Programs,* E.H. Kaplan and R. Brookmeyer, Eds., Yale University Press, New Haven, CT.

[89] Holtgrave, D. and S. Pinkerton (1997). Updates of cost of illness and quality of life estimates for use in economic evaluations of HIV prevention programs. *Journal of Acquired Immune Deficiency Syndrome and Human Retrovirology,* 16, 54-62.

[90] Hirth, R., M. Chernew, E. Miller, A. Fendrick, and W. Weissert (2000). Willingness to pay for a quality-adjusted life year: In search of a standard. *Medical Decision Making,* 20, 332-342.

[91] Brown, K. and N. Crofts (1998). Health care costs of a continuing epidemic of Hepatitis C virus infection among injecting drug users. *Australian and New Zealand Journal of Public Health,* 22, 384-388.

[92] Wong, J.B., W.G. Bennett, R.S. Koff, and S.G. Pauker (1998). Pretreatment evaluation of chronic Hepatitis C: Risks, benefits, and costs. *Journal of the American Medical Association,* 280, 2088-2093.

[93] Kaplan, E. and R. Heimer (1994). A circulation theory of needle exchange. *AIDS,* 8, 567-574.

[94] Tengs, T., et al. (1995). Five-hundred life-saving interventions and their cost-effectiveness. *Risk Analysis,* 15, 369-390.

[95] Strang, J., P. Griffiths, B. Powis, J. Abbey, and M. Gossop (1997). How constant is an individual's route of heroin administration? Data from treatment and non-treatment samples. *Drug and Alcohol Dependence,* 46, 115-118.

[96] Blower, S., A. Aschenbach, and J. Kahn (2003). Predicting the transmission of drug-resistant HIV: Comparing theory with data. *Lancet Infectious Diseases,* 3, 10-11.

[97] Cohen, C. (1999). *Boundaries of Blackness: AIDS and the Breakdown of Black Politics.* University of Chicago Press, Chicago, IL.

[98] Lurie, P. and E. Drucker (1997). An opportunity lost: HIV infections associated with lack of a national needle-exchange programme in the USA. *Lancet,* 349, 604-608.

[99] Freud, S. (1928). *Future of an Illusion.* W. W. Norton, New York, NY.

15 THE COST EFFECTIVENESS OF PARTIALLY EFFECTIVE HIV VACCINES

Douglas K. Owens[1,2], Donna M. Edwards[3], John F. Cavallaro[4] and Ross D. Shachter[4]

[1] VA Palo Alto Health Care System
Palo Alto, CA 94303

[2] Center for Primary Care and Outcomes Research
Stanford University School of Medicine
Stanford, CA 94305

[3] Sandia National Laboratories
Livermore, CA 94550

[4] Department of Management Science and Engineering
Stanford University
Stanford, CA 94305

SUMMARY

Development of a vaccine remains the best hope for curtailing the worldwide pandemic caused by human immunodeficiency virus (HIV) infection. Due to the complex biology of HIV infection, there is increasing concern that an HIV vaccine may provide incomplete protection from infection. In addition to reducing susceptibility to disease, an HIV vaccine may also prolong life in people who acquire HIV despite vaccination, and may reduce HIV transmission. We evaluated how varying degrees of vaccine efficacy for susceptibility, progression of disease, and infectivity influence the costs and benefits of a vaccine program in a population of men who have sex with men, We found that the health benefits, and thus cost effectiveness, of HIV vaccines were strikingly dependent on each of the types of vaccine efficacy. We also found that vaccines with even modest efficacy provided substantial health benefits and were cost effective or cost saving. Although development of an HIV vaccine has been extremely difficult, even a partially effective HIV vaccine could dramatically change the course of the HIV epidemic.

KEY WORDS

HIV, Vaccines, Cost-effectiveness analysis

15.1 INTRODUCTION

At the end of 2002, 42 million people were living with human immunodeficiency virus (HIV) infection. New infections were occurring at about 14,000 per day [1]. By 2010, an additional 45 million people will become infected with HIV if current trends continue [1]. Highly active antiretroviral therapy is very effective, but it is unavailable in most low-income countries where 95% of the HIV infections occur. Development of an HIV vaccine remains the best hope for curtailing the worldwide pandemic.

Despite intensive effort, development of an HIV vaccine has remained elusive. Many candidate vaccines have undergone clinical trials, but only one vaccine, AIDSVAX, has undergone large-scale, Phase III efficacy trials that are required for vaccine licensing. Preliminary results of the first AIDSVAX trial, reported in early 2003, indicated that the vaccine failed to reduce HIV infection rates in the overall group of vaccine recipients. In subgroup analyses, the manufacturer reported that the vaccine reduced HIV infection rates by 67% in non-Hispanic minorities, and by 78% among black recipients. The subgroup analyses were highly controversial because of small sample sizes. Even if these results become accepted, however, they would further confirm the belief among many experts that if a vaccine becomes available, it would likely provide only partial protection from HIV. This view led the Centers for Disease Control and Prevention and the World Health Organization to hold consultations to examine how partially effective HIV vaccines should be used [2].

The increasing concern that an HIV vaccine would be only partially effective has led to considerable interest in how to model vaccine efficacy (VE) for HIV vaccines. Vaccines for HIV may act to reduce the burden of disease in three ways. First, the vaccine may reduce *susceptibility* to disease, as do most familiar vaccines. In a framework developed by Longini and colleagues [3-5], this component of vaccine efficacy is termed the vaccine efficacy for susceptibility, $\mathbf{VE_S}$. Because the $\mathbf{VE_S}$ is likely to be less than 100% (because of incomplete protection), a person who has been vaccinated may subsequently become infected with HIV. Unlike some traditional vaccines, an HIV vaccine may also ameliorate disease in those who become infected. The vaccine would likely work by improving the ability of the immune system to suppress HIV viral replication. This suppression would lead to the two additional means by which the vaccine could reduce the disease burden from HIV infection: the vaccine could slow progression of HIV disease and decrease the likelihood of transmission of HIV. Transmission would likely decrease because the probability of transmission is related to the level of virus in the blood: transmission occurs less readily if the level of virus in blood (and other body fluids) is low [6, 7]. Thus, an HIV vaccine may have efficacy to reduce *progression* of disease, VE_p, and efficacy to reduce *infectivity* VE_i [3-5], in addition to efficacy to prevent infection (VE_s).

Should a partially effective vaccine be used? How good must a vaccine be before public health officials recommend its use? These questions are complex, in part because a vaccine with low VE_s might still have substantial health benefit if either VE_p or VE_i were high. In addition, because a vaccine program would require substantial resources, the question of whether to use the vaccine also depends on the costs of the program. To address these questions, we developed a dynamic transmission model to represent the effects of a vaccine in a population, and an economic model to assess the costs associated with the vaccine program [8-10]. We modeled two types of vaccines: a preventive vaccine ($VE_s > 0$, $VE_p = 0$, $VE_i = 0$) that would be given to uninfected people, and a therapeutic vaccine ($VE_s = 0$, $VE_p \geq 0$, $VE_i \geq 0$) that would be given to people known to have HIV. By evaluating both types of vaccines, we can understand how VE_s, VE_p, and VE_i influence both the health benefits and costs of a vaccine program. We evaluated the costs and benefits of these vaccine programs in a population of men who have sex with men (MSM) designed to reflect the population in San Francisco, California.

This chapter builds on previous work we have done in evaluating potential HIV vaccines [8-10]. We have recast our previous work into a framework for analysis of vaccines that has recently developed. This framework conceptualizes vaccine efficacy in terms of efficacy for susceptibility, for progression, and for infectivity. In addition, the work in this chapter assumes no behavior change (positive or negative) in the base case. Arguments have been made about why risk behavior might increase or decrease with a vaccine program, but recent evidence has not supported the more pessimistic assumption that we used in earlier work that risk behavior would increase. Additionally, we updated costs to reflect 2003 dollars.

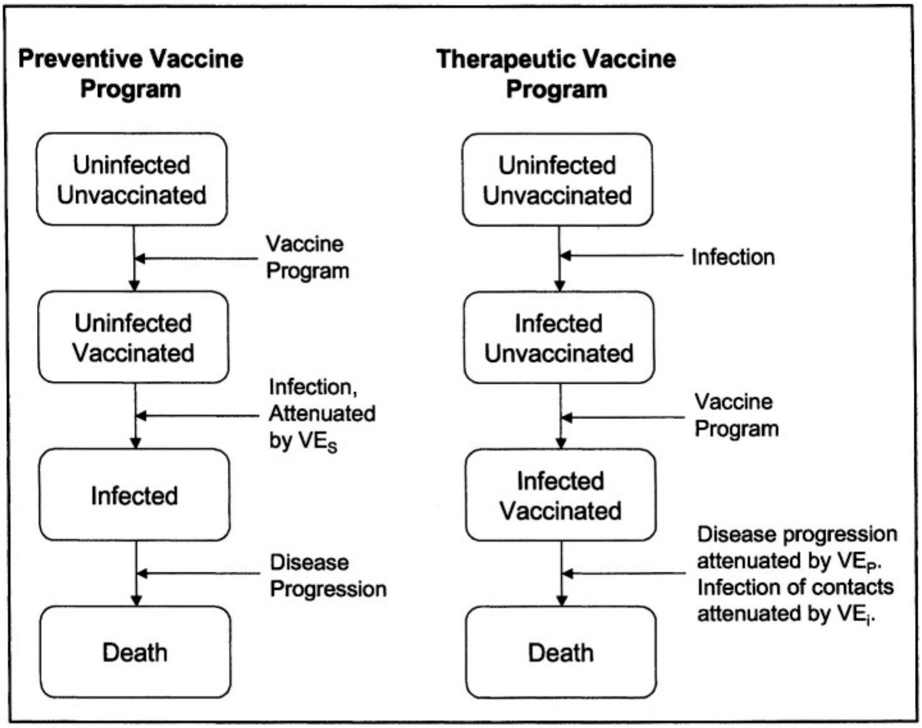
15.2 METHODS

15.2.1 Model and data

Details about the model structure, input data, and validation are available elsewhere [8-10], so we provide an abbreviated overview here. A schematic depiction of the model is shown in Figure 15.1. The diagram in Figure 15.1 is substantially simplified but indicates the important relationships captured in the model. The figure shows the vaccinated cohort for both preventive and therapeutic vaccines. For a preventive vaccine, infection is attenuated.

For a therapeutic vaccine, progression of disease is attenuated. Infectivity may also be reduced by a therapeutic vaccine.

Figure 15.1 Schematic of model structure



The model simulates both vaccinated and unvaccinated cohorts that evolve over time. The transitions between the compartments in the model (boxes in Figure 15.1) were determined by deterministic differential equations [8-10]. The simulation determined the number of people in each compartment after a specified time interval. Health outcomes were measured in terms of number of HIV infections in the population, the prevalence of HIV, and the total quality-adjusted life years (QALYs) lived in the population. A QALY is a measure of length of life adjusted for changes in quality of life [11]; QALYs are a standard outcome for cost-effectiveness analyses. We calculated QALYs by multiplying the time in a health state by a quality adjustment that ranges from 0 to 1. For example, if the quality of life with asymptomatic HIV infection is 0.83, then a year spent with asymptomatic HIV infection is equal to 1 × 0.83 = 0.83 QALYs. The model estimated the health outcomes and costs for the vaccinated and unvaccinated cohorts. We assessed the efficacy of the vaccine program in terms of infections averted, changes in prevalence, or changes in the QALYs for the vaccinated cohort relative to the unvaccinated cohort. As recommended by guidelines for the conduct of cost-effectiveness analyses [12], we discounted both health and economic outcomes. We discounted outcomes at 5%; expenditures are expressed in 2003 dollars.

We modeled the effects of vaccines and progression of disease as changes in the transition rate from one model compartment to another (Figure 15.1). For a preventive vaccine program, the rate of infection in individuals in the vaccinated cohort is attenuated by VE_s, which we defined as the proportion of vaccine recipients who are protected from infection. For a therapeutic vaccine, we assumed for simplicity that the efficacy of the vaccine in reducing progression of disease (that is, in prolonging life), VE_p, occurred via prolongation of life during the asymptomatic phase of infection. We varied the degree of this increase from one year to ten years. We also modeled changes in infectivity of vaccine recipients (VE_i) as reductions in transmission to contacts. We modeled disease progression (for both types of vaccine programs) from asymptomatic disease, to symptomatic disease, and then to AIDS (not shown in Figure 15.1). In addition, we modeled interactions of the vaccinated cohort with the uninfected people in the population. In the analyses we report here, we assumed that vaccine recipients would not change risky behavior. We have evaluated the importance of behavior change previously [8-10].

We calculated the incremental cost effectiveness of the vaccine program as the difference in costs between the vaccinated and unvaccinated cohorts, divided by the difference in health benefit. For example, to estimate the cost effectiveness in dollars per QALY gained, we calculated the cost-effectiveness ratio as:

$$(\$_{\text{vaccinated}} - \$_{\text{unvaccinated}}) / (\text{QALYs}_{\text{vaccinated}} - \text{QALYs}_{\text{unvaccinated}})$$

If the vaccine increased both costs and health benefit, we calculated the cost-effectiveness ratio. If the vaccine provided benefits while reducing costs, we said that vaccination dominated the strategy of no vaccination.

15.2.2 Model inputs

We estimated inputs for the model from published and unpublished data about the population of MSM in San Francisco [8-10]. Key input data for the model are shown in Table 15.1.

Table 15.1 Input Variables

Variable	Base-Case Value (range)
Preventive Vaccine	
Efficacy for susceptibility, VE_s	10% -90%
Efficacy for progression of disease, VE_p	0%
Efficacy for infectivity, VE_i	0%
Duration, years	5 to 50 years
Proportion of population vaccinated	75% (10%-100%)
Per-person cost	$1,000
Therapeutic Vaccine	
Efficacy for susceptibility, VE_s	0%
Efficacy for progression of disease, VE_p	1-10 years of additional life
Efficacy for infectivity, VE_i	0%-90% reduction in infectivity
Proportion of population vaccinated	75% (10%-100%)
Per-person cost	$1,000
Population Parameters	
Initial size	55,800
Prevalence, late-stage epidemic	49%
Mean age, years	30

Sources and detailed input data are provided elsewhere [9-11]. We estimated transmission probabilities from epidemiologic studies and model-based estimates. We evaluated vaccine programs in two types of epidemics, an early-stage epidemic and a late-stage epidemic. In an early-stage epidemic, the prevalence of HIV infection is relatively low (10%) but increasing. Such an epidemic may reflect younger MSM who have higher levels of risky behavior and higher number of annual partnerships. In a late-stage epidemic, the prevalence is relatively high (approximately 50%) and is decreasing. Such an epidemic may reflect older MSM who have lower levels of risky behavior and fewer partnerships. We report here results for the late-stage epidemic; we evaluated early-stage epidemics elsewhere [8-10]. Because we estimated the parameters for the model in the era prior to the advent of highly active antiretroviral therapy, the model underestimates length of life for patients who receive treatment under current regimens. We discuss the implications of newer therapies on our results in Section 15.4.

The characteristics of HIV vaccines are, of course, unknown. Therefore, we evaluated vaccines with many plausible combinations of efficacy and duration of action. We arbitrarily assumed that a vaccine would cost $1,000, and varied this value widely.

15.3 RESULTS

15.3.1 Preventive vaccine

We evaluated vaccine efficacy for susceptibility **(VE_s)** that ranged from 10% protection to 90%, with vaccine durations of 5, 10, and 50 years (Figures 15.2 and 15.3). Figure 15.2 shows the health and economic outcomes for vaccines with varying efficacy and duration. The figure indicates the net increase in QALYs and expenditures (or savings) in millions of dollars ($M) for a preventive vaccine after 150 years, assuming no change in risk behaviors, in a late-stage epidemic, with 75% of the population vaccinated. Each point on the polygon represents a preventive vaccine with different efficacy and duration. Squares on the top line of the polygon represent preventive vaccines with efficacy of 10% to 90% and a duration of 5 years. Squares on the bottom line represent a vaccine with a duration of 50 years.

The dotted lines indicate cost-effectiveness thresholds of $50,000 and $10,000 per QALY gained. A vaccine represented by a point on the polygon that falls between the two dotted lines has a cost-effectiveness ratio between $50,000 and $10,000 per QALY gained. A vaccine represented by a point to the right of the $10,000 per QALY line cost less than $10,000 per QALY gained. Points on the polygon below the horizontal axis represent vaccines that reduce net expenditures, and therefore dominate the no-vaccination strategy. A vaccine with any combination of efficacy and duration within the ranges noted will fall within the polygon in Figure 15.2. Points to the right of these lines cost less than the threshold rate per QALY gained.

Our analyses indicate that vaccines need not be highly effective to have substantial health benefit with reasonable expenditures (Figure 15.2). For example, a vaccine with only 10% efficacy and duration of 5 years (the top left point on the polygon), resulted in expenditures of about $83 million dollars and a net increase of about 3,600 QALYs at a cost of less than $50,000 per QALY gained (Figure 15.2). A vaccine with an efficacy of 90% and duration of 50 years (that is, lifelong protection, represented by the

Figure 15.2 Long-term outcomes of a preventive vaccine

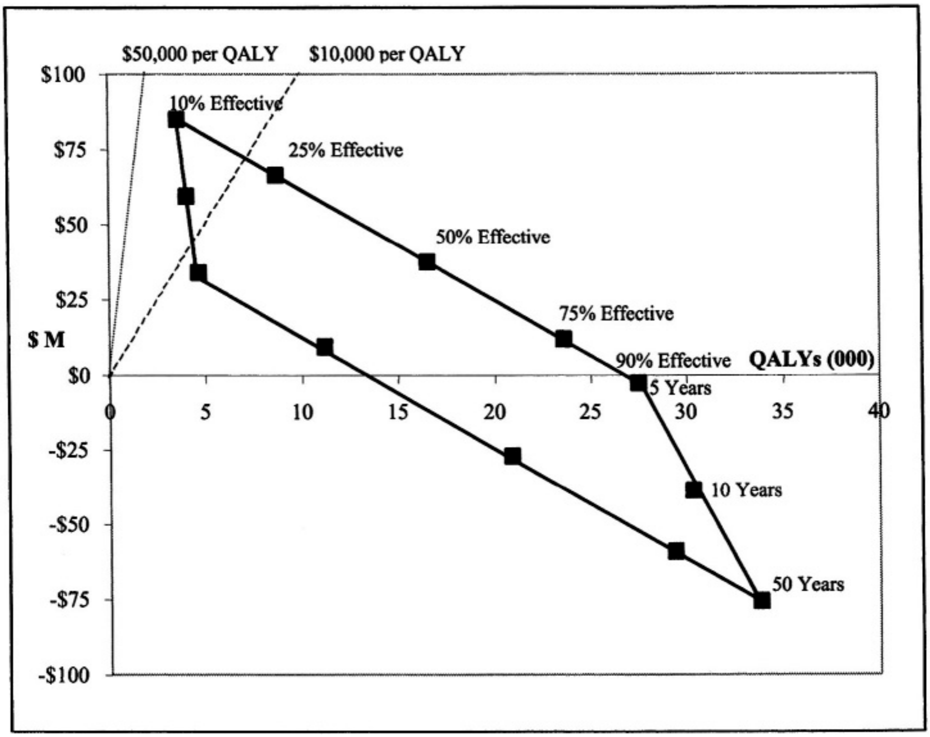
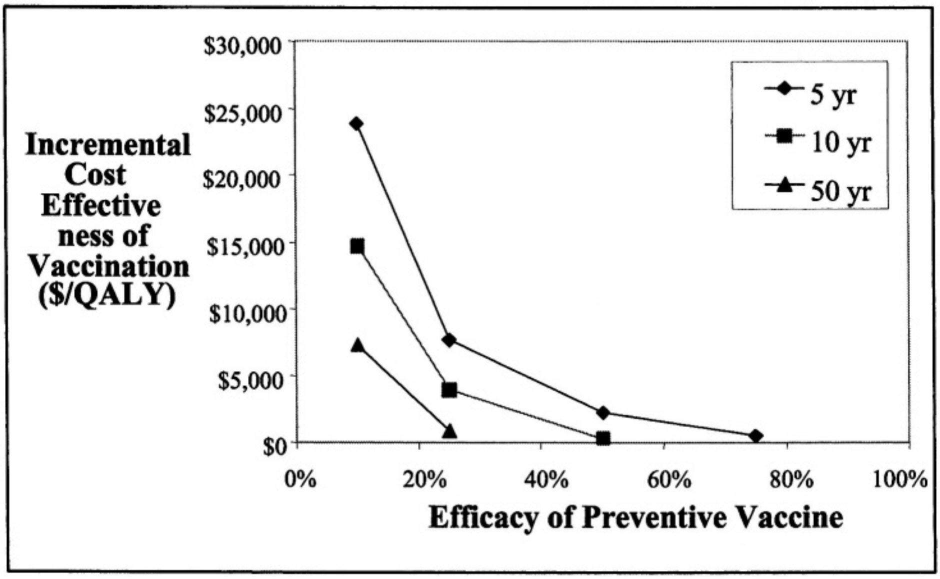


Figure 15.3 Effect of vaccine efficacy for susceptibility on the cost effectiveness of a preventive vaccine



rightmost point on the graph) reduced expenditures by approximately $75 million and increased QALYs in the vaccinated cohort by approximately 34,000. Thus, a vaccine with low efficacy was cost effective by the standards of high-income countries, and highly effective vaccines dominated the no-vaccination strategy by reducing expenditures while providing very large health benefit. More effective vaccines reduced total expenditures because the cost of the vaccine program was outweighed by the savings associated with prevention of HIV infection.

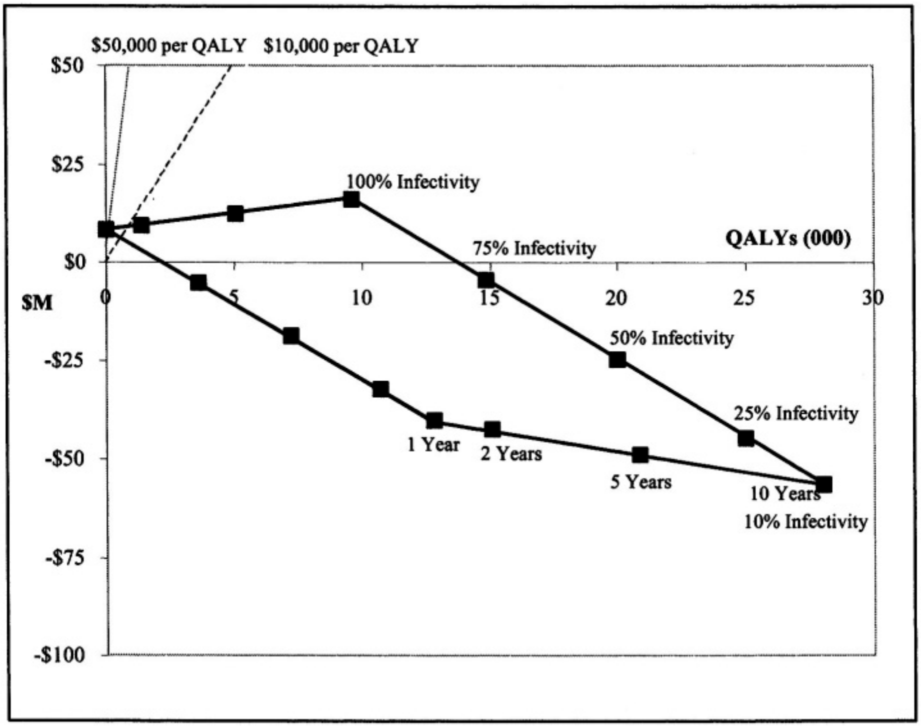
In Figure 15.3, we indicate more directly how changes in VE_s influenced the cost-effectiveness ratio. The figure shows the incremental cost effectiveness of a vaccine program relative to the no-vaccination strategy in dollars per QALY gained. Vaccines with duration of 5, 10, and 50 years are shown by different lines. The values represent long-term outcomes, with no behavior change, and 75% of the population vaccinated. With a vaccine efficacy of 10%, cost effectiveness varied from about $7,000 per QALY gained (duration of 50 years) to approximately $24,000 per QALY gained (duration of 5 years). As vaccine efficacy increased, the vaccine program became increasingly cost effective. When the lines in Figure 15.3 reach the x-axis, it indicates that at higher efficacy, the vaccine strategy dominated the no-vaccination strategy. Vaccination dominated the no-vaccination strategy when efficacy reached approximately 35%, 55%, and 90% for a vaccine with duration of 50 years, 10 years, and 5 years respectively.

15.3.2 Therapeutic vaccine

For therapeutic vaccines ($VE_s = 0$, $VE_p \geq 0$, $VE_i \geq 0$), we evaluated vaccines that prolonged life by 1, 2, 5, and 10 years, and that decreased infectivity by 0%, 25%, 75% and 90% (Figure 15.4). The figure indicates the net increase in QALYs and expenditures (or savings) in millions of dollars ($M) for a therapeutic vaccine after 150 years, assuming no change in risk behaviors, in a late-stage epidemic, with 75% of the population vaccinated. Each point on the polygon represents a therapeutic vaccine with different efficacy for prolongation of life (1, 2, 5, and 10 years) and infectivity. The square labeled 100% infectivity represents a vaccine that does not reduce infectivity; the square labeled 10% infectivity represents a vaccine that reduces infectivity by 90%. As in Figure 15.2, the dotted lines indicate cost-effectiveness thresholds of $50,000 and $10,000 per QALY gained. A vaccine represented by a point on the polygon that falls between the two dotted lines has a cost-effectiveness ratio between $50,000 and $10,000 per QALY gained. A vaccine represented by a point to the right of the $10,000 per QALY line cost less than $10,000 per QALY gained. Points on the polygon below the horizontal axis represent vaccines that reduce net expenditures, and therefore dominate the no-vaccination strategy.

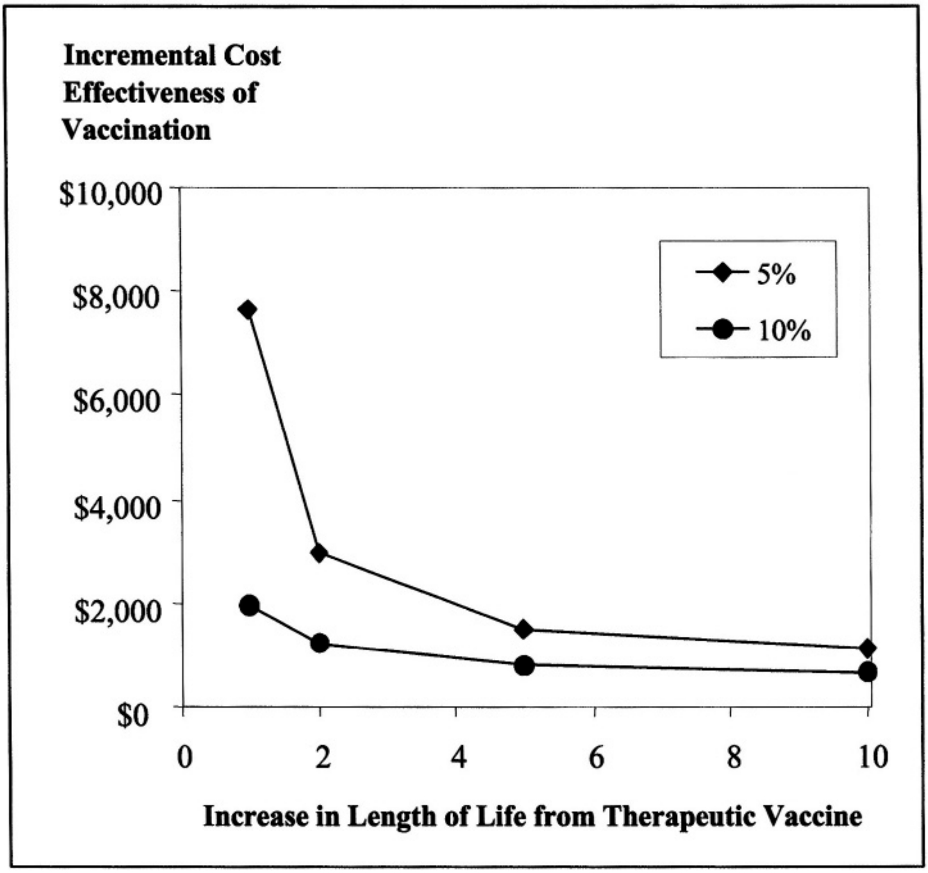
A therapeutic vaccine that extended life by two years cost less than $10,000 per QALY gained, even if it provided no reduction in infectivity (Figure 15.4). In contrast, a vaccine that reduced infectivity by 90% and increased length of life by 10 years (the rightmost point in Figure 15.4) resulted in large cost savings and a gain of about 28,000 QALYs in the vaccinated cohort. Thus, both the degree to which the vaccine prolonged life, VE_p, and the degree to which it reduced infectivity, VE_i, had a large influence on costs, benefits, and the cost effectiveness of the vaccine program.

Figure 15.4 Long-term outcomes for a therapeutic vaccine



The influence of VE_p on cost effectiveness is shown in Figure 15.5 for a therapeutic vaccine that reduced infectivity (VE_i) by 5% or 10%. The figure shows the incremental cost effectiveness of a vaccine program relative to the no-vaccination strategy in dollars per QALY gained. Vaccines that reduce infectivity by 5% and 10% are shown. The values represent long-term outcomes, with no behavior change, and 75% of the population vaccinated. A vaccine program resulted in expenditures of $8,000 per QALY gained if the vaccine only increased length of life by one year and reduced infectivity by 5% (Figure 15.5). Relatively small changes in infectivity had substantial influence on the cost effectiveness of a vaccine program (Figure 15.5). Vaccines that reduced infectivity by more than about 15% dominated the no-vaccination strategy, and are therefore not shown in Figure 15.5.

Figure 15.5 Cost effectiveness of a therapeutic vaccine



15.4 DISCUSSION

An HIV vaccine may reduce susceptibility to disease, may prolong life in people who acquire HIV despite vaccination, and may reduce HIV transmission. A vaccine could have only one of these mechanisms of protection (for example, reducing susceptibility), or it could have all three.

We evaluated preventive vaccines that provided protection from infection but did not prolong life or reduce transmission, and therapeutic vaccines that could both prolong life and reduce transmission, but provided no protection from infection. We evaluated these types of vaccines because of ongoing research to develop such vaccines. Although a vaccine could have all three mechanisms of protection, we can learn much about how the mechanism of protection influences costs and health benefits by assessing preventive and therapeutic vaccine programs independently.

The first major finding of our analysis is that each of the types of efficacy, VE_s, VE_p, and VE_i is extremely important for HIV vaccines. The health benefits, and thus cost effectiveness, of HIV vaccines were strikingly dependent on each of the types of vaccine efficacy. Our analyses indicated that a preventive vaccine that provided only protection from infection was cost effective or cost saving under many plausible scenarios. Likewise, a therapeutic vaccine that provided no protection from infection was also cost effective or cost saving under most scenarios.

An important implication of this finding is that an understanding of the effect of a vaccine in a population depends on assessing all three types of vaccine efficacy [3-5]. Longini and colleagues have discussed design of vaccine trials and how to augment trials so that all parameters of vaccine efficacy are assessed [3-5]. For a therapeutic vaccine, a trial that assessed only VE_p would leave substantial uncertainty about the health benefit of a vaccine. As noted in Figure 15.4, for a given prolongation of life (VE_p), the health benefit from the vaccine varies by a factor of three or more based on variation in VE_i. These considerations have led to the addition of sexual partner studies to vaccine trials to help assess changes in infectivity. Because infectivity correlates with the level of virus in blood, another strategy is to assess HIV viral load as a surrogate measure for infectivity. Direct assessment of transmission is preferable, however, if feasible.

A second major finding of our analysis is that even vaccines with modest efficacy provided substantial health benefits and were cost effective or cost saving. By traditional standards, a vaccine that prevented infection in only 25% to 50% of recipients would be considered a failure. In contrast, an HIV vaccine with these characteristics provided large health and economic benefit. In part, HIV vaccines need not meet high standards of efficacy because the mortality from HIV is so high. In addition, our analyses assumed that vaccine recipients did not increase their risky behavior; we previously have shown that increased risk behavior reduces the benefit of a vaccine program. As improvements in therapy reduce HIV mortality, or if studies show that vaccine recipients increase risky behavior, the efficacy of vaccines may need to increase to provide similar cost effectiveness. Nonetheless, our analysis indicates that vaccines of modest efficacy would provide great benefit.

Our work has two important limitations. First, we developed our model prior to the development of highly active antiretroviral therapy. Highly active antiretroviral therapy has sharply reduced the mortality of HIV infection. The costs of HIV care are also now substantially different than when our model was developed. Highly active antiretroviral therapy has reduced hospitalizations and associated costs. However, the cost of highly active antiretroviral therapy may reach $15,000 per year, so drug costs have increased substantially. We are extending the analyses discussed here to account for both the reduced mortality and changing patterns of expenditures on HIV care. Although the quantitative results will certainly be somewhat different, we expect that the main qualitative findings of the current analyses will remain largely unchanged: all types of vaccine efficacy will be important, and vaccines with modest protection will likely provide substantial benefit. Second, as noted, recent developments in HIV vaccine research suggest that a vaccine may have all three mechanisms of protection. In future work, we will evaluate vaccines that provide all three components of vaccine efficacy.

HIV now ranks as one of the most devastating pandemics in history. The development of a preventive or therapeutic vaccine, or a vaccine with hybrid characteristics, is an extraordinary public-health priority. Our evaluation indicates that a full understanding of the health benefit of HIV vaccines will require assessment of all three types of potential protection. Empiric assessment of the degree to which a vaccine protects from infection, and reduces progression of disease or transmission, will require long, complex, and expensive clinical trials. Fortunately, our analyses indicate that a vaccine need not be perfect or nearly so to have great health benefit. Even a partially effective HIV vaccine could dramatically change the course of the HIV epidemic.

Acknowledgments

This work was supported by Grant R01 DA15612-01 from the National Institute on Drug Abuse and supported in part by the Department of Veterans Affairs.

References

[1] UNAIDS Joint United Nations Programme on HIV/AIDS (2002). *AIDS Epidemic Update,* http://www.unaids.org, Accessed February 16, 2003.

[2] Hu, D.J., C.R. Vitek, B. Bartholow, and T.D. Mastro (2003). Key issues for a potential human immunodeficiency virus vaccine. *Clinical Infectious Diseases,* 36, 638-644.

[3] Longini, I.M., Jr., S. Datta, and M.E. Halloran (1996). Measuring vaccine efficacy for both susceptibility to infection and reduction in infectiousness for prophylactic HIV-1 vaccines. *Journal on AIDS and Human Retrovirology,* 13, 440-447.

[4] Halloran, M.E., C.J. Struchiner, and I.M. Longini, Jr. (1997). Study designs for evaluating different efficacy and effectiveness aspects of vaccines. *American Journal of Epidemiology,* 146, 789-803.

[5] Datta, S., M.E. Halloran, and I.M. Longini, Jr. (1998). Augmented HIV vaccine trial design for estimating reduction in infectiousness and protective efficacy. *Statistics in Medicine,* 17, 185-200.

[6] Gray, R.H., et al. (2001). Probability of HIV-1 transmission per coital act in monogamous, heterosexual, HIV-1-discordant couples in Rakai, Uganda. *The Lancet*, 357, 1149-1153.

[7] Quinn, T.C., et al. (2000). Viral load and heterosexual transmission of human immunodeficiency virus type 1. Rakai Project Study Group. *New England Journal of Medicine,* 342, 921-929.

[8] Edwards, D.M., R.D. Shachter, and D.K. Owens (1998). A dynamic HIV-transmission model for evaluating the costs and benefits of vaccine programs. *Interfaces,* 28, 144-166.

[9] Owens, D.K., D.M. Edwards, and R.D. Shachter (1998). Population effects of preventive and therapeutic HIV vaccines in early- and late-stage epidemics. *AIDS,* 12, 1057-1066.

[10] Owens, D.K., D.M. Edwards, and R.D. Shachter (2001). Costs and benefits of imperfect HIV vaccines: Implications for vaccine development and use, in *Quantitative Evaluation of HIV Prevention Programs,* E.H. Kaplan and R. Brookmeyer, Eds., Yale Press, New Haven, CT.

[11] Owens, D.K. and H.C. Sox (2000). Medical Decision Making: Probabilistic Medical Reasoning, in *Medical Informatics: Computer Applications in Health Care and Biomedicine,* E.H. Shortliffe, *et al.,* Eds., Springer-Verlag, New York.

[12] Gold, M., J.E. Siegel, L.B. Russell, and M.C. Weinstein, Eds. (1996). *Cost-Effectiveness in Health and Medicine.* Oxford University Press, New York.

16 DESIGNING PEDIATRIC FORMULARIES FOR CHILDHOOD IMMUNIZATION USING INTEGER PROGRAMMING MODELS

Sheldon H. Jacobson[1] and Edward C. Sewell[2]

[1] Department of Mechanical and Industrial Engineering
University of Illinois
Urbana, IL 61801

[2] Department of Mathematics and Statistics
Southern Illinois University Edwardsville
Edwardsville, IL 62026

SUMMARY

Several challenges have arisen in childhood immunization programs as vaccine manufacturers have become more successful in developing new vaccines for childhood diseases. Their success has created a combinatorial explosion of choices for health-care providers and other purchasers of pediatric vaccines, which in turn has created a new set of economic problems and issues related to determining which vaccines should be combined into single injections and how to design lowest overall cost formularies for pediatric immunization. This chapter provides a review of how operations research modeling and analysis tools can be used to address a variety of economic issues surrounding pediatric vaccine formulary design and pricing. A summary is presented of the pediatric immunization problems that have been studied using integer programming models, as well as the assumptions used to model such problems. A description of the methodologies used is provided. A summary of the results obtained with these models for a particular pentavalent combination vaccine that recently gained Food and Drug Administration (FDA) approval for pediatric immunization is presented. Concluding comments and directions for future research are also discussed.

KEY WORDS

Pediatric immunization, Combination vaccines, Economics, Integer programming models

16.1 INTRODUCTION

The United States recommended childhood immunization schedule has become increasingly complex. For example, the 2002 schedule required no less than five clinic visits and 19 injections over the first 18 months of life [1], with each clinic visit scheduled around the specifications and requirements associated with the vaccines. Other constraints that may lead to additional clinic visits include a child's tolerance to multiple injections during a single clinic visit [2], and parents' (or guardians') ability to take the time from their jobs to make immunization visits [3]. These obstacles often lead to noncompliance with the recommended childhood immunization schedule, increasing the risk to children of contracting the diseases that the vaccines were designed to prevent, resulting in a tremendous cost burden on the nation's already stressed health-care system.

These problems are further exacerbated by biotechnological advances that have led to new pediatric vaccines that must be incorporated into the already overcrowded recommended childhood immunization schedule. For example, the four recommended doses of oral polio vaccine (OPV) in the 1996 recommended childhood immunization schedule were replaced with four injections of inactivated polio vaccine (IPV) in the January 2002 schedule [1]. In 2000, four doses of a new 7-valent conjugate vaccine for pneumococcal disease ($\mathrm{PNU_{cn\text{-}7}}$) were recommended to be included in the recommend childhood immunization schedule [4]. Meeting the guidelines set forth in the recommended childhood immunization schedule may now require up to five injections at each of three recommended immunization visits (2, 4, and 6 months) in the first year of life.

New pediatric vaccines that gain Food and Drug Administration (FDA) approval and are added to the recommended childhood immunization schedule by the Advisory Committee on Immunization Practice (ACIP) will exert pressure to increase both the volume and the frequency of immunization visits, and hence further escalate the costs associated with well-baby care (i.e., routine medical care check-ups during the first few years of life). One approach to circumvent this approaching problem is to create a single-dose oral vaccine that immunizes children at birth from all childhood diseases [5]. A more realistic solution is to combine two or more vaccines to reduce the required number of injections and clinic visits [6, 7]. Assuming equivalent efficacy of combination vaccines compared to their monovalent counterparts, significantly less time for clinic visits would be required of parents. This may in turn result in higher immunization compliance rates and an ensuing decrease in the number of children afflicted with childhood diseases.

Combination vaccines have their own unique set of problems [8]. From the vaccine manufacturers' point of view, determining which vaccines to combine (based on biological compatibility of the antigens) and how to administer the vaccines (both in sequence and in timing) to ensure that immunity is achieved without compromising safety are important questions that need to be resolved. Moreover, the issue of extra vaccination (i.e., administering vaccine components that are not required by the childhood immunization schedule during specific vaccination periods) and the degree to which it should be tolerated (so as to minimize any associated negative side effects and the cost of administering unneeded vaccine components) must be addressed. Lastly, the overall objective of designing an economical package of vaccine types and brands to stock for a particular immunization environment must be addressed. Identifying solutions to these problems can be daunting for even the most experienced pediatric health-care researchers and professionals. Nonetheless, combining antigens that provide protection against multiple pediatric diseases into a single injection has been recognized by pediatric health-care providers to be an advance of significant benefit. In fact, licensed combination vaccines are officially preferred over their individual component vaccines because they reduce the pain and suffering associated with multiple injections [9].

This chapter reviews the application of integer programming models to address the design of economical pediatric vaccine formularies. These models were initially introduced to provide quantitative tools for use by the Centers for Disease Control and Prevention (CDC), health-care providers, insurance companies, and parents. Such tools can help them make well-informed vaccine formulary decisions [10, 11]. The models are designed using the principle that decisions based on purchase price alone, ignoring the economic value of distinguishing features among competing vaccine products, can be more costly in the long run. The resulting integer programming models assemble from among all available vaccine products at their market prices the vaccine formulary that provides the best value within the constraints of the immunization schedule, achieving the lowest overall cost to society or to any other desired perspective. The models select from among a set of monovalent (i.e., single antigen) and combination vaccines those products that should be used at which scheduled visits within the recommended childhood immunization schedule [1].

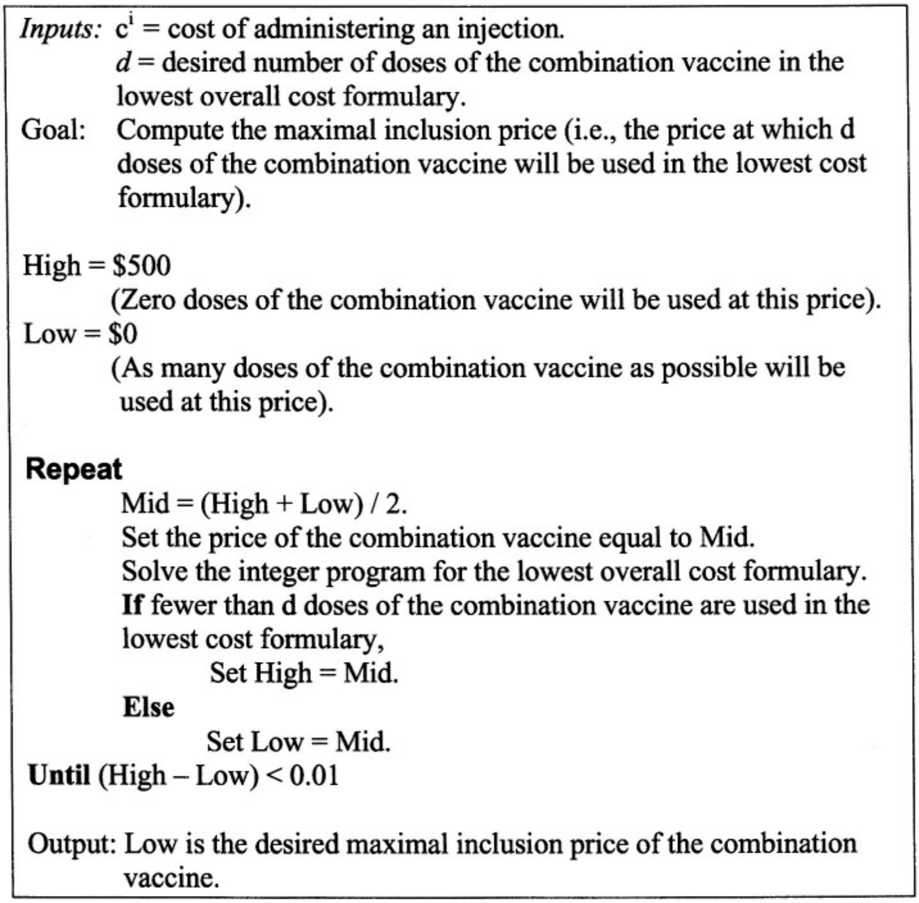
The chapter is organized as follows: Section 16.2 summarizes the pediatric immunization problems that have been studied using integer programming models. Section 16.3 provides a description of the methodologies used, as well as the assumptions used to model the problems. Section 16.4 summarizes the results obtained with these models for a particular pentavalent combination vaccine that is well positioned to become available for pediatric immunization in the near future. Section 16.5 provides concluding comments and directions for future research.

16.2 PROBLEM DESCRIPTION

Weniger et al. [10] report the results of a pilot study that shows how integer programming models can be used to design optimal pediatric vaccine formularies for a subsection of the recommended childhood immunization schedule. The concept "optimal" here refers to a vaccine formulary that provides the best economic value, considering more than just vaccine prices alone. The authors present an integer programming model to assist vaccine purchasers in making decisions about which vaccine products to include in their formulary (i.e., to stock in their inventory). The model takes into account not only the price of the vaccines, but also the cost of a clinic visit, the time to prepare a vaccine for injection, and the cost of administering injections. Jacobson et al. [11] report the technical details of these models. They also list several different vaccine formularies obtained by solving the model under a variety of economic criteria. The model does not make decisions about a specific vaccine product in isolation but, rather, assembles from among all competing monovalent and combination vaccines the formulary that satisfies the recommended childhood immunization schedule at the lowest overall cost to society (or, if desired, to any more limited perspective, such as the payer of direct health costs). The model's design is based on the principle that purchase price alone is just one of many factors with economic consequences that should be taken into account. The key contribution of this study is the result that it may be myopic to use vaccine prices as the sole factor to determine which vaccines should be purchased and that other costs within the system can be captured and used to identify vaccine formularies that provide a good overall value.

To encourage and evaluate new investment and research by the pharmaceutical industry for innovative and new vaccine products, Sewell et al. [12] use the integer programming models to *reverse engineer* the price of various combination vaccines using an iterative bisection search algorithm [13]. This algorithm is detailed in Figure 16.1. The algorithm determines the maximum inclusion price of each combination vaccine, with and without the perinatal dose of hepatitis B (i.e., a dose administered at birth), across five injection cost variations. This involves dividing arbitrary, extreme upper and lower values for the maximal price into equal-sized upper and lower ranges, and then solving the integer programming model to determine which of these ranges contains the maximal price. The identified range is then divided in half, and the process is iteratively repeated until the algorithm converges when the final upper and lower range is less than $0.01, revealing the maximal price to the nearest one cent. Note that for a bisection

Figure 16.1 Reverse engineering algorithm



```
Inputs: $c^i$ = cost of administering an injection.
        $d$ = desired number of doses of the combination vaccine in the
        lowest overall cost formulary.
Goal:   Compute the maximal inclusion price (i.e., the price at which d
        doses of the combination vaccine will be used in the lowest cost
        formulary).

High = $500
        (Zero doses of the combination vaccine will be used at this price).
Low = $0
        (As many doses of the combination vaccine as possible will be
        used at this price).

Repeat
        Mid = (High + Low) / 2.
        Set the price of the combination vaccine equal to Mid.
        Solve the integer program for the lowest overall cost formulary.
        If fewer than d doses of the combination vaccine are used in the
        lowest cost formulary,
                Set High = Mid.
        Else
                Set Low = Mid.
Until (High – Low) < 0.01

Output: Low is the desired maximal inclusion price of the combination
        vaccine.
```

search to be effective, a well-behaved cost function is required, such as one that is convex over the feasible region of possible formularies, which was the case for this study.

Reverse engineered prices for combination vaccines provide vaccine manufacturers with guidelines on how well future vaccine combinations may compete in the market, and hence provide information that can be used to determine how long it may take to recoup research and development investments in such products. In recent years, several pentavalent and hexavalent vaccines built around diphtheria, tetanus, and acellular pertussis **(DTP_a)** backbones have been under development and their developers are positioning them to gain FDA approval [14]. Sewell and Jacobson [15] provide technical details of the reverse engineering algorithm that determines the maximum price at which different combination vaccines provide an overall economic advantage, and hence belong in a lowest overall cost formulary. Jacobson and Sewell [16] incorporate the reverse engineering algorithm into a Monte Carlo simulation model to determine probability distributions for the price of four combination vaccines. Health-care providers and parents each place a different value (hence cost) on each injection administered (or avoided). Therefore, for a given set of injection costs there exists a *maximal price* at which a combination vaccine joins the lowest overall cost formulary (i.e., provides a good economic value). This maximal price can be determined by iteratively solving the model in Jacobson and Sewell [16]. Monte Carlo simulation is used to sample the injection costs from a set of probability distributions, where each probability distribution corresponds to the values that a population of parents may place on administering or avoiding an injection. The resulting set of maximal prices for each combination vaccine is used to create an empirical distribution that estimates the probability distribution of maximal prices for that combination vaccine. This probability distribution can be used, for example, to estimate the proportion of a population of parents who are willing to purchase the combination vaccine at a given price. Therefore, the maximal price probability distribution provides marketing information for vaccine manufacturers. Jacobson and Sewell [16] also use different injection cost probability distributions to assess the sensitivity of the maximal price distribution to the form of this probability distribution.

16.3 METHODOLOGY AND ASSUMPTIONS

A generic integer programming model is presented that captures the first 12 years of the 2002 recommended childhood immunization schedule for immunization against any subset of childhood diseases covered by the recommended childhood immunization schedule (which currently includes hepatitis B, diphtheria, tetanus, pertussis, *Haemophilus* influenzae type B, polio, measles, mumps, rubella, varicella, and pneumococcus). This model is an extension of the integer programming model introduced in the pilot study reported in [11]. The generic integer programming model is as follows:

Parameters

V = set of vaccines that may be administered

B = $\{AVP, GSK, MRK, WYE\}$ = set of manufacturers (brands)

b_v = brand of vaccine $v \in V$

T^* = $\{HBV, DTP_a, Td, HIB, IPV, MMR, VAR, PCV\}$ = set of standard sets of antigens

T_v = set of standard sets of antigens contained in vaccine $v\in V$

M^* = $\{0\text{-}1, 2, 4, 6, 12\text{-}18, 60, 144\}$ = set of months in which vaccines may be administered

M_v = set of months in which vaccine $v\in V$ may be administered

c^i = cost of administering an injection

c^V = cost of visiting a clinic

c_v = cost of vaccine $v\in V$ including the preparation cost = price of vaccine plus preparation cost

X = $\{(v,m)\colon\ v\in V,\ m\in M_v\}$ = set of pairs (v,m) such that it is permissible to administer vaccine $v\in V$ in month $m\in M_v$

X_t = $\{(v,m)\in X\colon t\in T_v\}$ for all $t\in T^*$

Variables

$$x_{v,m}=\begin{cases}1 \text{ if vaccine } v \text{ is given in month } m\\ 0 \text{ else}\end{cases} \quad \text{for all } (v,m)\in X$$

$$HibSkip6=\begin{cases}1 & \text{if HIB can be skipped in month 6 due to using MRK Hib}\\ & \quad\text{in months 2 and 4}\\ 0 & \text{otherwise}\end{cases}$$

s_m = number of shots (injections) given in month m, for all $m\in M^*$

$$\mathrm{v}_m=\begin{cases}1 \text{ if the clinic is visited in month } m\\ 0 \text{ else}\end{cases} \quad \text{for all } m\in M^*$$

Objective Function

$$\min \sum_{(v,m)\in X} c_v x_{v,m} + \sum_{m\in M^*} c^v \mathrm{v}_m + \sum_{m\in M^*} c^i s_m \tag{1}$$

Constraints

$$\sum_{(v,0)\in X_{HBV}} x_{v,0} = \begin{cases}1 \text{ if perinatal dose of HBV is to be given}\\ 0 \text{ else}\end{cases} \tag{2}$$

$$\sum_{(v,2)\in X_{HBV}} x_{v,2} + \sum_{(v,4)\in X_{HBV}} x_{v,4} \geq 1 \tag{3}$$

$$\sum_{(v,6)\in X_{HBV}} x_{v,6} + \sum_{(v,12-18)\in X_{HBV}} x_{v,12-18} \geq 1 \tag{4}$$

$$\sum_{(v,m)\in X_{HBV}:m=0-1,2,\text{ or }4} x_{v,m} \geq 2 \tag{5}$$

$$\sum_{(v,m)\in X_{DTPa}} x_{v,m} \geq 1 \quad,\ m = 2, 4, 6, 12\text{-}18, 60 \tag{6}$$

$$\sum_{(v,144)\in X_{Td}} x_{v,144} \geq 1 \tag{7}$$

$$\sum_{(v,m)\in X_{HIB}} x_{v,m} \geq 1 \quad,\ m = 2, 4, 12\text{-}18 \tag{8}$$

$$HibSkip6 + \sum_{(v,6)\in X_{HIB}} x_{v,6} \geq 1 \tag{9}$$

$$\sum_{(v,m)\in X_{IPV}} x_{v,m} \geq 1 \quad,\ m = 2, 4, 60 \tag{10}$$

$$\sum_{(v,m)\in X_{IPV}:m=6\text{ or }12-18} x_{v,m} \geq 1 \tag{11}$$

$$\sum_{(v,m)\in X_{PCV}} x_{v,m} \geq 1 \quad,\ m = 2, 4, 6, 12\text{-}18 \tag{12}$$

$$\sum_{(v,m)\in X_{MMR}} x_{v,m} \geq 1 \quad,\ m = 12\text{-}18, 60 \tag{13}$$

$$\sum_{(v,15)\in X_{VAR}} x_{v,15} \geq 1 \tag{14}$$

$$\sum_{\substack{(v,2)\in X_{DTPa}:\\ b_v=b}} x_{v,2} \geq \sum_{\substack{(v,m)\in X_{DTPa}:\\ b_v=b}} x_{v,m} \quad,\ m = 2, 4, 6, 12\text{-}18, 60 \text{ for all } b \in B \tag{15}$$

$$s_m = \sum_{(v,m)\in X} x_{v,m} \quad \text{for all } m \in M^* \tag{16}$$

$$10\nu_m \geq s_m \quad \text{for all m} \in M^* \tag{17}$$

$$\sum_{(v,m)\in X_{HIB}:b_v=MRK} x_{v,m} \geq HibSkip6 \quad , \quad m=2,4 \tag{18}$$

$$\sum_{\substack{(v,m)\in X_{HIB}:\\ b_v=MRK,m=2\text{ or }4}} x_{v,m} - HibSkip6 \leq 1 \tag{19}$$

$$\sum_{(v,m)\in X_{DTPa}} x_{v,m} = 5 \tag{20}$$

$$\sum_{(v,m)\in X:t\in T_v} x_{v,m} \leq 1 \quad \text{for all } t\in T^*, \text{ for all } m\in M^* \tag{21}$$

The objective function of the integer programming model is given by (1). The constraints of the model are as follows. Constraint (3) requires that HBV must be given in month 2 or 4, and constraint (4) requires that HBV must be given in month 6 or month 12-18. Constraint (5) says that the first two doses of HBV must be given in months 0-1, 2, or 4. Constraint (6) says that $\mathbf{DTP_a}$ is required in months 2, 4, 6, 12-18, and 60. Constraint (7) says that Td is required in month 144. Constraint (8) says that HIB is required in months 2, 4, and 12-18, while constraint (9) says that HIB is required in month 6 unless MRK HIB is used in months 2 and 4. Constraint (10) says that IPV is required in months 2, 4, and 60, while constraint (11) says that IPV is required in months 6 or 12-18. Constraint (12) says that PCV is required in months 2, 4, 6, and 12-18. Constraint (13) says that MMR is required in months 12-18 and 60. Constraint (14) says that VAR is required in month 15. Constraint (15) enforces $\mathbf{DTP_a}$ brand matching. Constraint (16) calculates the number of shots given in each month. Constraint (17) ensures that the clinic is visited in each month in which a vaccine is administered. Constraint (18) says that HIB can be skipped in month 6 only if MRK HIB is used in months 2 and 4; if this is so, then constraint (19) sets the variable *HibSkip6* to 1. Constraint (20) says that $\mathbf{DTP_a}$ extravaccination is not permitted. Finally, constraint (21) says that two doses of the same standard set of antigens are not allowed in the same month.

To illustrate the use of this model, the pentavalent combination vaccine – comprising vaccines for diphtheria, tetanus, acellular pertussis, hepatitis *B,* and inactivated polio (labeled $\mathbf{DTP_a}$**-HBV-IPV)** – is analyzed to determine the number of doses of the vaccine that earn a place in the lowest overall cost formulary at varying price levels. This particular combination vaccine was chosen since it is well-positioned to become available for pediatric immunization in the near future.

Several assumptions are made that provide boundaries for the scope of the results presented. Unless otherwise noted, the assumptions in Sewell et al.

[12] are used. The federally negotiated vaccine price list (effective August 9, 2002) is used by four pharmaceutical companies (labeled AVP = Aventis Pasteur, MRK = Merck, GSK = GlaxoSmithKline, WYE = Wyeth-Lederle) that manufacture all the vaccines that are currently licensed and under federal contract with the CDC for childhood immunization. These four vaccine manufacturers produce 14 vaccine products that protect against the 11 diseases (labeled HBV for the hepatitis B vaccine, $\mathbf{DTP_a}$ for the diphtheria, tetanus, acellular pertussis vaccine, HIB for the *Haemophilus* influenzae type B vaccine, IPV for the inactivated polio vaccine, MMR for the measles, mumps and rubella vaccine, $\mathbf{PNU_{cn\text{-}7}}$ for the pneumococcal vaccine, and VAR for the varicella vaccine).

The cost function for the integer programming model includes

- the purchase price of all licensed vaccines under federal contract,
- the cost of each clinic visit,
- the cost of vaccine preparation by medical staff,
- the cost of administering an injection.

Values used for these costs are shown in Table 16.1. The vaccine purchase prices are the federally negotiated prices as of August 9, 2002 (see Table 16.1). The cost of a clinic visit is set at $40, the same value used in the CDC pilot study and the more recent studies [10-12, 15, 16]. This cost includes the direct and indirect costs associated with a clinic being able to administer vaccines [10].

Vaccine preparation is assumed to require 3.0 minutes per dose for powdered vaccines [p]. This preparation requires vaccine reconstitution: diluent is drawn into a syringe, transferred into the vaccine vial, which is shaken, and then the liquefied vaccine is withdrawn. Liquid vaccines in vials [v] requiring entry with a needle to draw up into a syringe were assigned 1.5 minutes, and ready-to-administer prefilled syringes [s] were assigned 0.5 minutes. These assumptions were distributed around the average times observed in previous studies [17, 18]. Note that these times are also consistent with unpublished developing-country estimates of around 1.0 minute for filling and administering injections from disposable syringes and 80 seconds for resterilizable syringes [19]. Labor costs are set at $0.50 per minute, as in previous studies [12, 13]; this is equivalent to an annual total compensation of $60,000 for a 2,000 hour work year. Therefore, the resulting preparation costs for powders [p], vials [v], and syringes [s] are $1.50, $0.75, and $0.25 per dose, respectively (see Table 16.1). Three of the vaccine products, one brand of $\mathbf{DTP_a}$, one brand of HBV, and one brand of IPV, are available in both pre-filled syringes and liquid vial formulations.

Table 16.1 Vaccine list of existing products used in the vaccine selection model, including actual formulation and packaging features and federal contract vaccine price (as of August 9, 2002), and assumed nursing time preparation cost

Vaccine *	Formulation/ packaging †	Mfr.	8/9/02 Federal Purchase Price/ Dose §	Prep-aration Cost/ Dose	Subtotal Cost/ Dose ¶
DTP$_a$	[v]	AVP	\$11.75	\$0.75	\$12.50
DTP$_a$	[s]	GSK	\$12.00	\$0.25	\$12.25
DTP$_a$-HIB	[p]	AVP	\$23.40	\$1.50	\$24.90
IPV	[s]	AVP	\$8.80	\$0.25	\$9.05
HIB-HBV	[v]	MRK	\$21.83	\$0.75	\$22.58
HBV	[s]	GSK	\$9.25	\$0.25	\$9.50
HBV	[v]	MRK	\$9.00	\$0.75	\$9.75
HIB	[p]	AVP	\$7.25	\$1.50	\$8.75
HIB	[v]	MRK	\$8.32	\$0.75	\$9.07
HIB	[v]	WYE	\$7.33	\$0.75	\$8.08
MMR	[p]	MRK	\$15.64	\$1.50	\$17.14
PNU$_{cn-7}$	[v]	WYE	\$45.99	\$0.75	\$46.74
VAR	[p]	MRK	\$41.44	\$1.50	\$42.94
Td	[v]	AVP	\$8.76	\$0.75	\$9.51

* Abbreviations: DTP$_a$ – Diphtheria toxoid, tetanus toxoid, and acellular pertussis vaccine; HIB – *Haemophilus influenzae* type B conjugate vaccine; DTP$_a$-HIB – Diphtheria toxoid, tetanus toxoid, acellular pertussis, and *Haemophilus influenzae* type B conjugate vaccine; HBV – Hepatitis B virus vaccine; HIB-HBV – *Haemophilus influenzae* type B conjugate vaccine and hepatitis B virus vaccine; IPV – polivirus vaccine, inactivated; MMR – Measles-mumps-rubella; PNU$_{cn-7}$ – 7-valent conjugate pneumococcal vaccine; VAR – Varicella; Td – Tetanus toxoid

† [v] = liquid vaccine in vial; [s] = liquid vaccine in prefilled syringe; [p] = powdered (lyophilized) HIB vaccine component requiring reconstitution by addition of diluent or the liquid antigens of the remainder of the combination vaccine.

§ Current federal-contract and private-sector vaccine prices are available at: www.cdc.gov/nip/vfc/cdc_vaccine_price_list.htm

¶ Excludes clinic visit cost and injection cost.

Since the pre-filled syringes require one less minute of preparation time, only the packaging that results in the lowest overall cost (the price of the vaccine plus the cost of vaccine preparation by medical staff) is included in the model. Therefore, for these three vaccine products, only the pre-filled syringes are considered in the analysis. Note that the total cost per dose listed in Table 16.1 does not include the cost of administering each injection.

Weniger et al. [10] observe that the cost associated with administering an injection can be broken down into several components. The first component is the actual direct cost of administering the vaccine. Lieu et al. [9] suggest this cost to be approximately $5 per injection. The second component is the direct cost for repeat clinic visits if injections are refused by the parents (e.g., when four or more injections are required at a particular clinic visit). This cost is estimated to be approximately $3 per injection. The third component is the indirect cost of lost time from work by parents for repeat clinic visits if they refuse injections. This cost is estimated to be approximately $12 per injection (which is averaged over all injections administered during a single clinic visit as well as all the parents of the children being immunized). The fourth component is the indirect cost of “pain and emotional distress” associated with each injection, as measured by a parent’s “willingness-to-pay” to avoid such pain. This cost has been estimated to be as high as $25 per injection [9] or more conservatively, as $8 per injection [20]. The results reported in [21] independently support these values.

The following assumptions are used in analyzing the **DTP_a-HBV-IPV** combination vaccine. These assumptions were all used in the CDC pilot study reported in [10, 11] as well as the more recent studies reported in [12, 15, 16].

(i) The recommended childhood immunization schedule was followed for immunization against 11 diseases: hepatitis B, diphtheria, tetanus, pertussis, *Haemophilus* influenzae type B, polio, measles, mumps, rubella, varicella, and pneumococcus.

(ii) Injections can be administered in months 0-1, 2, 4, 6, 12-18, 60, and 144, providing seven months (or periods) to administer vaccines. Only one clinic visit can occur in each of these months (or periods). All injections in a given month (or period) are administered in a single clinic visit.

(iii) Only the 14 currently available and under federal contract vaccines are included in the model, with the exception of the combination vaccine being studied.

(iv) *Haemophilus* influenzae type B vaccines can only be administered in month 2 or later.

(v) The first hepatitis B vaccine injection must be administered in either month 0-1 (referred to as the perinatal dose of hepatitis B) or month 2.

(vi) If *Haemophilus* influenzae type B vaccine products by Merck are administered in both months 2 and 4, then no *Haemophilus* influenzae type B vaccine is required in month 6.

(vii) Manufacturer brand matching is required for diphtheria, tetanus, and pertussis vaccines, but not for *Haemophilus* influenzae type B and hepatitis B vaccines.

These assumptions are based on the guidelines set forth in the 2002 recommended childhood immunization schedule [1]. Vaccination plans that deviate from this schedule are not considered.

Users of the integer programming models have the flexibility to set any value for each cost included, such as the cost of an injection. Although the model was designed principally for use by major vaccine purchasers, such as public health agencies, health-care organizations, and large private clinics, the reverse engineering approach [15] also provides vaccine manufacturers with a tool for determining the price at which a proposed new vaccine product would win a place in the lowest overall cost formulary. Such a tool may be used to guide investment decisions and develop priorities towards products with the most beneficial economic impact within the pediatric immunization market.

The integer programming model determines the lowest overall cost formulary that satisfies the recommended childhood immunization schedule and the assumptions listed in this section. The lowest overall cost formulary is the set of vaccines that satisfies the immunization schedule at minimum cost, where cost includes the factors described above. The model has been designed to consider all monovalent and combination vaccines (to prevent the diseases covered by the schedule) that are licensed in the United States and under federal contract for purchase by the CDC. Therefore, all the analysis in this chapter focuses on vaccines under federal contract (and their associated federal contract prices).

For monopoly suppliers of particular vaccines, where there is no choice between competing products, the integer programming model solution trivially contains the single product available. Therefore, the polio vaccine, the measles-mumps-rubella vaccine, the varicella vaccine, and the 7-valent conjugate vaccine for pneumococcal disease are all chosen to be part of the lowest overall cost formulary. Moreover, for simplicity, hepatitis A is excluded from the analysis since it is not widely recommended as a pediatric vaccine, except for a number of western states [22]. The hepatitis A vaccine also does not yet appear in any pediatric combination vaccine. Therefore, assuming no great disparity from the range of injection cost estimates used here, and no major changes in purchase pricing, the integer programming model solution would include the two-dose regimen (versus three-dose) for the hepatitis A vaccine with the lowest overall purchase price and preparation cost.

Computer experiments were conducted using the integer programming model to assess the economic value of the $\mathbf{DTP_a}$**-HBV-IPV** combination vaccine, Pediarix ©, manufactured by GSK. These experiments use the reverse engineering algorithm to determine the number of doses of the vaccine that earn a spot in the lowest overall cost formulary as a function of the cost of administering an injection, given the prices for the vaccines currently under federal contract (at their federal contract prices). The preparation cost for this combination vaccine was set at $0.25 per dose (assuming that it is available in prefilled syringes [s]).

The two scenarios of administering and not administering the perinatal dose of hepatitis B were considered and analyzed. The price regions obtained for the $\mathbf{DTP_a}$**-HBV-IPV** combination vaccine are a function of the number of doses of the combination vaccine that earn a place in the lowest overall cost formulary. In general, the lower the vaccine is priced, the more doses it earns in the lowest overall cost formulary.

The four objective function components described in Section 16.2 were used to determine the overall cost for immunization that satisfies the 2002 recommended childhood immunization schedule [1]. As noted in Section 16.2, the cost of administering an injection is highly subjective and may include only direct medical costs, or both direct and indirect costs, based on the perspective of the payer [11]. Therefore, this cost was varied to determine a range of maximal prices for the $\mathbf{DTP_a}$**-HBV-IPV** combination vaccine.

16.4 RESULTS

Reverse engineering experiments were conducted to determine the number of doses of the $\mathbf{DTP_a}$**-HBV-IPV** combination vaccine that earn a place in the lowest overall cost formulary, as a function of the price of the combination vaccine and the cost of an injection (between $0 and $40), for the two scenarios of whether or not the perinatal dose of hepatitis B is administered.

Each experiment without the perinatal dose of hepatitis B resulted in zero, one, three, four, and five dose regions for the combination vaccine, while the experiments with the perinatal dose of hepatitis B resulted in zero, one, two, four, and five dose regions for the vaccine. All the experiments were conducted using the web-adapted version of the integer programming model, which is available at www.vaccineselection.com.

The maximal price regions are displayed in Figures 16.2 and 16.3. In general, as the cost of an injection increased, the maximal prices increased. This follows from the observation that the combination vaccine saves as many as two injections in each vaccine period compared to the use of currently available vaccines to administer the same antigens. Once again, the higher the cost of an injection (direct medical, indirect societal, or both), the greater the value of higher valency vaccines.

As the number of doses of the combination vaccine to be administered in the lowest overall cost formulary rose, the maximal prices decreased. In general, across varying cost assumptions and product formulations, the higher the maximal price, the greater the economic value that is being captured by the DTP_a-HBV-IPV combination vaccine in comparison to its monovalent counterparts. At fewer doses, the higher maximal prices result from the combination vaccine being highly competitive with its monovalent counterparts when it entirely replaces all the separate vaccines indicated at a specific visit. However, at a higher number of doses, the reduced maximal price reflects the case in which a monovalent vaccine can be administered for only the indicated antigens and thus avoid the economic inefficiency of extra vaccination. For example, the $12.00 price needed to achieve five doses of the combination vaccine (see Figures 16.2 and 16.3) reflects the large amount of extra vaccination that is required to administer five doses of the vaccine [23, 24]. Since the integer programming model is designed to place value only on those vaccine components that are required to satisfy the recommended childhood immunization schedule, and sets a value of zero for those components that correspond to extravaccinated doses, requiring five doses of the combination vaccine with the perinatal dose of hepatitis B administered is not economical. In essence, the maximal price of a combination vaccine absorbs any premium arising from savings associated with the cost of an injection or vaccine preparation costs. Therefore, the reverse engineering process shifts such costs (e.g., savings to the health-care system) into the price of the combination vaccine.

Figure 16.4 shows the difference between the maximal price for the DTP_a-HBV-IPV combination vaccine for the one-, three-, and four-dose cases (without the perinatal dose of hepatitis B administered) and the sum of the prices of the individual vaccines (GSK DTP_a, GSK HBV, AVP IPV), as a

Figure 16.2 Number of doses of DTP$_a$-HBV-IPV combination vaccine (without the perinatal dose of hepatitis B administered)

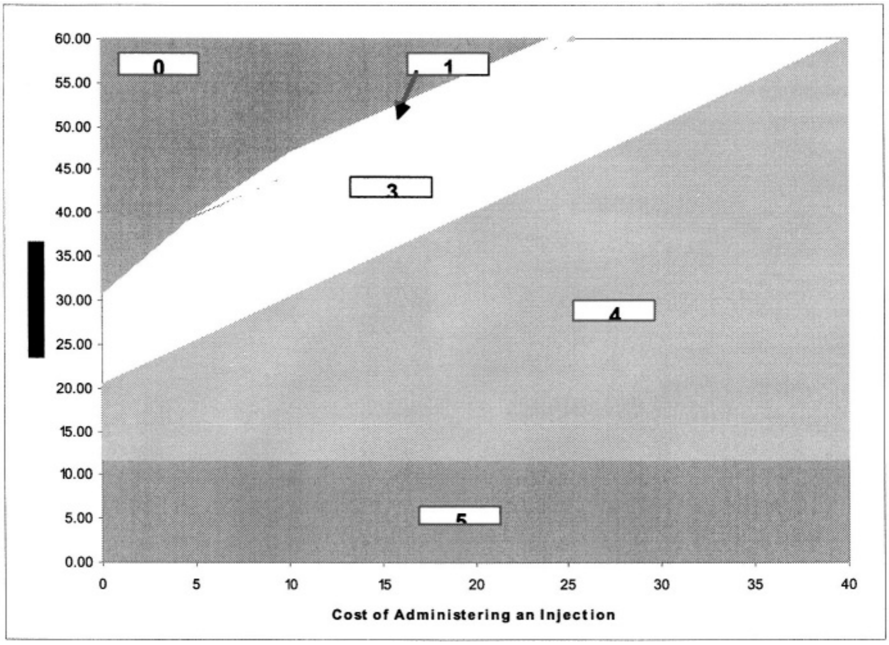


Figure 16.3 Number of doses of DTP$_a$-HBV-IPV combination vaccine (with the perinatal dose of hepatitis B administered)

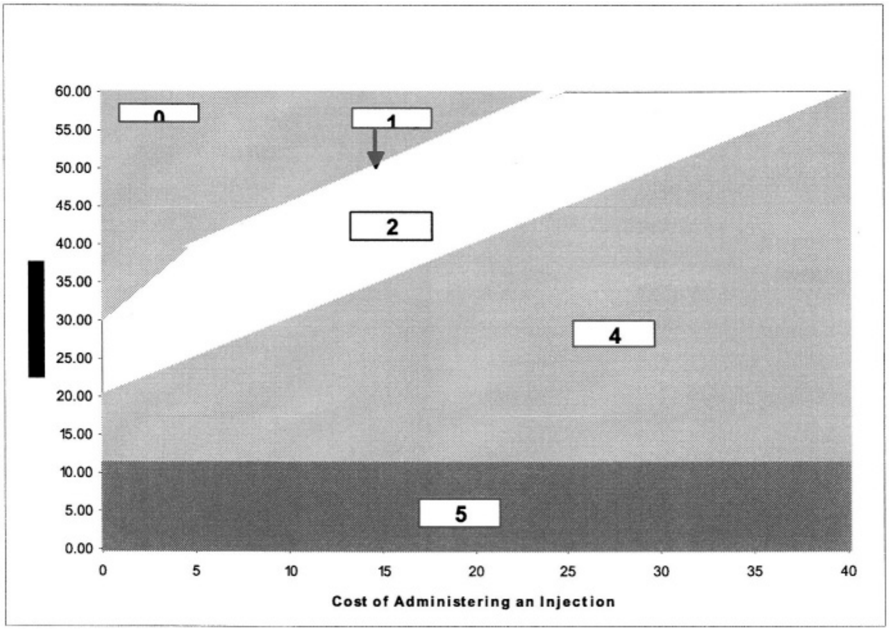
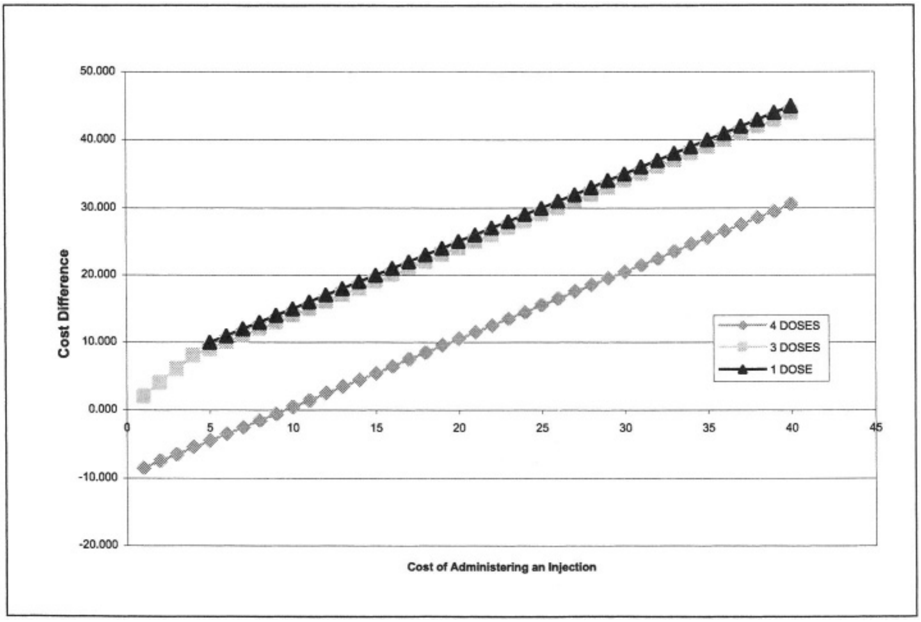


Figure 16.4 Difference between the maximal price for the DTP$_a$-HBV-IPV combination vaccine (without the perinatal dose of hepatitis B administered) and the sum of the prices of the individual vaccines



function of the cost of an injection. The five-dose case is not considered since the five-dose maximal price for the DTP$_a$-HBV-IPV combination vaccine is $12.00, which is significantly below $30.05, the sum of the prices of the individual vaccines (due to extravaccination). Therefore, it would not be feasible for a vaccine manufacturer to sell this combination vaccine at such a price. The four-dose maximal price for the DTP$_a$-HBV-IPV combination vaccine is also below $30.05 for all injection cost values below $9.00. The three- and one-dose maximal prices for the DTP$_a$-HBV-IPV combination vaccine are greater than $30.05 for all the injection cost values. Therefore, it is reasonable to expect that the best marketing strategy (from the manufacturers' point of view) for the DTP$_a$-HBV-IPV combination vaccine is to expect no more than three doses of the vaccine per child. All these differences are the same as when the perinatal dose of hepatitis B is not administered, except the three-dose case becomes a two-dose case.

As in the original study [10-11], the integer programming model excludes several potential factors of economic consequence to the overall cost of immunization. This is due to the dearth of reliable data. For example, the model ignored potential differences between competing products in vaccine efficacy, adverse events frequency, shelf life, thermal storage requirements, and wastage associated with single-dose versus multi-dose packaging.

Moreover, the model did not account for any potential savings in administrative overhead and inventory handling resulting from a reduction in the number of separate products included in a formulary. Lastly, the model ignored factors for which the economic impact, if any, is difficult to quantify, such as purchasing many vaccines from the same manufacturer to benefit from volume discounts, packaging similarities, color coding schemes, and brand loyalty.

One limitation of the approach taken here is the sensitivity of the maximal prices for the DTP_a-HBV-IPV combination vaccine to small changes in the monovalent vaccine prices, assumed costs of various kinds, and indications for the monovalent vaccines. Such changes would likely affect the lowest overall cost formulary and thus the resulting maximal prices for the combination vaccine. Therefore, any changes in the list of vaccines under federal contract, or changes in any cost factors used in the analysis, will require the analysis to be redone with this new data, resulting in new maximal prices for the combination vaccine. However, a change in the assumed fixed cost of a clinic visit ($40 here) does not affect this maximal price; it will only affect the total cost of immunizing a typical child.

16.5 CONCLUSIONS

The integer programming approaches discussed in this chapter were initially developed to support the needs of conventional vaccine purchasers – private pediatric and family practice clinics, HMOs, and local and state immunization programs buying vaccines under the federal contract system. These buyers face increasingly difficult procurement choices among a "combination chaos" of competing products with overlapping, non-complementary antigens [10]. Such individuals and groups could use the model (via the web-adapted version available at www.vaccineselection.com) to design their lowest overall cost formulary, using the actual prices (private sector or federal) available to them for existing and new vaccines, and adjusting other cost assumptions and parameters based on their particular health-care environment.

Reverse engineering the integer programming model provides a methodology that can most benefit vaccine manufacturers interested in developing combination vaccines and who need to understand the premium inherent in such products over their monovalent counterparts. Moreover, the federal government (the single largest purchaser of pediatric vaccines in the United States) and large HMOs can use the reverse engineered prices to assess the value of combination vaccines to their constituents, and hence decide at what price a given combination vaccine provides them with good value. Note that the reverse engineered price of a particular combination vaccine may be too low for a vaccine manufacturer to recoup its cost of developing such a product. Information from the integer programming model would be of significant value to guide investment decisions by such vaccine manufacturers, and hence avoid large research and development expenditures that may not be recouped.

The economic analysis of the DTP_a-HBV-IPV combination vaccine presented here provides a new approach to evaluating the value inherent in combination vaccines, and hence in determining whether they offer sufficient value at the price at which they are offered. The results presented here can benefit any manufacturer of a DTP_a-HBV-IPV combination vaccine by establishing the value of its vaccine based on how a particular segment of the market values the cost of an injection. Other combination vaccines that are in either phase 2 or phase 3 studies for pediatric immunization, such as DTP_a-HIB-HBV-IPV, DTP_a-HIB-IPV, and MMR-VAR [14], can also be analyzed using the models and tools discussed. Moreover, the results of such analyses can be used by very large pediatric vaccine purchasers in negotiating discounted prices for combination vaccines.

Acknowledgments

The authors wish to thank Bruce G. Weniger, M.D., M.P.H., Assistant Chief for Vaccine Development within the Vaccine Safety and Development Branch of the National Immunization Program, Centers for Disease Control and Prevention, for his encouragement for this line of research, as well as feedback on previous versions of the results discussed in this chapter. The authors also wish to thank Austral Engineering and Software, Inc. for providing access to their web site www.vaccineselection.com and for providing web support to conduct the experiments reported here, and to acknowledge Ms. Tamana Karnani for creating Figures 16.2-16.4. This research has been supported in part by the National Science Foundation (DMI-0222554, DMI-0222597). The first author has also been supported in part by the Air Force Office of Scientific Research (F49620-01-1-0007).

References

[1] Centers for Disease Control and Prevention (2002). Notice to readers: Recommended childhood immunization schedule – United States, 2002. *Morbidity* and *Mortality Weekly Report,* 51, 31-33.

[2] Madlon-Kay, D. and P. Harper (1994). Too many shots? Parent, nurse and physician attitudes toward multiple simultaneous childhood vaccinations. *Archives of Family Medicine,* 3, 610-613.

[3] Dietz, V.J., J. Stevenson, E.R. Zell, S. Cochi, S. Hadler, and D. Eddins (1994). Potential impact on vaccination coverage levels by administering vaccines simultaneously and reducing dropout rates. *Archives of Pediatric and Adolescent Medicine,* 148, 943-949.

[4] Centers for Disease Control and Prevention (2000). Preventing pneumococcal disease among infants and young children. *Morbidity and Mortality Weekly Report,* 49, 1-38.

[5] Institute of Medicine (1993). *The Children's Vaccine Initiative: Achieving the Vision.* National Academy Press, Washington, DC.

[6] Parkman, P.D. (1995). Combined and simultaneously administered vaccines: a brief history, in: Combined Vaccines and Simultaneous Administration: Current Issues and Perspectives. *Annals of the New York Academy of Sciences,* 754, 1-9.

[7] Centers for Disease Control and Prevention (1999). Combination vaccines for childhood immunization. Recommendations of the Advisory Committee on Immunization Practices (ACIP), American Academy of Pediatrics (AAP), and American Academy of Family Physicians (AAFP). *Morbidity and Mortality Weekly Report,* 48, 1-15.

[8] Decker, M.D. and K.M. Edwards (1999). Combination vaccines. In *Vaccines* (S.A. Plotkin and W.A. Orenstein, Eds.), Third Edition. W.B. Saunders Company, Philadelphia, PA, 508-530.

[9] Lieu, T.A., S.B. Black, G.T. Ray, K.E. Martin, H.R. Shinefield, and B.G. Weniger (2000). The hidden costs of infant vaccination. *Vaccine,* 19, 33-41.

[10] Weniger, B.C., R.T. Chen, S.H. Jacobson, E.G. Sewell, R. Deuson, J.R. Livengood, and W.A. Orenstein (1998). Addressing the

challenges to immunization practice with an economic algorithm for vaccine selection. *Vaccine,* 16, 1885-1897.

[11] Jacobson, S.H., E.C. Sewell, R. Deuson, and B.G. Weniger (1999). An integer programming model for vaccine procurement and delivery for childhood immunization: a pilot study. *Health Care Management Science,* 2, 1-9.

[12] Sewell, E.C., S.H. Jacobson, and B.G. Weniger (2001). Reverse engineering a formulary selection algorithm to determine the economic value of pentavalent and hexavalent combination vaccines. *Pediatric Infectious Disease Journal,* 20, S45-S56.

[13] Burden, R.L. and J.D. Faires (1997). *Numerical Analysis.* Sixth Edition. Brookes-Cole Publishing Company, New York, NY.

[14] Infectious Diseases in Children (2002). Almost 200 new drugs in development for use in children. 15, 41-44.

[15] Sewell, E.C. and S.H. Jacobson (2003). Using an integer programming model to determine the price of combination vaccines for childhood immunization. *Annals of Operations Research,* 119, 261-284.

[16] Jacobson, S.H. and E.C. Sewell (2002). Using Monte Carlo simulation to determine combination vaccine price distributions for childhood diseases. *Health Care Management Science,* 5, 135-145.

[17] LeBaron, C.W., L. Rodewald, and S. Humiston (1999). How much time is spent on well-child care and vaccination? *Archives of Pediatric and Adolescent Medicine,* 153, 1154-1159.

[18] Pellissier, J.M., P.M. Coplan, L.A. Jackson, and J.E. May (2000). The effect of additional shots on the vaccine administration process: results of a time-motion study in two settings. *American Journal of Managed Care,* 6, 41-47.

[19] Program for Appropriate Technology in Health (PATH) (1996). Cost analysis of various vaccine delivery systems. In *HealthTech Product Development Plan – MEDiVAX™ Low-workload Jet Injector,* Seattle, WA, June 5.

[20] Meyerhoff, A.S., B.G. Weniger, and R.J. Jacobs (2001). The economic value to parents of reducing the pain and emotional distress

of childhood vaccine injections. *Pediatric infectious Disease Journal,* 20, S57-S62.

[21] Kuppermann, M., R.F. Nease, L.M. Ackerson, S.B. Black, H.R. Shinefield, and T.A. Lieu (2000). Parents' preferences for outcomes associated with childhood vaccinations. *Pediatric Infectious Disease Journal,* 19, 129-133.

[22] Centers for Disease Control and Prevention (1999). Prevention of hepatitis A through active or passive immunization: recommendations of the Advisory Committee on Immunization Practices (ACIP). *Morbidity and Mortality Weekly Report;* 48, 1-37.

[23] Centers for Disease Control and Prevention (1997). Pertussis vaccination: use of acellular pertussis vaccines among infants and young children – recommendations of the Advisory Committee on Immunization Practices (ACIP). *Morbidity and Mortality Weekly Report,* 46, 1-25.

[24] Centers for Disease Control and Prevention (2000). Poliomyelitis prevention in the United States: updated recommendations of the Advisory Committee on Immunization Practices (ACIP). *Morbidity and Mortality Weekly Report,* 49,1-22.

17 ALLOCATING RESOURCES TO CONTROL INFECTIOUS DISEASES

Margaret L. Brandeau

Department of Management Science and Engineering
Stanford University
Stanford, CA 94305

SUMMARY

How can decision makers best choose among competing epidemic control programs and populations? The problem of resource allocation for epidemic control is complex, and differs in a number of significant ways from traditional resource allocation problems. A variety of OR-based methods have been applied to the problem, including standard cost-effectiveness analysis, linear and integer programming, simulation, numerical procedures, optimal control methodologies, nonlinear optimization, and heuristic approaches. This chapter reviews a number of these models. This chapter does not aim to be an exhaustive review of the literature; rather, we discuss an illustrative subset of existing models. We conclude with discussion of promising areas for further research.

KEY WORDS

Resource allocation, Epidemic control

17.1 INTRODUCTION

According to the World Health Organization, the world is facing “an infectious disease crisis of global proportions” [1]. Infectious diseases are the world’s largest killer of children and young adults, accounting for more than 13 million deaths per year and many more cases of illness [1]. The greatest number of infectious disease deaths are due to HIV/AIDS, a disease that is almost invariably fatal. In 1999, 2.8 million people worldwide died from AIDS. At least 34 million people are currently infected with HIV and the epidemic continues to grow rapidly, with no cure or vaccine in sight [2]. Diarrheal diseases, tuberculosis, malaria, measles, and influenza and pneumonia also cause significant numbers of infectious disease deaths. Some of these diseases are becoming increasingly prevalent because of HIV: the weakened immune systems of HIV-infected individuals make them prone to opportunistic infections such as tuberculosis and diarrheal diseases. In addition to existing infectious diseases, new infectious diseases, such as SARS (Severe Acute Respiratory Syndrome) continue to emerge and to spread rapidly via global travel [3].

Infectious diseases can cause death directly. They can also cause death indirectly by increasing the chance that an individual will contract other diseases such as cancer. For example, infection with human papilloma virus can lead to cervical cancer; schistosomiasis can lead to bladder cancer; and infection with Hepatitis B or C can lead to liver cancer.

Although infectious diseases pose a serious threat to public health, resources for controlling infectious diseases are limited. Decision makers must determine how to allocate limited epidemic control budgets among competing programs and populations so as to achieve the greatest health benefit given the available prevention resources.

Competing epidemic control programs may include vaccination, prevention programs, and treatment programs. Prevention programs include behavioral and nonbehavioral programs. Behavioral programs are generally aimed at inducing uninfected and/or infected individuals to change risky behavior. Nonbehavioral programs include programs for ensuring the safety of the healthcare system (e.g., universal precautions for health care workers; or screening of donated blood, organs and tissues), immigration restrictions and quarantine programs (e.g., quarantine of infected individuals), and environmental abatement programs (e.g., treatment of outdoor areas with insecticides or microbicides). Treatment programs may reduce the spread of an infectious disease by reducing the infectiousness of infected individuals.

Decision makers must also choose which populations to target with epidemic control programs. Different subgroups of a population may have different risk of infection depending on their level of exposure to the infection and their susceptibility to infection. They may also have differing propensity to change their risky behavior. Some prevention programs may be targeted to infected individuals, whereas other programs may be targeted to uninfected individuals.

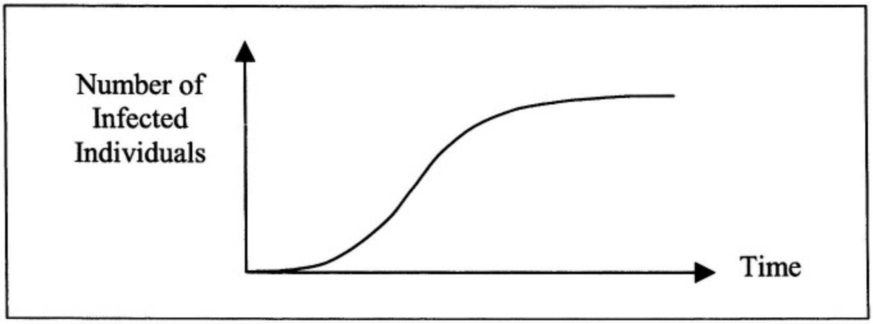
How can decision makers best choose among competing epidemic control programs and populations? As we discuss in the following section, the problem is complex, and differs in a number of significant ways from traditional resource allocation problems.

A variety of OR-based methods have been applied to the problem of resource allocation for epidemic control, including standard cost-effectiveness analysis, linear and integer programming, simulation, numerical procedures, optimal control methodologies, nonlinear optimization, and heuristic approaches. This chapter reviews a number of these models. This chapter does not aim to be an exhaustive review of the literature; rather, we discuss an illustrative subset of existing models. We conclude with discussion of promising areas for further research.

17.2 EPIDEMIC CONTROL

The problem of allocating resources for epidemic control is complex. A key reason is that epidemics of infectious disease are inherently nonlinear and dynamic: the rate of new infections is proportional to the product of the number of infected people and the number of uninfected people, and these quantities change over time. As illustrated in Figure 17.1, a typical epidemic follows an S-shaped curve. At first, few people are infected. As new people become infected, the rate of new infection increases. Eventually, as more people become infected, the growth of the epidemic slows.

Figure 17.1 Growth of a typical epidemic



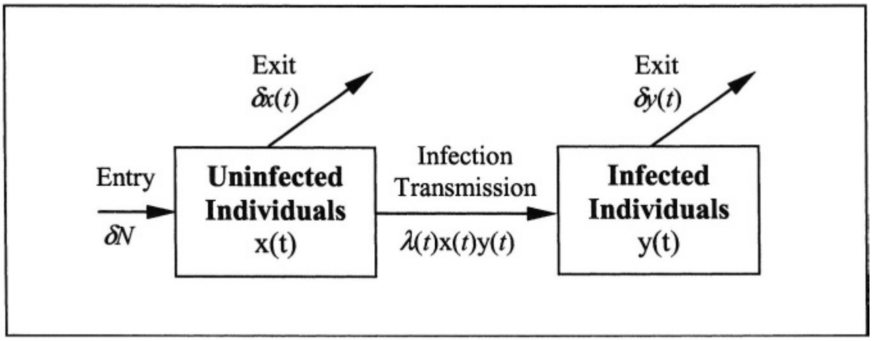
Epidemics are typically modeled using a system of nonlinear difference or differential equations. To provide some insight into epidemic dynamics and the problem of epidemic control, we present a very simple epidemic model. A comprehensive discussion of epidemic models can be found elsewhere [4, 5].

Consider the following simple compartmental model that describes the spread of an infectious disease in a single closed population. This model is illustrated schematically in Figure 17.2. Let $x(t)$ denote the number of uninfected individuals in the population at time t and let $y(t)$ denote the number of infected individuals in the population at time t (these individuals are assumed to be infectious). Every individual is either uninfected (and thus counted as part of $x(t)$) or infected (and thus counted as part of $y(t)$). Let N denote the (constant) size of the population. Let $\lambda(t)$ denote the rate of infection-transmitting contacts at time t; this rate is referred to as the sufficient contact rate. Let δ denote the rate of entry into and exit from the population. The epidemic model can be written as

$$\frac{dx(t)}{dt} = \delta N - \lambda(t)x(t)y(t) - \delta x(t) \qquad (1a)$$

$$\frac{dy(t)}{dt} = \lambda(t)x(t)y(t) - \delta y(t) \qquad (1b)$$

Figure 17.2 A simple compartmental epidemic model



The population size is constant: individuals enter the population at rate δN and leave at rate $\delta(x(t)+y(t))$, where $x(t)+y(t) = N$. Infection occurs at rate $\lambda(t)x(t)y(t)$ at time t; these individuals leave the uninfected group and enter the infected group. All infected individuals are equally likely to mix with all uninfected individuals; this is known as homogeneous mixing.

Suppose now that two programs for slowing the epidemic have been implemented, a vaccination program for uninfected individuals and an educational program that aims to eliminate risky behavior among infected individuals. Suppose that uninfected individuals are immunized at rate $\mu(t)$ at time t and that infected individuals are removed from the infectious population at rate $\gamma(t)$ at time t. Let $z(t)$ denote the number of individuals who can neither acquire nor transmit infection at time t (these are vaccinated susceptibles, and infected individuals who have been removed from the infectious population). The above model can be rewritten as

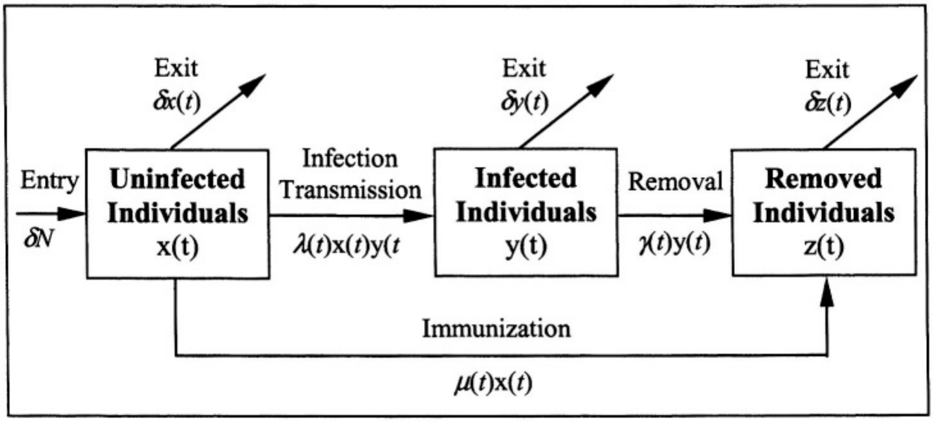
$$\frac{dx(t)}{dt} = \delta N - \lambda(t)x(t)y(t) - \delta x(t) - \mu(t)x(t) \tag{2a}$$

$$\frac{dy(t)}{dt} = \lambda(t)x(t)y(t) - \delta y(t) - \gamma(t)y(t) \tag{2b}$$

$$\frac{dz(t)}{dt} = \mu(t)x(t) + \gamma(t)y(t) - \delta z(t) \tag{2c}$$

This model is illustrated schematically in Figure 17.3. As before, the population size is constant: individuals enter the population at rate δN and leave at rate $\delta(x(t)+y(t)+z(t))$, where $x(t)+y(t)+z(t) = N$. Uninfected individuals can become infected (the terms $\lambda(t)x(t)y(t)$ in (2a) and (2b)) or immunized (the terms $\mu(t)x(t)$ in (2a)). Infected individuals can be removed from the infectious population (the terms $\gamma(t)y(t)$ in (2b) and (2c)).

Figure 17.3 A simple compartmental epidemic model with controls



The above models are quite simple. More realistic models may include features such as subdivision of the population by risk group and disease stage, nonhomogeneous mixing, different types of infectious contacts, nonconstant population size, different rates of exit/death from different groups, and stochastic parameters. Whatever the form, however, all epidemic models share the basic element of nonlinear dynamic infection transmission.

Because epidemics grow nonlinearly and different population subgroups often have different risk of infection, saving a high-risk person from infection today may save scores of people from being infected later. However, allocating all prevention resources to high-risk individuals may allow a significant epidemic to occur among low-risk individuals. Thus, the optimal resource allocation may not involve allocating all resources to high-risk individuals.

A second complexity of the resource allocation problem is the relationship between resources expended and program outcomes; we shall refer to this relationship as a "production function". In some cases, the production function for a prevention program may be linear. For example, each additional dose of a fully effective protective vaccine removes one additional person from the susceptible population. However, in many cases, a program's production function may not be linear. A program may have diminishing returns, for example, if people reached when more money is invested in a program are less likely to change their risky behavior than people reached when less is invested in the program. A prevention program may have increasing returns to scale if a minimum level of investment is necessary before the program has any significant impact on the spread of the epidemic.

Additionally, epidemic control programs may not be independent: investment in one program may change the effectiveness of other epidemic control programs. For example, a general education campaign may increase awareness of a disease and thus increase the effectiveness of other prevention programs.

Finally, the time horizon considered by the decision maker can have a significant impact on the best decision: the allocation of resources that eradicates an epidemic in the long term may not be the same as the allocation of resources that yields the maximum health benefit in the next year.

We now describe different approaches that have been used to model the problem of allocating epidemic control resources.

17.3 RESOURCE ALLOCATION MODELS

17.3.1 General resource allocation models

Resource allocation has long been studied by operations researchers [6]. For example, the classical knapsack problem can be framed as one of selecting from a set of potential investments to maximize benefit subject to a budget constraint. The problem can be formulated as an integer program or a linear program. Such formulations generally assume that the effects of investment are independent; that is, investment in one program does not affect the benefits of other programs. Linear programming formulations assume that programs are perfectly divisible (i.e., it is possible to invest in any fraction of a program) with linear returns to scale (i.e., investing in 50% of a program yields 50% of the benefit that would be obtained from full investment in the program).

These assumptions are usually not realistic for epidemic control programs. The relationship between resources invested and health benefits (such as infections averted or years of life gained) is likely to be nonlinear, as discussed above. Moreover, interventions often cannot be considered independently because the health benefits that accrue from one intervention may depend on the amount of money invested in other interventions.

17.3.2 Implicit resource allocation via cost-effectiveness analysis

Cost-effectiveness analysis, a standard tool in health economics, is a way of evaluating the costs and benefits of one health intervention compared with another [7]. For an evaluation of two interventions, A and B, the incremental cost effectiveness of B relative to A is given by:

$$(\mathrm{Costs}_B - \mathrm{Costs}_A) / (\mathrm{Effectiveness}_B - \mathrm{Effectiveness}_A)\ .$$

The above ratio expresses the cost per additional unit of health benefit conferred by intervention B compared to intervention A. Costs represent total expenditure on an intervention, plus resulting changes in health care costs. Effectiveness can be expressed in natural units of outcome such as new infections averted or years of life saved. To facilitate comparison of alternative health care investments, health outcomes can be expressed in terms of quality-adjusted life years (QALYs) lived [7].

The goal in allocating resources among health interventions is to maximize health benefits subject to available funds. Standard cost-effectiveness analyses call for spending money on those programs that yield the greatest "bang for the buck" (measured as health benefits per dollar invested) until the budget is spent. This is based on the solution to the following simple optimization problem. Let E_i denote the effectiveness that accrues from investment C_i in intervention *i*, $i = 1, \ldots, n$, and let *B* be the total available budget. Let x_i, $i = 1, \ldots, n$ denote the level of investment in intervention *i*: when x_i is 1, the maximum amount C_i is invested in intervention *i*. The problem can be written as

$$\begin{aligned} & \max \sum_{i=1}^{n} E_i x_i \\ s.t. \quad & \sum_{i=1}^{n} C_i x_i \le B \\ & 0 \le x_i \le 1 \quad i = 1, \ldots, n \end{aligned}$$

The above problem is a linear programming (LP) knapsack problem. The optimal solution is to allocate resources to interventions in order of increasing cost-effectiveness ratios (C_i/E_i) until the budget is spent.

The solution to the above LP knapsack problem is the optimal resource allocation only if the following three conditions hold: (1) The interventions are perfectly divisible (i.e., it is possible to invest in any fraction of a program). (2) The interventions have constant returns to scale: thus, for example, doubling the investment in a particular intervention doubles the health benefits that accrue. (3) The interventions are independent: investment in one intervention does not change the incremental cost effectiveness of any other interventions. Because these conditions are not likely to be met for an epidemic control problem, standard cost-effectiveness analysis has limited applicability for such problems.

17.3.3 Linear programming approaches

Several authors have proposed linear and mixed integer linear programming formulations for the resource allocation problem that do not require the first and second conditions (divisibility and constant returns to scale). If interventions are not divisible, the above LP can be replaced with an integer program (IP) where the decision variables $\{x_i\}$ are constrained to be either 0 or 1 [8, 9]. If an intervention is partially divisible – for example, if investment in a program is constrained to one of several discrete levels, or if investment in a program is constrained to be either nothing or at least p% of full investment – then the constraint $0 \le x_i \le 1$ can be replaced with integer constraints or mixed integer linear constraints [10]. If interventions do not have constant returns to scale – for example, if a minimum level of investment is required before any health benefits can be realized, or if programs have increasing or decreasing returns to scale – the optimization problem can be modified by introducing integer indicator variables (to include fixed cost) and/or by using piecewise linear approximations of the cost-effectiveness functions (to include nonconstant returns to scale) [10]. However, such formulations do not capture the nonlinearities in cost effectiveness caused by epidemic dynamics, nor do they capture possible interactions between investment in one intervention and the effectiveness of other interventions.

17.3.4 Optimal control approaches

An alternative approach that does capture epidemic nonlinearity applies optimal control to an epidemic model. Such an analysis determines the optimal application of epidemic control over time. The goal is to obtain analytical results that characterize the form of the optimal solution. Thus, most analyses consider a single epidemic control program applied in a single population. Closed-form expressions for the compartment size functions (e.g., the functions $x(t)$ and $y(t)$ in equations (1)) are only known for certain very simple epidemic models [4, 5]. Thus, optimal control analyses use relatively simple epidemic models such as that in equations (2), but usually assuming only one type of control. Examples of controls typically considered include vaccination (which increases the rate $\mu(t)$ in (2)), treatment or removal of infectious persons (which increases the rate $\gamma(t)$ in (2)), and programs aimed at reducing the sufficient contact rate (the rate $\lambda(t)$ in (2)).

A typical objective in the application of such control might be to minimize the cost of control (e.g., variable vaccination cost plus the fixed cost of establishing the vaccination program) plus a cost associated with the number of individuals who become infected (e.g., treatment cost). Except for the fixed cost of establishing the control program, costs are usually assumed to be linear: the cost of control is a constant multiplied by the change in the value of the affected parameter, and the cost of disease is a constant multiplied by the number of people who become infected. Use of simple epidemic models and a linear cost function allows for characterization of the form of the optimal solution(s): for example, vaccination of susceptible individuals until the disease prevalence is reduced below a certain level.

Such an approach has been applied to quarantine and removal programs (e.g., [11, 12]), vaccination programs (e.g., [13, 14]), and other epidemic control programs (e.g., [15-18]). The optimal control approach provides an elegant solution, but has limited applicability since it is generally limited to a single epidemic control program in a single population.

17.3.5 Equilibrium analysis

If the goal of the control program is complete disease eradication or to optimize some function of the long-term state of the epidemic (and a sufficiently long time horizon is considered), then one need not consider short-term epidemic dynamics but only the long-term epidemic equilibrium. (Disease eradication corresponds to an equilibrium state with no infected individuals.) This allows for the use of more complex epidemic models than those used in the optimal control approach. For example, some authors have based equilibrium analyses on epidemic models with different age groups or different risk of becoming infected.

A typical equilibrium analysis determines the minimum level of control (for example, the minimum number of individuals in each population group who must be vaccinated) such that the disease is eradicated (e.g., [19, 20]). One analysis determined the amount of a fixed Influenza A vaccine to distribute among different age groups to optimize a function of the final (equilibrium) state of the epidemic [21]. Two different objective functions were considered: that of minimizing expected epidemic costs (health costs, costs of lost wages, and costs of early mortality) and that of minimizing expected years of life lost due to early mortality. The optimal vaccine distribution was found via a numerical search procedure.

The equilibrium approach is limited by the assumption of a time horizon sufficiently long for equilibrium to be reached and by the assumption that the goal of the epidemic control program is to minimize some function of the equilibrium state of the epidemic.

17.3.6 Simulation analysis and numerical procedures

Much of the above described work is theoretical in nature: one can generate insights into the structure of the optimal control for a single epidemic control program, but the assumptions underlying the analyses are quite limiting. An alternative approach that generally requires fewer limiting assumptions is to consider a finite set of resource allocation alternatives and simulate their effects using a more realistic epidemic model. Although the results may not be transferable to other populations or other epidemics, they are likely to be useful for the specific epidemic and population considered. Simulation has been applied, for example, to determine the most effective programs for HIV control in different regions of Africa [22, 23]. Those analyses compare the effects (in terms of reduced HIV prevalence) of different combinations of interventions, but do not explicitly consider the cost of the interventions.

Some simulation analyses use compartmental epidemic models (e.g., such as those in equations (1) and (2)) in which the population is divided into mutually exclusive, collectively exhaustive compartments. Other analyses simulate an epidemic by simulating the health state of each individual in the population. This latter approach requires significantly more computation than the compartmental model approach.

A related approach for solving the resource allocation problem is the use of numerical procedures in conjunction with an epidemic model. One analysis evaluated the effectiveness of six different methods for preventing the spread of gonorrhea [24]. Effectiveness was measured as reduction in equilibrium gonorrhea prevalence among women. The authors numerically analyzed a compartmental epidemic model to determine the equilibrium prevalence associated with each control program. The authors did not consider cost of the programs in the analysis, but mentioned that program costs and benefits would have to be compared before the best control program could be chosen. Another analysis used a simple epidemic model to evaluate the impact of targeting an entire HIV prevention budget to different (noninteracting) populations [25]. The prevention programs were assumed to have linear production functions. The author showed that targeting prevention funds to high-risk populations could avert significantly more HIV infections than targeting funds to low-risk populations.

17.3.7 Optimization approaches

Use of simulation for solving the resource allocation problem limits the analysis to consideration of a finite set of alternatives. Use of equilibrium analysis limits the solution to a sufficiently long time horizon. A variety of optimization approaches that overcome these limitations have been developed. The optimization approaches usually employ either a very simple epidemic model or an approximation of a more complex epidemic model.

One analysis considered the optimal application of three types of programs to control tuberculosis (vaccination, prophylaxis, and therapy) [26]. The epidemic was modeled by a compartmental model with nine compartments. The epidemic equations were approximated by linear equations for each year in the time horizon. A schedule was set for the reduction in active cases each year. The goal was to determine the allocation of resources that achieves the schedule at lowest cost. The problem was formulated and solved as an LP.

Some authors have developed optimization models for the resource allocation problem that allow for the possibility of nonlinear production functions; thus, parameters describing the epidemic (e.g., the sufficient contact rate) can change nonlinearly as a function of investment in an epidemic control program. These models have the following general form. Assume that n different prevention programs are available, with total funds B that can be invested. Let v_i denote the investment that is made in prevention program i, $i=1, \ldots, n$, and let $\mathbf{v} = (v_1, v_2, \ldots, v_n)$. Denote the upper limit on investment in program i by V_i, $i = 1, \ldots, n$. Let IA($\mathbf{v}$) denote the number of infections that are averted over the time horizon of the problem given investment $\mathbf{v}$, and let QALY($\mathbf{v}$) denote the number of quality-adjusted life years (QALYs) gained over the time horizon of the problem given investment $\mathbf{v}$. These functions are determined from the epidemic model and from the production functions that describe how parameters of the epidemic model change in response to investment in the prevention programs. In some cases it may not be possible to write these functions in closed form, since many epidemic models do not have a closed-form solution.

The resource allocation problem can be written as

$$\begin{aligned} \max_{\mathbf{v}} \quad & IA(\mathbf{v}) \text{ or } QALY(\mathbf{v}) \\ s.t. \quad & \sum_{i=1}^{n} v_i \leq B \\ & 0 \leq v_i \leq V_i \quad i=1,\ldots,n \end{aligned}$$

The objective is to allocate a fixed budget among prevention programs to maximize either infections averted or QALYs gained, subject to upper limits on investment in each program. (The general formulation also allows for nonzero lower limits on investment in each program, if desired.) The models described below differ in their assumptions about the epidemic model, the production functions, and the timing of investment in the prevention programs. However, all have the same general form: maximization of a nonlinear health-benefits function subject to linear constraints on investment.

One analysis considers the allocation of epidemic control resources to multiple non-interacting populations [27]. The epidemic in each population is described by a simple epidemic model with two compartments (susceptibles and infectives). Resources spent combating the epidemic reduce the sufficient contact rate of the disease. A separate prevention program is available for each population. Each prevention program reduces the sufficient contact rate according to a general production function. The authors developed analytical results characterizing the optimal solution for both objective functions (maximizing IA(**v**) or QALY(**v**)) for the case of a sufficiently long time horizon.

A slightly more complex epidemic model was used in determining the optimal allocation of HIV prevention resources between two independent populations (injection drug users and non-users) [28]. Using data from a local hospital, the authors estimated production functions for three different prevention programs (HIV testing with routine counseling, HIV testing with intensive counseling, and methadone maintenance for injection drug users). The goal was to maximize the number of HIV cases averted over a fixed time horizon. The optimal resource allocation was determined via a numerical search procedure.

In other work, a more comprehensive optimization framework was developed that allows for a general compartmental epidemic model with interacting populations and interacting prevention programs [29]. The authors showed that, for both objective functions (maximizing IA(**v**) or QALY(**v**)), for the special case of linear production functions and first-order approximations of the compartment size functions, the problem reduces to a knapsack LP that has a greedy solution. The authors presented several heuristics for solving the general resource allocation problem, and showed that they yield near-optimal solutions.

The general framework [29] was applied to determine the allocation of a limited budget among three types of HIV prevention programs (needle exchange programs, methadone maintenance treatment, and condom availability programs) in a population of injection drug users and non-users [30]. The analysis allowed for interacting populations (injection drug users could acquire HIV from non-users and vice versa) and interacting prevention programs (the effectiveness of a prevention program could depend on how much had been invested in the other programs). The optimal resource allocation for each objective (maximizing IA(**v**) and QALY(**v**)) was determined using the heuristic methods previously developed [29]. The authors showed that simpler allocation methods (for example, allocation of resources to population groups based on HIV incidence) might lead to allocations that do not yield the maximum health benefit.

The above analyses [27-30] assume that resources are allocated at the beginning of the time horizon and that the allocation cannot be changed during the time horizon. This assumption was relaxed in later work [31], which allowed for a limited epidemic control budget to be allocated over multiple time periods, with funds allocated at the beginning of each time period. For certain special cases with two time periods, multiple independent populations, and linear production functions, the authors showed that the optimal solution involves investing in each period as much as possible in some of the populations and nothing in all the other populations. They presented heuristic algorithms for solving the general problem, and showed that good allocations can be made based on some fairly simple heuristics. The authors also showed that allowing for some reallocation of resources over the time horizon of the problem, rather than allocating resources just once at the beginning of the time horizon, can lead to significant increases in health benefits. They concluded that allowing for reallocation of funds might generate more health benefits than use of a sophisticated model for one-time allocation of resources.

17.3.8 Heuristic approaches and tools for decision makers

Many of the above models, particularly the optimization models described in Section 17.3.7, involve the development and use of nonlinear dynamic models and the application of sophisticated optimization techniques. Such models may not be readily accepted by practitioners, and often require data that are unavailable or difficult to obtain. An alternate stream of research aims to develop resource allocation models that are simple for practitioners to understand and implement. Much of this work has been done in the context of HIV prevention.

A recent report from the Institute of Medicine [32] suggested the following simple model for determining the allocation of HIV prevention resources that maximizes the number of infections averted. Assume that a number of different HIV prevention programs are available that can target risk groups, indexed by *j*, in different geographic areas, indexed by *i*. Let v_{ij} denote the amount of money invested in the program targeted to risk group *j* in geographic area *i*. Let n_{ij} denote the number of people in risk group *j* and geographic area *i*, and let $I_{ij}(0)$ denote the baseline number of new HIV infections that will occur in that group over the time horizon of the problem in the absence of additional investment in prevention. Let c_j denote the cost per person of an intervention that targets risk group *j*, f_j the maximum fraction of risk group *j* that can be reached with such an intervention, and e_j the percentage reduction in the rate of new HIV infections for those in programs targeted to risk group *j*. The model can be written as

$$\max_{\mathbf{v}} \quad \sum_i \sum_j \frac{v_{ij}}{c_j} \frac{I_{ij}(0)}{n_{ij}} e_j$$

$$\text{s.t.} \quad \frac{v_{ij}}{c_j} \le f_j n_{ij}$$

$$\sum_i \sum_j v_{ij} \leq B$$

$$0 \leq v_i \leq V_i \qquad i = 1,\ldots,n$$

The objective is to maximize the number of infections averted, which in group *(i,j)* is the product of the number of people reached by the prevention program targeted to that group (v_{ij}/c_j) multiplied by the number of infections averted per program participant $(I_{ij}(0)e_j/n_{ij})$. Investment is constrained by the number of people who could possibly be reached in each group $(f_j n_{ij})$ as well as the overall budget constraint. The model is a knapsack LP that has a greedy solution.

The above formulation assumes that the number of new infections that will occur over the time horizon of the problem in the absence of investment $(I_{ij}(0))$ is known exogenously, that populations and programs are independent, and that the production functions for the prevention programs are linear. An alternative formulation allows for general production functions [33]. We drop the subscript *j,* and assume that target populations and prevention programs are indexed by *i* only. Let $\alpha_i(v_i)$ denote the fraction of infections in population *i* that will be averted by investment v_i in the prevention program targeted to population *i.* The model can be written as

$$\max_{\mathbf{v}} \quad \sum_{i=1}^{n} I_i(0)\alpha_i(v_i)$$

$$s.t. \quad \sum_{i=1}^{n} v_i \leq B$$

$$v_i \geq 0 \qquad i = 1,\ldots,n$$

If desired, nonzero lower limits on investment in each program can be included. This model is a knapsack problem that can be solved using nonlinear optimization techniques or dynamic programming.

Another formulation requires linear production functions but allows for a general epidemic model, interacting populations and prevention programs, and either objective function (maximizing infections averted or QALYs gained) [29]. The model is created by approximating the general resource allocation problem [29] using first-order approximations for the compartment size functions in the epidemic model. The resulting problem is a knapsack LP of the form

$$
\begin{aligned}
\max_{\mathbf{v}} \quad & \sum_{i=1}^{n} d_i v_i \\
\text{s.t.} \quad & \sum_{i=1}^{n} v_i \le B \\
& v_i \ge 0 \qquad i = 1,\ldots,n
\end{aligned}
$$

where, as before, v_i is the level of investment in program i. The coefficient d_i is an approximate measure of the number of infections that will be averted (or the number of QALYs that will be gained) in population i per dollar invested in the prevention program that reaches population i. Although the coefficients d_i are calculated from a first-order approximation of the epidemic model, they reflect more than just a linear incidence rate [29]. This model has a greedy solution that can be determined from a simple ordering of the objective function coefficients.

17.4 CONCLUSIONS AND FUTURE RESEARCH

Many of the resource allocation models described in this chapter are grounded in theory from welfare economics and cost-effectiveness analysis: they aim to maximize health benefit subject to a budget constraint. In practice, the problem of allocating resources to control epidemics is more complex. Policy makers may face political and social objectives. These may include achieving equity among population groups or programs, targeting resources to underserved populations, restricted access to certain programs, and “earmarking” of funds from different sources (e.g., see [34]). Further work could identify important considerations in real-world resource allocation problems and extend existing models to reflect such situations.

Policy makers need accessible models for making resource allocation decisions. Section 17.3.8 described several simple models that have been developed for solving the problem [29, 32, 33]. An important next step is to translate simple models of that type into tools that can help decision makers make informed resource allocation decisions. Such a tool might take the form of a spreadsheet model that users could tailor to their own particular needs by specifying parameter values – for example, parameters that describe the target populations, the epidemic that is to be controlled, and risk behaviors within those populations, and parameters that describe the production functions of the prevention programs. Ideally such a model would determine the allocation of resources that maximizes health benefits, with and without constraints on expenditure, so that decision makers could understand the cost (in terms of foregone health benefits) of different social constraints.

Determining how best to allocate resources for epidemic control can be difficult. The dynamics of epidemic growth in the population at large and in various population subgroups may be complex. Policy makers must typically choose between competing epidemic control programs (e.g., vaccination, prevention, and treatment programs) that can be targeted to different population subgroups. The relationship between resources expended and program outcomes may not be linear. The epidemic control programs may not be independent. In addition to the goal of maximizing health benefits, social considerations such as equity may be important.

Recently the U.S. government approved a five-year, $15 billion global AIDS package aimed at combating AIDS, particularly in Africa. The government must determine how this money should be allocated between treatment of HIV-infected individuals, hospice care for those dying of AIDS, and prevention. A recent news report indicated that, in trying to agree on how to spend the money, members of Congress could not "get past basic questions such as whether it's more important to advise people to abstain from risky sex or to give them condoms" [35]. Moreover, some of the money would be allocated to the United Nations and World Health Organization's Global Fund to Combat AIDS, Tuberculosis, and Malaria, which some conservative groups object to because they say it supports groups that condone abortion [35].

Despite the difficulty of making decisions about allocating epidemic control resources, decisions must be made. Infectious diseases are a critical public health problem worldwide. OR-based models can help determine the allocation of resources that maximizes health benefits, thus providing important input to such decisions.

Acknowledgments

This work was supported by a grant from the National Institute on Drug Abuse, National Institutes of Health (R-01-DA-15612).

References

[1] World Health Organization (1999). *Removing Obstacles to Healthy Development,* http://www.who.int/infectious_disease_report, Accessed November 21, 2002.

[2] World Health Organization (2000). *Report on the Global HIV/AIDS Epidemic - June 2000,* http://www.unaids.org/epidemic_updated/report, Accessed November 21, 2002.

[3] World Health Organization (2003). *Severe Acute Respiratory Syndrome (SARS) - multi-country outbreak - Update 26,* World Health Organization, http://www.who.int/csr/don/2003_04_10/en/, Accessed April 10, 2003.

[4] Anderson, R.M. and R.M. May (1991). *Infectious Diseases of Humans: Dynamics and Control.* Oxford University Press, Oxford.

[5] Bailey, N.J. (1975). *The Mathematical Theory of Infectious Diseases and its Applications.* Hafner Press, New York.

[6] Hillier, F.S. and G.J. Lieberman (1993). *Operations Research.* Holden-Day, New York.

[7] Gold, M.R., J.E. Siegel, L.B. Russell, and M.C. Weinstein, Eds. (1996). *Cost-effectiveness in Health and Medicine.* Oxford University Press, New York.

[8] Torrance, G.W., W.H. Thomas, and D.L. Sackett (1972). A utility maximization model for evaluation of health care programs. *Health Services Research,* 7, 118-133.

[9] Birch, S. and A. Gafni (1992). Cost effectiveness/utility analyses. Do current decision rules lead us to where we want to be? *Journal of Health Economics,* 11, 279-296.

[10] Stinnett, A.A. and A.D. Paltiel (1996). Mathematical programming for the efficient allocation of health care resources. *Journal of Health Economics,* 15, 641-653.

[11] Sethi, S.P. (1978). Optimal quarantine programmes for controlling an epidemic spread. *Journal of the Operational Research Society,* 29, 265-268.

[12] Greenhalgh, D. (1988). Some results on optimal control applied to epidemics. *Mathematical Biosciences,* 88, 125-158.

[13] Muller, J. (1998). Optimal vaccination patterns in age-structured populations. *SIAM Journal of Applied Mathematics,* 59, 222-241.

[14] Greenhalgh, D. (1986). Control of an epidemic spreading in a heterogeneously mixing population. *Mathematical Biosciences,* 80, 23-45.

[15] Sethi, S.P. and P.W. Staats (1978). Optimal control of some simple deterministic epidemic models. *Journal of the Operational Research Society,* 29, 129-136.

[16] Sethi, S.P. (1974). Quantitative guidelines for communicable disease control programs: A complete synthesis. *Biometrics,* 30, 681-691.

[17] Blount, S., A. Galamboski, and S. Yakowitz (1997). Nonlinear and dynamic programming for epidemic intervention. *SIAM Journal of Applied Mathematics and Computation,* 86, 123-136.

[18] Behncke, H. (2000). Optimal control of deterministic epidemics. *Optimal Control Applications and Methods,* 21, 269-85.

[19] Hethcote, H.W. and J.W. Van Ark (1987). Epidemiological models for heterogeneous populations: Proportionate mixing, parameter estimation, and immunization programs. *Mathematical Biosciences,* 84, 85-118.

[20] May, R.M. and R.M. Anderson (1984). Spatial heterogeneity and the design of immunization programs. *Mathematical Biosciences,* 72, 83-111.

[21] Longini, I.M., E. Ackerman, and L.R. Elveback (1978). An optimization model for influenza A epidemics. *Mathematical Biosciences,* 38, 141-157.

[22] Bernstein, R.S., et al. (1998). Simulating the control of a heterosexual HIV epidemic in a severely affected East African city. *Interfaces,* 28, 101-126.

[23] Robinson, N.J., D.W. Mulder, B. Auvert, and R.J. Hayes (1995). Modelling the impact of alternative HIV intervention strategies in rural Uganda. *AIDS,* 9, 1263-1270.

[24] Hethcote, H.W. (1982). Gonorrhea modeling: A comparison of control methods. *Mathematical Biosciences,* 58, 93-109.

[25] Kahn, J.G. (1996). The cost-effectiveness of HIV prevention targeting: How much more bang for the buck? *American Journal of Public Health,* 86, 1709-1712.

[26] ReVelle, C., F. Feldmann, and W. Lynn (1969). An optimization model of tuberculosis epidemiology. *Management Science,* 16, B190-B211.

[27] Brandeau, M.L., G.S. Zaric, and A. Richter (2003). Optimal resource allocation for epidemic control among multiple independent populations: Beyond cost effectiveness analysis. *Journal of Health Economics*, 22, 575-598.

[28] Richter, A., M.L. Brandeau, and D.K. Owens (1999). An analysis of optimal resource allocation for HIV prevention among injection drug users and nonusers. *Medical Decision Making,* 19, 167-179.

[29] Zaric, G.S. and M.L. Brandeau (2001). Resource allocation for epidemic control over short time horizons. *Mathematical Biosciences,* 171, 33-58.

[30] Zaric, G.S. and M.L. Brandeau (2001). Optimal investment in a portfolio of HIV prevention programs. *Medical Decision Making,* 21, 391-408.

[31] Zaric, G.S. and M.L. Brandeau (2002). Dynamic resource allocation for epidemic control in multiple populations. *Mathematical Medicine and Biology,* 19, 235-255.

[32] Ruiz, M., et al., Eds. (2001). *No Time to Lose: Getting More from HIV Prevention.* National Academy Press, Washington, DC.

[33] Kaplan, E.H. (1998). Economic evaluation and HIV prevention community planning: A policy analyst's perspective. In *Handbook of Economic Evaluation of HIV Prevention Programs,* D.R. Holtgrave, Ed., Plenum Press, New York.

[34] Johnson-Masotti, A.P., S.D. Pinkerton, D.R. Holtgrave, R.O. Valdiserri, and M. Willingham (2000). Decision making in HIV prevention community planning: An integrative review. *Journal of Community Health,* 25, 95-112.

[35] Abrams, J. (2003). Bickering threatens to delay approval of global AIDS bill. *Seattle Times,* Seattle, April 12, A12.

18 MICROBIAL RISK ASSESSMENT FOR DRINKING WATER

Stephen E. Chick[1], Sada Soorapanth[2] and James S. Koopman[3,4]

[1] INSEAD
Technology Management Area
Fontainebleau, France

[2] Department of Industrial and Operations Engineering
University of Michigan
Ann Arbor, MI 48109

[3] Department of Epidemiology
University of Michigan
Ann Arbor, MI 48109

[4] Center for the Study of Complex Systems
University of Michigan
Ann Arbor, MI 48109

SUMMARY

Infectious microbes can be transmitted through the drinking water supply. Recent research indicates that infection transmission dynamics influence the public health benefit of water treatment interventions, although some risk assessments currently in use do not fully account for those dynamics. This chapter models the public health benefit of two interventions: improvements to centralized water treatment facilities, and localized point-of-use treatments in the homes of particularly susceptible individuals. A sensitivity analysis indicates that the best option is not as obvious as that suggested by an analysis that ignores infection dynamics suggests. Deterministic and stochastic dynamic systems models prove to be useful tools for assessing the dynamics of risk exposure.

KEY WORDS

Microbial risk, Epidemic model, Water treatment, Stochastic infection model, Ornstein-Uhlenbeck process

18.1 INTRODUCTION

A cryptosporidiosis outbreak linked to *Cryptosporidium* oocysts in Milwaukee's drinking water caused over 400,000 cases of diarrhea and 1,000 hospitalizations in 1993. The outbreak played a role in the death of more than 50 individuals, primarily individuals with AIDS [1, 2]. The World Health Organization indicates that disease caused by these and other waterborne microbes is involved in the death of millions of people every year [3], and the illness of many more. One culprit is the lack of a safe water supply and basic sanitation [4]. Endemic infection is a significant concern, not just outbreaks.

This chapter reviews recent progress in merging infection transmission models with microbial risk assessments. The goal of that work is to better represent the dynamics of infection in such risk assessments. This chapter also presents sensitivity analyses that provide policy regions that indicate when it is better to use centralized water treatment alternatives versus local water treatment measures, as a function of infection transmission parameters. We also discuss how different model structures, including ordinary differential equations (ODEs), stochastic Markov chains of individual infection and recovery events, and Ornstein-Uhlenbeck (OU) diffusion approximations may be useful for policy region assessment and inference for parameters whose values are poorly understood.

Although the focus of this chapter is risk assessment for water treatment interventions and their public health consequences, the idea of modeling risk exposure as a dynamic function of a system's state is rather general. Other microbial applications include the protection of the food supply chain, and biological warfare preparedness. Specific issues that have attracted public interest recently include so-called "mad cow disease" (bovine spongiform encephalopathy) and the threat of anthrax and smallpox attacks. *E. coli* and Norwalk-like viruses can be found in both the water system and the food chain [5]. The importance of dynamics for risk exposure assessments is not exclusive to infectious diseases. For example, weather dynamics can influence risk exposure to radiation in the aftermath of nuclear accidents [6]. The need for dynamic systems models of risk exposure, then, has a much wider application than the scope presented here, and the tools available to approximate those exposures continue to be developed.

Drinking water can be protected from microbes with a series of barriers starting with source water protection, centralized municipal water treatment, filters or other local point-of-use treatments, and wastewater treatment. Centralized drinking water treatments improve water quality for the entire community. Options include filtration, chlorination, and ozone pretreatment.

Ozone pretreatment may reduce *Cryptosporidium* oocysts in water by 40-60%, but may be quite costly. A facility for a particular California reservoir is estimated to cost $154-190 million initially, and $3.8-5.2 million per year thereafter [7]. Local treatment can also be used for population subgroups that require particularly effective pathogen removal. Options include copper-silver ionization and chlorine dioxide generation in hospitals and nursing homes [8], and reverse osmosis filters in the homes of immunocompromised individuals. Such filters may costs hundreds of dollars per home, and require regular maintenance. These costs justify a formal assessment of the public health benefit of each treatment option.

A standard approach to risk assessment for chemicals and microbes is to identify hazards, quantify occurrence and exposure, assess the dose-response relationship, and identify human health consequences. Exposure is generally taken to be from drinking water in this context. The probability of infection is assessed with a dose-response curve, where dose is a function of microbes in consumed water. The health effects of any resulting disease are then quantified. But microbes present additional risk exposures that chemicals do not usually exhibit. Microbes can circulate through two secondary transmission routes: interpersonal human contact, and a water loop where infected individuals recontaminate water through recreational use or waste [9, 10].

Some analyses (e.g., [7]) account for secondary transmission in the water loop by using the prevalence of infection in the population to assess the amount of microbes shed into recreational water, then estimating increased contamination in drinking water. That approach is consistent with risk calculations used by the Environmental Protection Agency (EPA) [11]. Such an approach does not fully model the fact that effective water treatment changes the prevalence of infection, which is an input to the assumed risk exposure model. Although this indirect effect of treatment on risk exposure due to secondary transmission is not modeled by that approach, there is a recognized need to do so to inform water treatment policy [12].

Other analyses [13, 14] use dynamic systems models to represent the dynamics of risk exposure. The models are based on deterministic ordinary differential equations (ODEs). Such models find that the public health benefit of water treatment interventions depends strongly on how infection is circulated. For example, Milwaukee residents with AIDS suffered particularly extreme consequences from cryptosporidiosis during the 1993 epidemic [15]. Some have proposed that highly effective filters that eliminate *Cryptosporidium* oocysts would effectively protect individuals with AIDS from similar risks in the future. This would be the case if there were no additional exposure from secondary transmission to that subgroup from human contact. However, it is likely that some secondary transmission occurred [1, 13]. Depending on the average number of secondary transmissions, and the relative probability of infection given exposure for those with AIDS, improving a standard municipal facility by adding ozone pretreatment may be more effective than filters [14]. If secondary transmission is sufficiently high, ozone can reduce secondary exposure in the AIDS subgroup by reducing cryptosporidiosis prevalence in the general population – and that reduction can outweigh the benefits of completely effective filters on the water taps of individuals with AIDS.

Section 18.2 extends previous work [14] by presenting a sensitivity analysis for those policy regions (ozone pretreatment versus local filters) with respect to several infection parameters, and by using a more refined model of the natural history of infection of cryptosporidiosis. The policy region is quite sensitive to the efficacy of ozone for inactivating oocysts, but is not very sensitive to the size of the sensitive population subgroup, as long as it is not too large, nor to the rate of exogenous introduction of oocysts into the water supply. Ozone becomes less effective, relative to filters in the susceptible subpopulation, as the water loop becomes relatively more important for secondary transmission than human contact.

Deterministic infection models ignore variability that arises in real infection transmission systems. Further, standard ODE parameter fitting tools make normal distribution assumptions that may not be satisfied in practice [16]. Stochastic infection models explicitly account for this variability, and may provide a mechanism to further incorporate infection dynamics into the parameter inference process. Parameter inference is important because the secondary transmission parameters for a number of microbes of interest to the EPA are poorly understood at present. Several researchers have examined mechanisms to infer parameters of various infection models given outbreak or intervention trial data [13, 17-19], or using endemic data [16]. Those works attempt to incorporate the dynamics of infection into the likelihood model using a variety of approximations (e.g., binomial distributions for discrete-time models, normal approximations for larger populations using moment methods).

Section 18.3 extends that work by suggesting that diffusion process approximations be used to model the stochastic infection dynamics. The idea is to apply stochastic process results [20-23] to approximate the underlying discrete-state Markov chain model of infection and microbe contamination with a continuous-state Ornstein-Uhlenbeck (OU) process. We present diffusion approximation formulas for the stationary mean and covariance of the underlying infection model. Section 18.4 describes potential areas for further research for water treatment policy, risk analysis, and epidemic modeling research.

18.2 ODE MODELS TO EVALUATE POLICY REGIONS

Our goal is to develop a mathematical model that captures the dynamics of three modes of infection transmission: infection from microbes in the drinking water that come from exogenous sources, secondary infection from microbes in drinking water that result from contamination of source water from modeled individuals, and secondary transmission from human-to-human contact. The model must account for multiple subgroups with different infection susceptibility and outcome parameters, and further allow for the assessment of public health benefits of both local and municipal level interventions. We first describe an ODE infection model. Many parameters are not well understood for most microbes on the EPA's Candidate Contaminant List. We therefore present a sensitivity analysis that could be applied for those agents. The analysis here is consistent with current knowledge about cryptosporidiosis.

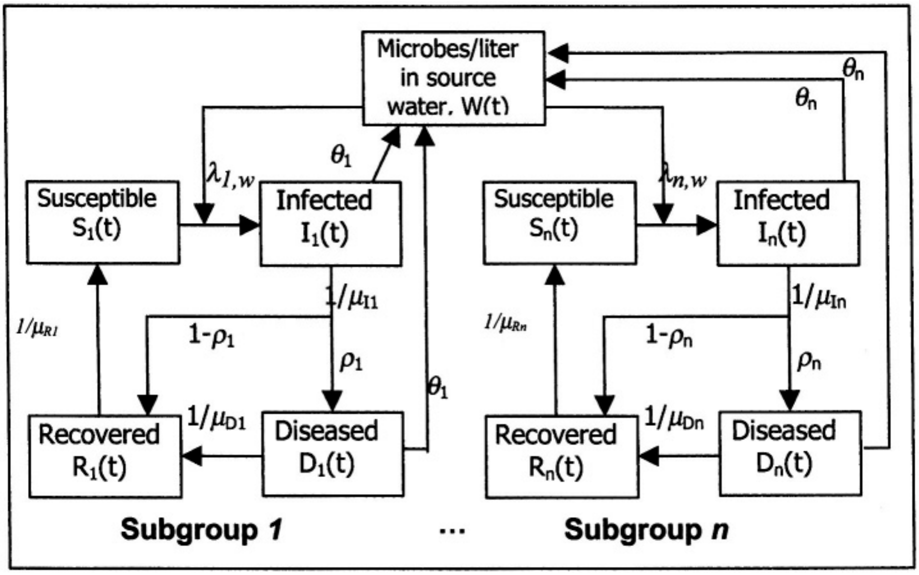
18.2.1 Deterministic infection transmission system model

Figure 18.1 illustrates that humans are assumed to change health status from susceptible (S), infected (I), diseased (D), and recovered (R) as a result of microbial infection. Microbes can be shed by infected and diseased individuals into the water supply, which in turn can reinfect susceptible individuals. We further assume that there are *n* different subgroups that interact according to a proportional mixing pattern [24]. Individuals in different subgroups may have different mixing and infection parameters. Here we are particularly interested in the case of two subgroups: immunocompetent and immunosuppressed individuals. A more detailed study might also model special characteristics of the young and the aged.

The N_i individuals in subgroup i are counted as to whether they are infected $I_i(t)$ (infectious, but asymptomatic), diseased $D_i(t)$ (infectious and symptomatic), recovered $R_i(t)$ (temporarily immune to reinfection), or susceptible $S_i(t)$. These values vary through time as the system evolves. For simplicity, the argument t is dropped below except when we wish to emphasize dependence of these values on time.

Microbe concentration in the water supply, $W(t)$, shown in the upper portion of the figure, is influenced by the rate γ of exogenous introduction of microbes, the rate α that microbes leave the system from water flow or inactivation, and the rate θ_i that infected individuals contaminate the water supply. This leads to the microbe concentration dynamic in equation (1).

Figure 18.1 An SIDRS/W infection model with water loop and proportional mixing



$$\frac{dW}{dt}=\gamma-\alpha W+\sum_{i=1}^{n}\theta_i(I_i+D_i)\,. \tag{1}$$

Each susceptible individual in subgroup i has the potential of becoming infected after being exposed. The rate of exposure for each susceptible individual depends on two main sources. Exposure from water consumption is determined by the number of microbes per unit volume in the source water, W, the fraction of microbes that remain after treatment τ_i, the volume of drinking water consumed per day ϕ_i, and the probability of infection per ingested microbe, r_i. Exposure from secondary transmission depends on the number of individuals in each subgroup, N_j, the number of contacts per day, c_j, and the probability β_i that a potentially infectious contact will infect an individual in subgroup i. This results in an overall exposure rate $\lambda_{i,W}$ for subgroup i, when the microbe concentration is W.

$$\lambda_{i,W}=r_i\tau_i\phi_iWS_i+\sum_{j=1}^{n}c_j(I_j+D_j)\frac{c_iN_i}{\sum_{k=1}^{n}c_kN_k}\frac{S_i}{N_i}\beta_i \tag{2}$$

The first term models exposure from drinking water. The second term sums the exposures from each subgroup to susceptibles in i: there are $c_j(I_j+D_j)$

potentially infectious contacts, of which a fraction $c_iN_i \,/\, \Sigma\, c_kN_k$ are with members of group *i*. The probability that a member of subgroup *i* is susceptible is S_i/N_i, and the probability of infection given the contact is β_i.

After becoming infected, only a fraction ρ_i become diseased; the rest recover and become immune for some duration of time, μ_{Ri}. The mean duration of infection is μ_{Ii}, and the mean duration of disease is μ_{Di}. Since the dynamics of microbial infection are on a much faster time scale than the lifetimes of humans, we assume a closed population.

$$\begin{aligned} dS_i/dt &= -\lambda_{i,w} + R_i/\mu_{Ri} \\ dI_i/dt &= \lambda_{i,w} - I_i/\mu_{Ii} \\ dD_i/dt &= \rho_i I_i/\mu_{Ii} - D_i/\mu_{Di} \\ R_i &= N_i - S_i - I_i - D_i \end{aligned} \qquad (3)$$

In summary, the infection transmission model is specified by equation (1) and equation (3). We refer to this as an SIDRS/W model. The parameters in the above equations, as well as values that are consistent with *Cryptosporidium,* are presented in Table 18.1. Parameters without base values are functions of other parameters, or are unknown or varied in the sensitivity analysis to follow. The term in brackets is the unit of measure for the parameter values in the table.

18.2.2 Policy regions for water treatment decisions

This section presents a sensitivity analysis for water treatment policy regions for centralized versus local treatment interventions. We consider *n=2* population subgroups, (1) immunocompetent and (2) immunocompromised individuals, and their exposure to *Cryptosporidium.* The centralized water treatment considered here is ozone pretreatment, which can remove 40-60% of *Cryptosporidium* oocysts from water. This has the effect of reducing τ_i by an appropriate percentage for the entire population. The local treatment considered here is a filter that essentially removes exposure from drinking water (as an extreme case) for the immunocompromised subgroup. This sets $\tau_2 = 0$ for the immunocompromised subgroup, but leaves τ_1 unchanged for the immunocompetent subgroup.

We define the 'better' treatment in this chapter as that which leads to the lowest endemic prevalence of cryptosporidiosis in the immunocompromised subgroup. This objective is motivated by the extreme effects of cryptospor-

Table 18.1 Summary of notation for SIDRS/W model, ranges for Cryptosporidium, and values used in a base analysis of the deterministic ODE model

Symbol	Meaning	Range	Base Value
n	Number of subgroups in human population	1, 2, …	2
N	Total # individuals in human population	> 0	$1.6\text{x}10^6$
N_i	Total # individuals in subgroup i	> 0	
γ	Rate of exogenous introduction of microbes [microbes/liter/day]	10^{-6}-10^2	10^{-6}
α	Rate microbes become inactivated [1/day]	.05	.05
θ_i	Rate an infectious individual sheds microbes [microbes /liter/day]	≥ 0	0
r_i	Probability an ingested microbe causes infection	.0021-.0076	.00428
ϕ_i	Water consumption [liters/day]	.017-2	1
τ_i	Fraction of microbes surviving water treatment	10^{-6}-1	10^{-3}
ρ_i	Probability that infection progresses to disease	.38-.81	.61
μ_{Ii}	Mean incubation period [days]	1-12	7
μ_{Di}	Mean duration of disease stage [days]	1-55	9
μ_{Ri}	Mean duration of recovered/immune stage [days]	60-120	90
	Fraction of oocysts viable after ozone pre-treatment	.2-.8	.4
c_i	Human contact rate for members of subgroup i [contacts/day]	≥ 0	
β_i	Probability a susceptible member of subgroup i becomes infected from a potentially infectious human contact	0.0-1.0	
$\lambda_{i,W}$	Force of infection to subgroup i, given microbe concentration W	See equation (2)	

idiosis in that subgroup during the 1993 Milwaukee outbreak. A similar analysis can be run for other outcome measures of merit, including quality-adjusted and disability-adjusted life years, and cost effectiveness ratios, but we do not do so here.

A risk assessment that ignores the dynamics of secondary transmission would conclude that the filter is more successful that ozone pretreatment for the immunocompromised subgroup. If secondary transmission is significant, however, secondary transmission from the immunocompetent subgroup can result in significant infection in the immunocompromised subgroup. In fact, if human-to-human secondary transmission is high enough, then removing all microbes from the water will still not prevent endemic transmission. In that case, water treatment makes almost no impact on the prevalence of infection.

Before presenting policy regions, we introduce notation to describe secondary transmission. Let $R_{0h,ij}$ be the mean number of secondary transmissions from human contact by an infected individual of subgroup *j* to individuals in subgroup *i,* assuming that all individuals in subgroup *i* are susceptible (c_j contacts per unit time, a fraction $c_i N_i / \Sigma c_k N_k$ of them with subgroup *i,* of which β_i are infective, for a mean duration of $\mu_{Ij} + \rho_j \mu_{Dj}$).

$$R_{0h,ij} = c_j \frac{c_i N_i}{\sum c_k N_k} \beta_i \left(\mu_{Ij} + \rho_j \mu_{Dj}\right). \tag{4}$$

Let $R_{0w,ij}$ be the analogous number of secondary transmissions through the water loop from an infective in subgroup *j* to individuals in subgroup *i,* assuming that all individuals in subgroup *i* are susceptible (see Appendix).

$$R_{0w,ij} = \frac{N_i r_i \tau_i \phi_i \theta_j}{\alpha} \left(\mu_{Ij} + \rho_j \mu_{Dj}\right). \tag{5}$$

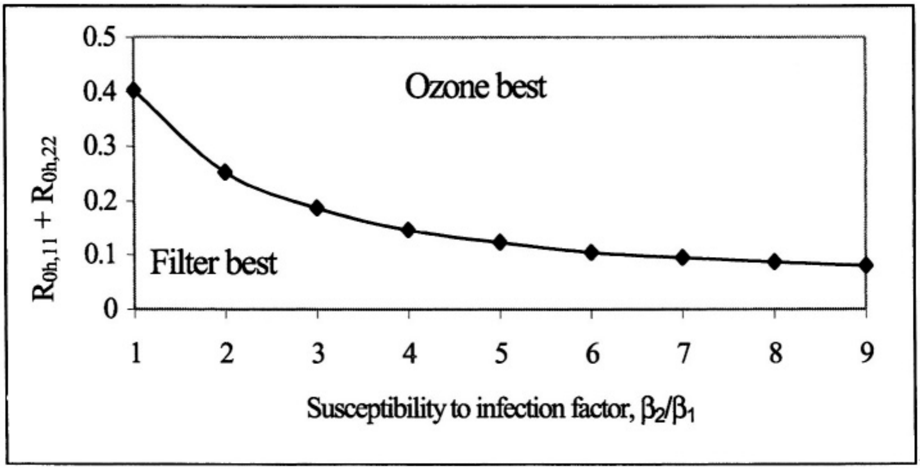
The Appendix proposes two different arguments to show that the basic reproduction number, R_0, is key to determining the infection dynamics. It can be related to the expected number of secondary infections needed to sustain endemic infection.

$$\begin{aligned} R_0 = R_{0h,11} + R_{0w,11} + R_{0h,22} + R_{0w,22} - R_{0h,11} R_{0w,22} - \\ R_{0h,22} R_{0w,11} + R_{0h,12} R_{0w,21} + R_{0h,21} R_{0w,12}. \end{aligned} \tag{6}$$

If $R_0 > 1$, then infection remains endemic even if no exogenous introduction of microbes occurs ($\gamma = 0$).

Somewhat surprisingly, it is still possible for a municipal improvement like ozone pretreatment to outperform filters on the taps of immunocompromised individuals, even if endemic infection is not sustainable through secondary transmission. The reason is that cryptosporidiosis prevalence in the immunocompetent subgroup can be significantly reduced with ozone pretreatment. This in turn reduces secondary exposure of cryptosporidiosis to the immunocompromised subgroup. Figure 18.2 illustrates that ozone pretreatment is more successful at reducing endemic cryptosporidiosis infection in the immunocompromised subgroup if the secondary transmission rate from human contact is high enough. This graph assumes that all secondary transmission occurs from human contact ($R_0 = R_{0h,11}+R_{0h,22}$, because $\theta_1 = \theta_2 = 0$), and that other parameters take on the base values for *Cryptosporidium* given in Table 18.1. If immunocompromised individuals are much more susceptible to cryptosporidiosis than immunocompetent individuals (larger β_2/β_1), then ozone pretreatment is attractive at even lower levels of secondary transmission. The values for $c_i\beta_i$ are chosen to give rise to the corresponding value of R_0 on the y-axis.

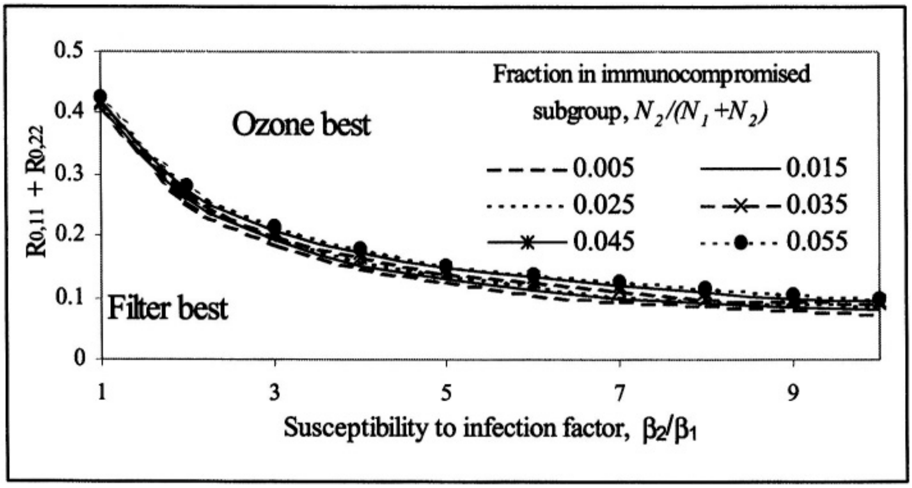
Figure 18.2 Ozone pre-treatment is better for larger values of secondary transmission or the relative susceptibility of immunocompromised individuals



These observations are qualitatively similar to results in our previous work [14]. The policy region boundary is somewhat lower here than in [14] for several reasons: the natural history of infection is more realistic here (including two infectious periods, the infected/asymptomatic and diseased/symptomatic states), a more effective ozone pretreatment process is assumed (60% of oocysts are removed rather than 50%), and a few other parameters are changed. The qualitative shape of the policy region, however, is the same. We now extend the results by assessing the sensitivity of the policy region to several parameters that may affect transmission dynamics.

Figure 18.3 shows that the policy region is relatively insensitive to the fraction $N_2 /(N_1 + N_2)$ of individuals in immunocompromised subgroup, at least when base case parameter values are used, and the fraction of immunocompromised individuals is relatively small (under 5% or so). If that fraction increases, the policy region boundary would rise, as direct exposure would become relatively more important than secondary transmission from the smaller immunocompetent subgroup. The policy region is similarly insensitive [25] to the rate γ of exogenous introduction of microbes, except if rates would lead to oocyst concentrations found during outbreaks with plant failures.

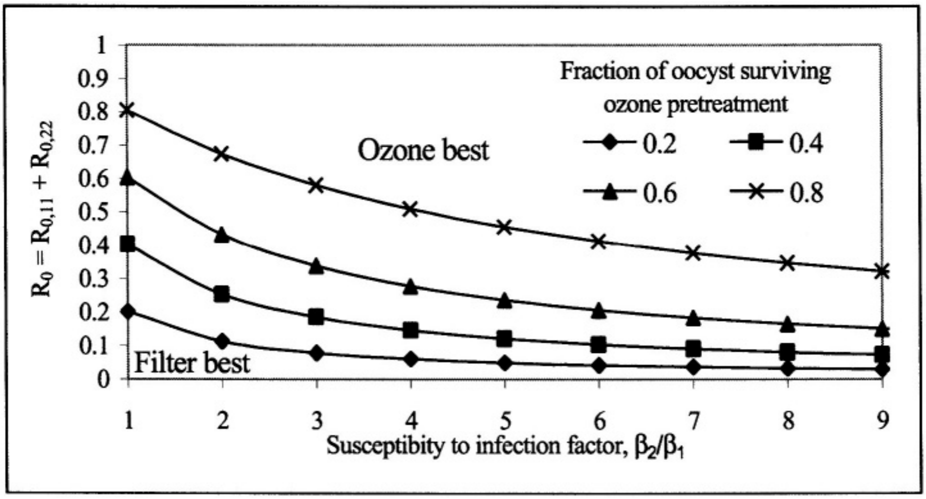
Figure 18.3 The policy region is relatively stable over a range of values for the fraction of population that is immunocompromised



The 1993 Milwaukee outbreak data has been used to estimate [13] the secondary transmission rate as R_0=0.15. The secondary transmission rate during endemic situations is unknown, but individuals may be more conscientious about secondary transmission during an outbreak than when infection is transmitted silently in the background. It seems reasonable to assume that immunocompromised individuals may be somewhat more susceptible to cryptosporidiosis infection due to human transmission ($\beta_2/ \beta_1 > 1$), but there is inconclusive data one way or the other [26].

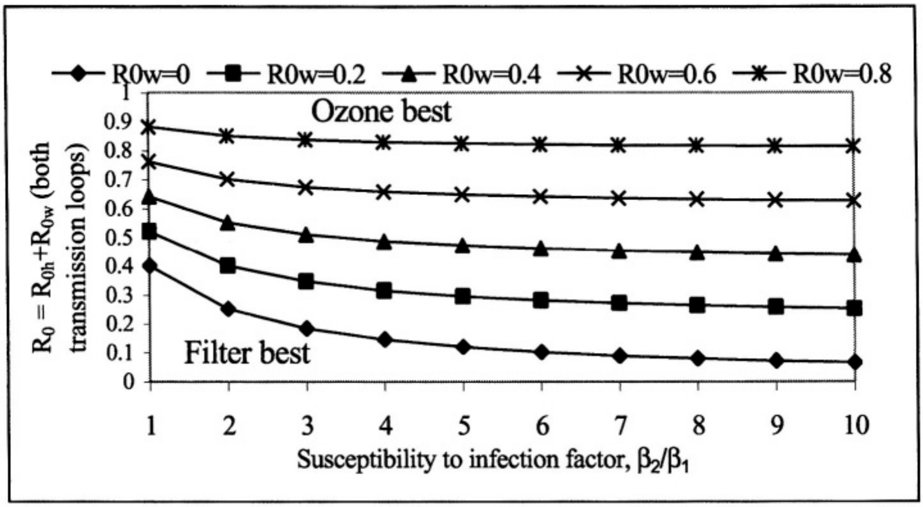
Figure 18.4 illustrates the sensitivity of the policy region to ozone pretreatment efficiency. Ozone pretreatment outperforms filters in this analysis even at relatively low values of secondary transmission, assuming that 80% of oocysts can be inactivated during the pretreatment. Although the values of secondary transmission parameters are not completely understood, this would put the treatment policy boundary near educated approximations for the parameter estimates. On the other hand, a risk assessment that assumes that there is no secondary transmission from interpersonal contact would indicate that filters are much more effective at reducing the endemic prevalence of cryptosporidiosis in the immunocompromised subgroup.

Figure 18.4 The policy region is highly dependent upon the effectiveness of ozone pre-treatment for removing oocysts



The graphs above assume that human contact is the sole exposure for secondary transmission, with no active water loop. This may be appropriate where there is no potential for recreational activities to contaminate source water. In some regions, however, recreational use can pose a distinct risk for water loop transmission [7]. Figure 18.5 shows that as the water loop increases in importance for transmission (increasing $R_{0w} = R_{0w,11} + R_{0w,22}$), the policy region boundary rises. Conceptually, this matches the notion that if *all* secondary transmission occurs through the water loop, with no human contact, then filters for the immunocompromised subgroup are more effective than ozone pretreatment in reducing cryptosporidiosis prevalence in that subgroup (filters are assumed here to be 100% effective, but ozone is only partially effective at removing oocysts). This means that filters are always more effective, relative to this objective, when there is no human-to-human transmission. Filters may still be an effective intervention if secondary transmission occurs primarily through the water loop.

Figure 18.5 Filters are much more effective if the water loop increases in importance relative to human-to-human secondary transmission



18.3 VARIATION IN INFECTION OUTCOMES

Infection and recovery times are stochastic, not deterministic; this is one source of variation in prevalence and microbial contamination data. How much variation in infection outcomes should one expect, even if all infection transmission parameters are known precisely? Another important related question is how to estimate unknown infection parameters, given field data. While the policy regions like those in Section 18.2 are useful for qualitative insights into the effects of treatment given transmission parameter assumptions, the precise values of parameters are still poorly understood for several microbes transmitted through the water system. A model of the random variation in infection prevalence and microbe concentration can be used as a likelihood function to help infer the unknown parameters. Ideally, such a model would be easy to simulate quickly.

18.3.1 Stochastic model background

Several authors have incorporated stochastic system dynamics to infer the parameters of infection models. Deterministic ODE infection models may have stochastic analogs that are derivable as large population limits [20-23, 27]. A continuous time stochastic analog of the deterministic SIS/W model with n closed subgroups, the model in Section 18.2 without the extra disease states, has a state $(S_1, \ldots, S_n, Y)$, where $Y=WN\Delta$ is the total oocyst count in the drinking water supply[1]. The state space is a lattice, $\{\prod_{i=1}^{n}\{0, 1, \ldots, N_i\}\}x\{0, 1, \ldots\}$. The state does not include I_i since $I_i = N_i - S_i$ by assumption here. State transition rates are determined by the associated rate in the ODE. For example, the transition rate from $(S_1,\ldots,S_i,\ldots, S_n,Y)$ to $(S_1,\ldots,S_i+1\ldots, S_n, Y)$ is I_i / μ_{Ii}, based on the recovery rate $1/\mu_{Ii}$ of each individual. Infection transitions to $(S_1, \ldots, S_i\text{-}1 \ldots, S_n, Y)$ occur with rate

$$r_i\phi_i\tau_i WS_i + \sum_{j=1}^{n} c_j I_j \frac{c_i N_i}{\sum_{k=1}^{n} c_k N_k} \frac{S_i}{N_i} \beta_i \,. \qquad (7)$$

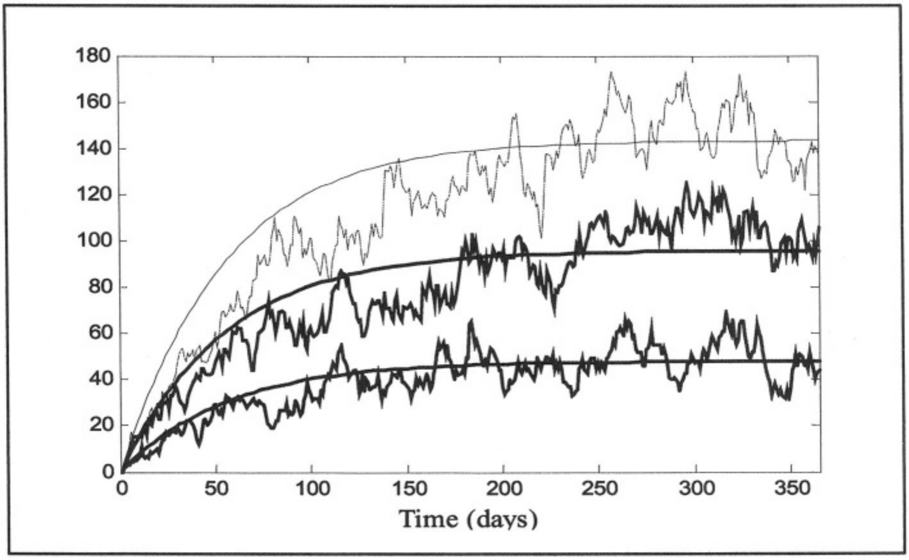
The analogy of these rates with equations (2) and (3) should be clear. Rate terms in the ODE dynamics correspond to infinitesimal state transition rates in the Markov chain model, and transition rates for microbe immigration and inactivation occur similarly. Figure 18.6 shows a sample path for the number infected in a *3*-subgroup model as it varies about the trajectory of the analogous ODE model. An alternate approach is to use a closely related discrete-time Reed-Frost epidemic model [18], or to also incorporate social network information into the state with a stochastic graph [19].

Several researchers (e.g., see [19] and references therein) have developed likelihood models for Bayesian inference that incorporate infection dynamics into the likelihood function for parameters as a function of data that might be obtained from tracing an outbreak, or closely monitoring an intervention trial. Interesting properties of quasistationary distributions [28], the long run distribution assuming that infection remains endemic, of infection models have been studied as well.

A recent proposal to infer infection parameters with endemic data provides a statistical tool that provides an alternative to waiting for, identifying, and measuring an outbreak [16]. The work uses stationary distributions of a combined stochastic-deterministic SIS/W infection model in a homogenous population to model endemic data. Infection and recovery events were assumed to be stochastic, but water contamination was assumed to be deterministic, given the number infected. Because a closed form for the stationary distribution is not known and there are situations when normal approximations used by standard ODE least square estimators are not fully justified, the authors develop two likelihood approximations.

[1] The total number of microbes, not microbes per volume, in a volume $N\Delta$ of water that scales with N is needed to obtain diffusion approximation results. See [20-23], Appendix A.4.

Figure 18.6 A sample path for the number infected in three subgroups for a stochastic model varies about the trajectory of the analogous ODE



The first uses the stationary distribution of a closely related lattice Markov chain whose state is the number infected. That likelihood approximation has good bias and root mean square error (RMSE) properties, but may be computationally intensive when extended to populations with multiple subgroups, or if the natural history of infection is more complex. The second likelihood approximation uses a normal distribution approximation that takes advantage of relationships between low order moments that are determined by the Kolmogorov forward equations, but is somewhat more biased or may give confidence regions that are too small, particularly near R_0 = 1 or when populations are small, where the normal approximation may be suspect. Further, the continuous dynamics for the water, combined with the moment relationships, may or may not give a full specification of the system with more complicated natural histories of infection, or with multiple subpopulations (e.g., higher dimensions).

Here we take an alternate approach to approximating the stationary distribution of the number of infections: a diffusion approximation [20-23]. While statistical bias issues may remain to be resolved if the populations are small or if R_0 is near 1, the approach appears to be more generalizable to higher dimensions. While the mixed stochastic/deterministic model in [16] cannot directly use diffusion approximation results, a slight change to use the stochastic model on the lattice introduced at the beginning of Section 18.3.1 makes those results applicable. In particular [20-23] illustrate that density-dependent processes, which include many epidemic models like the lattice-state model above, can be approximated (in law) by an Ornstein-Uhlenbeck (OU) process near an endemic equilibrium point, as N grows. The stationary mean is approximated by the ODE's asymptotically stable endemic infection level (which is positive if there is exogenous contamination, $\gamma > 0$, and other parameters are not 0), and the stationary covariance matrix Σ can be approximated by appropriately rescaling the solution to a Lyapunov equation, as overviewed in Appendix A.4.

18.3.2 Preliminary results for diffusion approximation

This chapter presents only preliminary results for the OU approximation. We simulated the continuous time SIS/W stochastic process with proportional mixing and compared sample statistics for the stationary mean and variance with the endemic ODE mean and OU approximation to the variance.

Simulated population sizes were 60, 600 and 6000 individuals in $n=3$ subgroups, with 1/6 of the individuals in subgroup 1, 1/3 in subgroup 2, and 1/2 in subgroup 3. Parameters were chosen so that a fair amount of secondary transmission would be observed. Parameters were chosen to be the same for each subpopulation, with $c_i=c$, $\beta_i=\beta$, $\mu_{Ii}=\mu_I=7$ days, etc, so that $R_{0h} = c\beta\mu_I = 0.875$, $R_{0w} = 0.05$. Table 18.2 provides some summary sample statistics for the stationary mean $\bar{i}_j$ and standard deviation σ_j of the number infected in subgroup j. The statistics were based on 150 years of simulated infection and water contamination. The means are time averages, and the standard deviations are based upon sampling the number infected once per month. The OU approximation for the mean equals the ODE endemic equilibrium, and the standard deviations are computed as described in the previous section and Appendix A.4. As observed elsewhere [29, 30], the mean number infected estimated by the simulations is lower than predicted by the deterministic model for smaller populations. Correlation is strong between subgroups and water contamination levels in simulations with significant secondary transmission, matching simulations with a single subgroup in [16].

The OU approximation for the mean and variance of the number infected provides yet another likelihood approximation for inferring infection parameters from endemic data, to complement the two approximations in [16] (the two ideas there were to compute the stationary distribution, and to use the Kolmogorov forward equations to establish relationships between moments). How strongly the bias in the estimates of the means and variances

Table 18.2 Comparison of some estimates from long simulation runs versus the Ornstein-Uhlenbeck (OU) approximation for the stationary mean and covariance

Population Size, N	Approx-imation	Statistic			
		$\bar{i}_2$	$\bar{i}_3$	σ_2	σ_3
60	Simulation	0.72	1.10	0.927	1.18
	OU	0.95	1.44	1.10	1.44
600	Simulation	9.28	13.9	3.72	5.32
	OU	9.56	14.4	3.81	5.25
6000	Simulation	95.7	143	13.5	18.82
	OU	95.5	143	14.3	20.97

might influence the bias and RMSE of parameter estimators based upon an OU likelihood approximation is an area for further study.

18.4 CONCLUSIONS AND FUTURE RESEARCH

18.4.1 Water treatment policy

Infection transmission dynamics can strongly influence the public health benefit of water treatment interventions. Ignoring secondary transmission in a risk assessment, or examining only first-order effects, can suggest misleading conclusions. System dynamics models can help quantify the complex infection dynamics that some microbes transmitted through the drinking water system may have. Policy decisions regarding the recreational use of public waterways that are source water directly influence the potential for secondary transmission, too.

While infection transmission parameters may be important determinants of the health benefit of interventions, their values are not well understood for a number of microbes. The use of stationary distributions as likelihood functions for unknown parameters allows endemic data to be used in the inference process. This complements tools by others to infer parameters in an intervention trial or with outbreak data [17-19].

Here we considered only one microbial agent. In reality, there are many strains of many microbes. A comprehensive risk management program must consider multiple microbes and multiple intervention options. Further, some coordination may be required between different governmental agencies. The Centers for Disease Control and Prevention are historically responsible for outbreak and infection data, whereas the EPA is historically responsible for water quality data.

18.4.2 Infection modeling

One advantage of the OU approximation, at least for large populations with nontrivial endemic levels, is that the mean and covariance matrix are readily computed. Furthermore, transient probability distributions can be estimated with this OU approximation under certain conditions [23]. In principle, this would allow for data from outbreaks, intervention trials and/or endemic data to be used to infer transmission parameters.

The OU approximation has a statistical bias when the population size is small, or there are small numbers of individuals per subgroup, such as occurs when family units or small work sites form the subgroups [30]. The stationary and quasistationary mean prevalence of the lattice-based Markov chain infection model may be lower than the scaled endemic equilibrium infection level. The difference goes to zero in the large population limit, but may be nontrivial for small populations. A rigorous exploration of this bias is an area for further research.

Such bias holds implications not only for parameter inference, but also for speeding up simulations of infection processes. The OU process might ignore every infection and recovery event in a large process, but may require small time steps to insure that bias is avoided. An interesting simulation question is to evaluate effective ways to simulate the approximating OU process in a way that faithfully represents important low order statistical properties of the original Markov process. Approximations that work well when almost everybody or almost nobody in a subgroup is infected are an open area of research, and have implications for simulating small populations and subgroups, such as family units or daycare centers that are participating in water treatment intervention or vaccine trials.

18.4.3 Other modeling applications

The food supply chain is complex and presents another potential route for the transmission of microbes. Some infection models have examined the dynamics of growth of microbes as food passes from the farm to the fork [31] in one context. Others have examined the dynamics of infection in herds [32, 22] and the ensuing impact on the livestock industry. One area for further development is the integration of infection dynamics models in animals, microbes in the food supply chain, and primary and secondary exposure in human populations.

Acknowledgments

We appreciate discussions with Brenda Boutin, PhD, and the financial support of the US EPA National Center for Environmental Assessment (CR-827427-01-0).

References

[1] Hoxie, N.J., J.P Davis, J.M Vergeront, R.Nashold, and K. Blair (1997). Cryptosporidiosis-associated mortality following a massive waterborne outbreak in Milwaukee, Wisconsin. *American Journal of Public Health,* 87, 2032-2035.

[2] MacKenzie, W.R., et al. (1994), A massive outbreak in Milwaukee of Cryptosporidium infection transmitted through the public water supply. *New England Journal of Medicine,* 331, 161-167.

[3] Iley, K. (2002). Aid groups urge action on water-borne diseases. Reuters News Service at www.planetark.org, March 25.

[4] Cowdy, H. (2002). Millions at risk from contaminated water. Reuters News Service at www.planetark.org, March 25.

[5] Centers for Disease Control and Prevention (2001). Norwalk-like viruses: Public health consequences and outbreak management. *Morbidity and Mortality Weekly Report,* 50(RR-9), 1-17.

[6] Cooke, R. and B. Kraan (2000). Processing expert judgements in accident consequence modelling. *Radiation Protection Dosimetry,* 90, 311-315.

[7] Stewart, M.H., M.V. Yates, M.A. Anderson, C.P. Gerba, J.B. Rose, R.DeLeon, and R.L. Wolfe (2002). Predicted public health consequences of body-contact recreation on a potable water reservoir. *Journal of the American Water Works Association,* 94, 84-97.

[8] Stout, J.E., Y-S E. Lin, A.M. Goetz, and R.R. Muder (1998). Controlling legionella in hospital water systems: Experience with the superheat-and-flush method and copper-silver ionization. *Infection Control and Hospital Epidemiology,* 19, 663-674.

[9] Current, W. (1994). Cryptosporidium parvum: Household transmission. *Annals of Internal Medicine,* 120, 518-519.

[10] Medema, G. and J. Schijven (2001). Modeling the sewage discharge and dispersion of Cryptosporidium and giardia in surface water. *Water Research,* 35, 4370-4316.

[11] Regli, S., J.B. Rose, C.N. Haas, and C.P. Gerba (1991). Modeling the risk from giardia and viruses in drinking water. *Journal of the American Water Works Association,* 83, 76-84.

[12] International Life Sciences Institute (2000). *Revised Framework for Microbial Risk Assessment.* ILSI Risk Science Institute workshop report, International Life Sciences Institute, Washington, DC.

[13] Eisenberg, J.E., E.Y.W. Seto, J.M. Colford Jr, A. Olivieri, and R.C. Spear (1998). An analysis of the Milwaukee Cryptosporidiosis outbreak based on a dynamic model of the infection process. *Epidemiology,* 9, 255-263.

[14] Chick, S.E., J.S. Koopman, S. Soorapanth, and M. E. Brown (2001). Infection transmission system models for microbial risk assessment. *Science of the Total Environment,* 274, 197-207.

[15] Frisby, H.R., D.G. Addiss, W.J. Reiser, et al. (1997). Clinical and epidemiologic features of a massive waterborne outbreak of WJ Cryptosporidiosis in persons with HIV infection. *Journal of Acquired Immune Deficiency Syndromes and Human Retrovirology,* 16, 367-373.

[16] Chick, S.E., J.S. Koopman, S. Soorapanth, and B.K. Boutin (2003). Inferring infection transmission parameters that influence water treatment decisions. *Management Science,* 49, 920-935.

[17] Brookhart, M.A., A.E. Hubbard, ME. van der Laan, J.M. Colford, and J.N.S. Eisenberg (2002). Statistical estimation of parameters in a disease transmission model. *Statistics in Medicine,* 21, 3627-3638.

[18] O'Neill, P.D. (2003). Perfect simulation for Reed-Frost epidemic models. *Statistics and Computing,* 13, 37-44.

[19] Britton, T. and P.D. O'Neill (2002). Statistical inference for stochastic epidemics in populations with random social structure. *Scandinavian Journal of Statistics,* 29, 375-390.

[20] Kurtz, T.G. (1970). Solutions of ordinary differential equations as limits of pure jump Markov processes. *Journal of Applied Probability,* 7, 49-58.

[21] Kurtz, T.G. (1971). Limit theorems for sequences of jump Markov processes approximating ordinary differential processes. *Journal of Applied Probability,* 8, 344-356.

[22] Clancy, D. and N.P. French (2001). A stochastic model for disease transmission in a managed herd, motivated by Neospora caninum amongst dairy cattle. *Mathematical Biosciences,* 170, 113-132.

[23] Pollett, P.K. (1990). On a model for interference between searching insect parasites. *Journal of the Australian Mathematical Society, Series B,* 31, 133-150.

[24] Nold, A. (1980). Heterogeneity in disease-transmission modeling. *Mathematical Biosciences*, 52, 227-240.

[25] Soorapanth, S. (2002). *Microbial Risk Models Designed to Inform Water Treatment Policy Decisions.* PhD Thesis, University of Michigan, Ann Arbor, MI.

[26] Osewe, P., D.G. Addiss, K.A. Blair, A. Hightower, M.L. Kamb, and J.P. Davis (1996). Cryptosporidiosis in Wisconsin: A case-control study of post-outbreak transmission. *Epidemiology and Infection,* 117, 297-304.

[27] Altmann, M. (1998). The deterministic limit of infectious disease models with dynamic partners. *Mathematical Biosciences,* 150, 153-175.

[28] Nåsell, I. (1996). The quasi-stationary distribution of the closed endemic SIS Model. *Advances in Applied Probability,* 28, 895-932.

[29] Chick, S.E. (2002). Approximations of stochastic epidemic models for parameter inference. Working Paper, INSEAD, Fontainbleau, France.

[30] Koopman, J.S., S.E. Chick, C.S. Riolo, C.P. Simon, and J.A. Jacquez (2002). Stochastic effects on endemic infection levels of disseminating versus local contacts. *Mathematical Biosciences,* 180, 49-71.

[31] McNab, B.W. (1998). A general framework illustrating an approach to quantitative microbial food safety risk assessment. *Journal of Food Protection,* 61, 1216-1228.

[32] Ferguson, N.M., C.A. Donnelly, and R.M. Anderson (2001). The foot-and-mouth epidemic in Great Britain: Pattern of spread and impact of interventions. *Science,* 292, 1155-1160.

[33] Simon, C.P. and J.A. Jacquez (1992). Reproduction numbers and the stability of equilibria of SI models for heterogeneous populations. *SIAM Journal of Applied Mathematics,* 52, 541-576.

Appendix

A.1 Algebraic stability conditions for the SIRS/W model

Consider first the SIRS/W infection transmission model in a homogenously mixing population (a special case of the general model, with $n=1$ subgroup, $\rho=0$, so $D=0$ and we drop subscripts in this section). The expected number R_{0h} of secondary transmissions, due to human contact, caused by one infective in an otherwise susceptible population, is the contact rate c, times the infection probability per contact β, times the duration of infection μ_I.

$$R_{0h} = c\beta\mu_I \,. \tag{A.1}$$

The analogous number of secondary transmissions through the water loop is qualitatively derived by noting that an infected individual raises the concentration of microbes by θ microbes per day for μ_I days, the microbes remain viable for $1/\alpha$ days, and each of N susceptibles consumes a fraction $\tau\phi$ of available microbes, each of which causes infection with probability r.

$$R_{0w} = \frac{Nr\tau\phi\theta}{\alpha}\mu_I \,. \tag{A.2}$$

Then $R_0=R_{0h}+R_{0w}$ is the total number of secondary transmissions, on average.

Theorem 1: If there is no exogenous source of microbes $(\gamma=0)$, and there is homogeneous mixing ($n=1$), then

- The disease free equilibrium ($S^* = N$, $I^* = R^* = W^* = 0$) is locally asymptotically stable if $R_0 < 1$, and unstable if $R_0 > 1$.
- The endemic equilibrium ($S^*=N/R_0$, $I^* = N(1\text{-}1/R_0)\mu_I,/(\mu_I + \mu_R)$, $W^*= \theta I^*/\alpha$) is locally asymptotically stable if $R_0 > 1$, and is not realizable if $R_0 < 1$.

Proof: The equilibrium values are determined by setting derivatives to 0. The stability result is proven by linearization in [25].

A.2 Algebraic stability conditions for the ODE in Section 18.2, (n=2)

By analogy with equation (A.1), let $R_{0h,ij}$ and $R_{0h,ij}$ be as in equations (4) and (5).

Theorem 2: Suppose there is no exogenous source of microbes $(\gamma = 0)$, and there are $n=2$ subgroups with proportional mixing, as in Section 18.1. Consider the following two conditions.

(i) $$R_{0h,11} + R_{0w,11} + R_{0h,22} + R_{0w,22} - R_{0h,11}R_{0w,22} - R_{0h,22}R_{0w,11} + R_{0h,12}R_{0w,21} + R_{0h,21}R_{0w,12} < 1$$

(ii) $$R_{0h,11} + R_{0w,11} + R_{0h,22} + R_{0w,22} < 1$$

Then:

- Conditions (i) and (ii) are sufficient for the disease free equilibrium to be asymptotically stable $(S_l{}^* = N_l,\ I_l{}^* = R_l{}^* = W^* = 0)$.
- If the inequality in condition (i) is reversed, then the zero equilibrium is not stable, resulting in positive endemic infection.

Proof: The stability result is proven by linearization in [25]. The two conditions are equivalent when $c_1\theta_2 = c_2\theta_1$. If $c_1\theta_2 \neq c_2\theta_1$, then the linearization leads to a quintic equation after some factorization, which is not solvable in closed form. The two conditions together are sufficient to insure that the dominant eigenvalue falls in the left hand complex plane.

We have not yet developed characterizations for $n>2$ subgroups when the water loop is active. [33] use Lyapunov functions to characterize stability for $n \geq 1$ subgroups with proportional and other mixing patterns for human-to-human transmission, but do not account for the water loop.

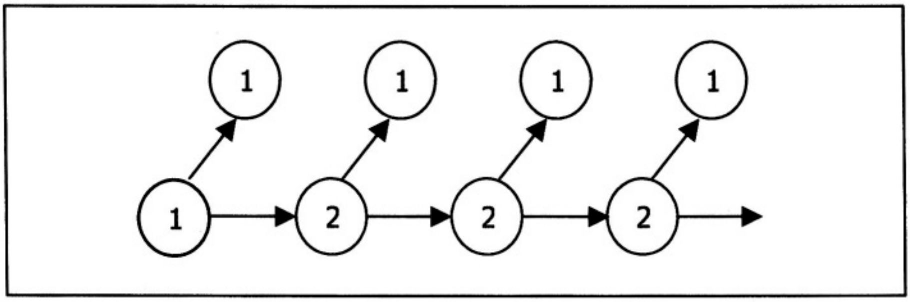
A.3 Alternate stability conditions for the ODE in Section 18.2

Sections A. 1 and A.2 above provide *population thresholds* to characterize stability based on an algebraic analysis. An alternate heuristic to assess whether endemic infection is sustainable even if no exogenous introduction of microbes occurs $(\gamma = 0)$ is to assess an *individual level* endemic threshold using probabilistic arguments. This section overviews such an argument for the $n = 2$ subgroup model. Let $R_{0h,ij}$ and $R_{0w,ij}$ be as in Section A.2, and denote the total mean number of secondary transmissions to a completely susceptible subgroup *i* from an index case in subgroup *j* by

$$R_{0,ij} = R_{0h,ij} + R_{0w,ij}\,. \qquad \text{(A.3)}$$

The individual level threshold is established by assessing whether the mean number of new infections in subgroup *i* caused by an initial index case in *i* is at least 1, when the whole chain of infection is considered. For example, an individual in subgroup 1 can infect someone in subgroup 2, who then infects another person, who then eventually infects someone in subgroup 1. The expected number of infections (directly or indirectly caused) in the chain (see Figure 18.7) should be at least 1 for at least one subgroup.

Figure 18.7 The chain of infection from an index case in subgroup 1 can result in infections in subgroup 1 directly, or indirectly through subgroup 2



If $R_{0,11} > 1$ or $R_{0,22} > 1$, then a given subgroup can sustain infection within itself, and therefore infection remains endemic. Consider the case where $R_{0,11}$ and $R_{0,22}$ are both at least 0 but neither exceeds 1. If a single individual in subgroup 1 is infected, and the population is otherwise susceptible, then the expected number R of additional cases through the whole chain of transmission that eventually reach subgroup 1 is

$$\begin{aligned} R &= R_{0,11} + R_{0,21}\left(R_{0,12} + R_{0,22}\left(R_{0,12} + R_{0,22}(\cdots)\right)\right) \\ &= R_{0,11} + R_{0,21}R_{0,12} + R_{0,21}R_{0,12}R_{0,22} + R_{0,21}R_{0,12}R_{0,22}^{2} + \cdots \qquad \text{(A.4)} \\ &= R_{0,11} + \frac{R_{0,21}R_{0,12}}{1 - R_{0,22}}. \end{aligned}$$

The last equation holds because $R_{0,22} \in [0,1)$. An individual level threshold says that endemics cannot be sustained without exogenous sources of infection if $R<1$, or $R_{0,11} + R_{0,22} - R_{0,11}R_{0,22} + R_{0,12}R_{0,21} < 1$. Substituting the definition of $R_{0,ij}$ in equation (A.3) gives an individual level threshold that is equivalent to the population threshold in condition (i) of Theorem 2 above.

A.4 Ornstein-Uhlenbeck approximation to SIS/W process

The OU approximation to the stochastic SIS/W model with proportional mixing can be derived by examining the ODE analog of that model along with the transmission rates of the stochastic model $(S_1,\ldots, S_n, W)$ summarized in Section 18.3.1. The idea (e.g., [22, 23]) is to first find a representation so that the state scales up with N, the total population size, then to look at a rescaled version of that process. The S_i already scale directly with N. To get the microbe contamination to scale with N, we model the total oocyst count Y in the drinking supply, rather than microbe concentration, and suppose that the drinking supply scales with the population size (e.g., contains a total of $N\Delta$ liters, and water drunk by individuals is replaced with fresh water so that the total volume remains constant). This means that $Y = N\Delta W$, γ is the oocyst contamination rate per unit time per Δ liters of water, and the rescaled process of interest is $x=(S_1,\ldots, S_n, Y)/N$.

Let $dx/dt = f(x_0)$ be a vector valued function that describes the dynamics of the scaled ODE model, and let x_0 be an asymptotically stable equilibrium in the interior of the scaled state space, with $f(x_0) = 0$. Let $A = \nabla f(x)\big|_{x=x_0}$ be the matrix containing the gradient of the dynamics $f(x)$ of the scaled process $x=(S_1,\ldots ,S_n, Y)/N$, evaluated at x_0, and let $\zeta_i = N_i/N$. Let the matrix G be the local covariance of a scaled version of the state over a short time δt, given that the state is currently x_0. For the SIS/W model with proportional mixing, G is a diagonal matrix, and is determined by evaluating the following at x_0.

$$G_{ii} = \frac{(\varsigma_i - x_i)}{\mu_{Ii}} + r_i\phi_i\tau_i w x_i + x_i \frac{c_i N_i \beta_i}{N_i \sum_{k=1}^{n} c_k N_k} \sum_{j=1}^{n} c_j(\varsigma_j - x_j), \mathit{for}\ i = 1,\ldots,n$$

$$G_{n+1,n+1} = \gamma + \sum_{j=1}^{n} \theta_j (N_j - S_j) + \alpha w \qquad \text{(A.5)}$$

The matrix G will have nonzero off-diagonal elements for the SIDRS/W model, since an increase in S_i means a decrease in I_i.

The stationary distribution for the OU process can be approximated with a normal distribution with mean Nx_0 and covariance matrix $N\Sigma$, where Σ solves equation (A.6) (e.g., see [22, 23] for similar models without water transmission).

$$A\Sigma + \Sigma A^T + G = 0 \qquad \text{(A.6)}$$

19 SCREENING FOR DIABETIC RETINOPATHY

Ruth Davies[1] and Sally C. Brailsford[2]

[1]Warwick Business School
University of Warwick
Coventry CV4 7AL, United Kingdom

[2]School of Management
University of Southampton
Southampton SO17 1BJ, United Kingdom

SUMMARY

A discrete event simulation describes the screening and natural history of eye disease in patients with diabetes, using the POST simulation software. Discrete event simulation, unlike other modeling techniques, can show the interaction between screening and the two main diabetic eye disease processes. Results show that there is a tradeoff between screening frequency, screening sensitivity and patient compliance. The extent to which screening is cost effective is not clear cut. We have little data on the cost of blindness and different views on the appropriate quality-of-life values to assume. The model can be extended to evaluate prevention and treatment for all the complications of diabetes.

KEY WORDS

Diabetes, Simulation, Blindness, Cost-effectiveness

19.1 INTRODUCTION

This chapter provides an example of the way in which disease modeling can be used to plan and influence health policy. Diabetes mellitus (DM) is a condition in which the body produces insufficient insulin to break down and absorb the glucose consumed in food. Type 1 DM, which usually appears in childhood, is generally treated with insulin injections. Type 2 DM, which is more common in middle aged and older people, is treated by diet, drugs, and in some cases insulin. Despite treatment, diabetes causes damage to tissues, producing complications. These include: retinopathy, nephropathy and neuropathy which may lead to blindness, end-stage kidney failure and limb amputations, respectively. These conditions are all serious, unpleasant and may be expensive to treat. The growing prevalence of Type 2 diabetes [1] will result in an increasing demand for treatment for these conditions.

There is little prospect, at present, of preventing diabetes altogether. New drugs and techniques to improve the control of existing drugs may help to prevent complications in the long term. Most interest at present is centered on screening programs, either to detect diabetes in apparently healthy individuals, or to detect early signs of complications that may be treated in order to prevent more serious future problems developing. The cost must be balanced against the resulting benefits. In this chapter we will review several screening models and then will describe our own model of screening to detect early indications of eye disease caused by diabetes. We will show how discrete event simulation can be applied to this problem and will provide some sample results.

19.2 DESCRIPTION OF APPLICATION

19.2.1 Background

Two main types of eye conditions may affect patients with diabetes: proliferative diabetic retinopathy and macular edema. The former is more prevalent in Type 1 and the latter in Type 2 DM patients.

The early stages of retinopathy, termed background retinopathy, are asymptomatic but changes can be detected by screening. Background retinopathy may lead to proliferative diabetic retinopathy which may, in turn, progress to vision loss through the formation of new blood vessels, bleeding and scar tissue formation. Treatment is given to patients with proliferative retinopathy, and sometimes pre-proliferative retinopathy, by scatter laser treatment. Although this treatment worsens the vision of some patients, for most patients it will slow, or completely stop, the progression to blindness.

In macular edema, blood leaks from the vessels of the eye to the central part of the retina. This diabetes complication is more common in the elderly and may lead to central vision loss. The Early Treatment Diabetic Retinopathy Study [2] identified a clinically significant stage of the condition in which treatment may be beneficial. Treatment, by focal laser treatment, is targeted at the affected part of the eye. For both macular edema and proliferative retinopathy, the eyes may progress to blindness at different rates, and several treatments may be needed for each eye.

Although the cumulative risk of retinopathy is substantial in Type 1 DM [3], reaching up to 60% for sight threatening disease over 20 years, the overall burden of preventable blindness due to Type 2 DM is greater [4] because its prevalence is higher. The eye diseases caused by diabetes fulfill the World Health Organization (WHO) criteria for screening [5]: in particular, they evolve through key recognizable stages in their progression to blindness, they represent an important health problem, there are valid and acceptable screening tests, and blindness can be prevented or visual decline slowed with laser photocoagulation.

Currently, significant variation exists in screening provision [6]; primary screening is undertaken using different staff and settings (optometrists, general practitioners, mobile cameras, hospital based diabetic physicians or ophthalmologists) and using various methods (e.g. mydriatic or non-mydriatic ophthalmoscopy or photography), and there is uncertainty about the appropriate screening intervals [7]. Our aim has been to provide a flexible model with which it would be possible to test a wide range of screening scenarios.

19.2.2 Other models

Models have been used to evaluate whether screening should be performed, how and of whom. The Wisconsin Epidemiologic Study of Diabetic Retinopathy (WESDR)[1] [3, 4, 8, 9] is the main study from which parameters for retinopathy transitions are derived. It is a widely published and highly respected cohort study that has been followed up since the early 1980s. Other studies providing data include: the Diabetic Retinopathy Study (DRS)[1] [10], the Early Treatment Diabetic Retinopathy Study (ETDRS)[1] [2], and the United Kingdom Prospective Diabetes Study (UKPDS)[1] [11]. Models differ in their use and interpretation of these data and in the assumptions they make. Some assumptions are enforced by the modeling approach adopted, while other assumptions are clinical or epidemiological. Examples of the

[1] These studies all produced a large number of other papers which we do not have room to reference here.

former include whether patients are considered as a group or as individuals, whether patients are assumed to belong to identical cohorts or whether a realistic population is modeled, whether multiple health states are possible, and the form of the mathematical expressions used to calculate transitions between states. Examples of clinical or epidemiological assumptions are the selection and definition of the health states, and the selection of the factors that influence transitions between these states. Models also differ in their outputs and endpoints. Cost-effectiveness models vary considerably in their assumptions about which costs should be included, how to measure benefits, and whether discounting should be used (and at what rates).

In the literature, the most widely used modeling approach for diabetic retinopathy, and other complications of diabetes, is the compartmental model (typically implemented in a spreadsheet). Compartmental models are based on homogeneous groups or cohorts of patients and are updated in equal time steps. Progression from state to state is based on the Markovian assumption that the probability of progression to a new state at time T depends only on the state occupied at time T – 1. If there are N patients in state A at time T – 1, and the probability of transition from state A to state B in one time step is *p,* then the number of patients who move from A to B at time T is simply *p*N. Such computations are easily carried out for successive periods in a spreadsheet. The resulting values are simply the mean numbers in each state. The variability in the numbers can be found using a stochastic simulation in which the numbers of individuals moving from state to state are sampled from random numbers in each time period.

Dasbach et al. [12] used a spreadsheet simulation based on four-year WESDR data for three population groups: Type 1 patients of five or more years duration, Type 2 patients taking insulin and Type 2 patients not taking insulin. The annual transition probabilities between four health states (low risk, high risk, blind and dead) depended solely on the start state and were derived mathematically from the WESDR four-year data. The model does not distinguish between macular edema and proliferative retinopathy. It simply calculates the number of people in each state at the end of each year; the states are homogeneous, there is no sampling and newly diagnosed cases are not added to the cohort.

Bachmann and Nelson [13] used an Excel spreadsheet model combined with Monte Carlo simulation, using the standard random number generator in Excel, to calculate the numbers of cases of treatable retinopathy detected and the cases of blindness prevented by a screening program. The model did not calculate the total years of sight saved, and assumed a constant population of diabetic patients, with deaths equal to new cases. This model used WESDR and relevant United Kingdom (UK) data. The authors assumed that no previous screening had been carried out, so that the initial yield would be considerably higher than in subsequent screening rounds. They found that "substantial disability" could be prevented by screening and early treatment of diabetic retinopathy.

Tomar et al. [14] developed a spreadsheet cohort Markov model for all major complications of Type 1 DM. They used six health states: healthy (no DM), DM without complications, DM with retinopathy, DM with neuropathy (with or without retinopathy), DM with nephropathy and any combination of retinopathy and neuropathy, and Dead. They used WESDR data to estimate the annual transition probabilities between the states. Cost data (or more precisely charge data) from a local health insurance company (University Health Care Inc, Wisconsin), were used to attach costs to the four intermediate states. Costs were discounted over the lifetime of the cohort.

In a more recent Markov cohort model for Type 2 DM by Vijan et al. [15], the progression rate up to pre-proliferative diabetic retinopathy depended on glycemic control, using data from UKPDS; after that, progression was assumed to be independent of glycemic control, based on data from the DRS and ETDRS. As in Bachmann's model [13], they assumed that no patients were screened prior to entry in the model. Although this was a deterministic spreadsheet model, Vijan et al. used Monte Carlo simulation to perform sensitivity analysis on the cost-effectiveness estimates. The main model output was cost per quality-adjusted life year (QALY) gained. They found that annual screening of all Type 2 patients was not cost effective and that screening every two years was adequate, except for patients with very poor glycemic control.

A recent model of this type by James et al. [16] compared a systematic screening program with existing practice. The systematic program involved a mobile screening unit combined with a dedicated hospital clinic; the "opportunistic" screening service was provided by optometrists, general practitioners and diabetologists, combined with referral to general hospital eye clinics. The output was given in terms of cost per case of eye disease detected.

The spreadsheet transition model approach is appealing because it is transparent – all the assumptions are made explicit in the spreadsheet – and such models are relatively quick to develop and run. However, transitions are assumed to take place only at distinct points in time (in the above examples, at the end of each year) and it is not possible to take into account multiple health states, interactions and co-morbidities. Additionally, all individuals must belong to a homogeneous group. In order to account for different age and risk groups, a compartmental model must include separate sets of states and transitions. The number of states can become very large and the model unwieldy. The Markovian assumption may also be limiting.

In order to overcome the problem of allocating people to states, some models incorporate individual variability. Javitt, with various colleagues, has been working in the area for many years [17-20], using a system called "PROPHET" (PROspective Population Health Event Tabulation). This is a cohort model, coded in Borland Pascal, which the authors say "combines features of decision trees, Markov processes and Monte Carlo simulation techniques". The model describes individuals whose progress is updated in two-monthly time steps; disease progression and mortality rates depend on age and disease severity. It incorporates both macular edema and proliferative retinopathy and uses data from WESDR, DRS, ETDRS and the Rochester study [21]. In the 1994 paper, the authors found that two-yearly screening until detection of background diabetic retinopathy (BDR) in both Types 1 and 2 DM appeared to be adequate, although they recommended yearly or even six-monthly screening thereafter [19].

The model of Type 2 diabetes that Javitt developed with Eastman et al. [22] is an Excel spreadsheet cohort model using the simulation add-in @Risk. It incorporates multiple complications: retinopathy, neuropathy, nephropathy, and coronary vascular disease. The model is updated annually. Every year, for each patient, the probability of death in the next year is derived, given that person's age and disease state. A random number is then sampled to determine whether the person survives. If they do, the probability of transition to the next state for each separate complication in the next year is calculated using the epidemiological data contained in the model. Another random number is then sampled to determine whether this transition actually occurs. Although in reality many patients will have multiple complications, the authors explain that the model does not collect data on compound health states "because of computer memory constraints". The model uses retinopathy data from WESDR and from the Rochester Diabetic Neuropathy Study [21]. The authors claim that the model is conservative. They discuss problems of validation, for example for particular ethnic groups. Although spreadsheets have the advantage of transparency, they are not a good medium for modeling individuals. A large number of parameters must be entered to describe the attributes of each individual, and it is time consuming to sample probabilities for each patient in each time period. Short time periods waste computation time but if the time periods are long, important transitions may be missed. Custom simulations with the same type of structure in high-level languages will have similar problems. Discrete event simulations overcome these problems because they are designed to describe individuals and progress from one individual patient event to the next, in time order.

All the models except those of Bachman and Nelson [13] and Tomar et al. [14] were based on incident cohorts of the populations who progressed through the model. All models except that of Tomar et al. [14] omitted current prevalent populations (who inevitably make a considerable demand on future resources) and did not include newly incident populations.

The models make no claim that the screening saves lives. The benefits must therefore be measured in terms of cases of eye disease detected, the number of people prevented from going blind, sight years saved or QALYs gained. The advantage of using QALYs is that results can be compared across different health systems. The disadvantage is that researchers use different quality adjustment values for the same health condition. Cost-effectiveness ratios are sensitive to the quality values that are chosen. Costs may also be difficult to determine; in particular, many of the studies cite difficulties in determining the cost of blindness.

The results from most of the studies indicate that annual screening for diabetic retinopathy is beneficial. The results from Vijan et al. [15], indicating that two-yearly screening after the detection of background retinopathy would be adequate for low risk patients, are controversial [23]. The authors claim that this conclusion is robust to the choice of the value of the quality adjustment for blindness. It is doubtful, however, whether it is realistic to split patients into risk groups based on glycemic control, as the risk group of any patient may vary over time.

We identified a need for a robust modeling approach that is capable of taking into account multiple diseases and that can evaluate a variety of screening policies that might be appropriate in the UK. The model should be able to compare the provision of different screening methods, different treatment modes and different levels of patient compliance.

19.3 METHODOLOGY USED

19.3.1 The simulation software

The model we created uses discrete event simulation which describes the progress of individuals from one event to the next. In our model, the individuals (entities) are patients who progress through different disease states, being screened, treated and so on. An event would be the start or finish of a disease state, treatment or investigation. All events take place in time order. Discrete event simulation is based on the concept of queuing networks, where entities progress through a series of queues and activities. Queues arise when events are constrained, such as admission to hospital. The major advantage of using discrete event simulation is that entities can be given characteristics that influence their flow through the system. In a disease simulation, these characteristics might include age, co-morbidity and patient disease history. The internal structure of a discrete event simulation package manages time advance, sampling and data collection.

We used a variant of discrete event simulation called patient oriented simulation technique (POST) [24, 25], available in Borland Delphi, using Pascal procedures and functions. In POST, an entity "owns" event notices and queue links that engage in events and wait in queues on the entity's behalf. This means that an entity can take part different events and wait in several queues at the same time. A strongly linked program structure enables event notices to be rapidly removed from events, or queue links to be removed from queues, as needed. An extreme example is death; if an event notice belonging to a patient entity is deemed to have reached an event representing death, then all the other event notices and queue links are withdrawn. Such a structure also enables screening to be modeled independently of the natural history of disease [28].

19.3.2 Natural history states

The simulation described in this chapter was designed to evaluate the cost-effectiveness of laser treatment in preventing severe sight loss in patients with Type 1 and Type 2 DM. More information can be obtained from our published papers and website [26-30]. The model describes progression through the two main types of eye conditions. Klein et al. [3, 8] have defined in detail the stages through which eyes progress in the natural history of the disease: these comprise 10 stages of progression from normal through proliferative retinopathy to unclassifiable and nine stages of progression for diabetic macular edema. In order to determine the cost-effectiveness of screening, we do not need so much detail. The important states to identify are those that set off (or may set off) some other activity. These are: no retinopathy, background retinopathy, proliferative diabetic retinopathy, macular edema, clinically significant macular edema and severe vision loss where sight is less than 20/200 on the Snellan scale [4].

19.3.3 Screening

In the model, the screening process starts as soon as a patient is diagnosed with diabetes. In practice, screening may take place in the community at a general practice surgery, at an optometrist or in a van using a mobile camera, or it might take place in hospital, either by a diabetologist or by an ophthalmologist. When problems are detected, we assume that patients are assessed by an ophthalmologist with the best equipment available, providing a "gold standard" service with 100% sensitivity and 100% specificity. We have previously assessed the relative merits of different modes of screening [29]. In this chapter, we will look at different levels of screening sensitivities without relating them to particular locations or techniques.

We divided the screening process into two parts:

1. primary screening until any retinopathy has been detected, which will normally be background retinopathy unless the eye disease has already progressed past that point, and

2. subsequent screening for those with background retinopathy until treatable retinopathy is detected, the patient has severe vision loss or dies.

Once any retinopathy is detected after primary screening, patients continue to the second part of screening if background retinopathy is confirmed in an ophthalmology clinic. If the detection is a false positive then the patient returns to primary screening. Once treatable retinopathy is detected, a patient is again assessed in an ophthalmology clinic and, if the diagnosis is confirmed, is treated; a patient with a false positive detection returns for more screening. Treated patients continue to be screened until they have been treated for both maculopathy and retinopathy, they suffer severe vision loss or they die.

In this chapter we will compare four modes of screening:

1. Primary screening takes place once a year and subsequent screening takes place six monthly at an ophthalmology clinic (gold standard);

2. Primary screening takes place once a year and subsequent screening takes place six monthly using the same screening mode as for the primary screening;

3. Primary screening and subsequent screening take place once a year using the same screening mode;

4. Primary screening takes place once every two years and subsequent screening takes place once a year using the same screening mode as for the primary screening.

Within these modes we varied sensitivity and patients' compliance with screening. We found that sensitivity may be as low as 52% for general practice screening [31] compared to over 80% for hospital screening [32]. James et al. [16] found compliance with screening to be 80%. It makes a difference, however, as to whether non-attendance is random or systematic.

We assumed that, among those who occasionally missed a screen, non-attendance was random, with compliance of 85% for Type 2 DM and 95% for Type 1 DM patients. We assumed that a further 5% of patients overall never attended any screening. The resulting 80% compliance for Type 2 DM and 90% compliance for Type 1 DM patients yielded an average of 82% compliance overall.

19.3.4 Model structure

Two very similar models were developed, one for Type 1 and one Type 2 DM. The main difference, apart from the parameters, was that the Type 2 model allowed for a proportion of patients to have some retinopathy at diagnosis, whereas Type 1 patients were assumed to have no retinopathy. The models describe the progression of diabetes from diagnosis to death (or the end of the simulation).

Age and the state of the eye disease influence progression through the simulation. Figure 19.1 shows that the two types of sight threatening retinopathy (clinically significant macular edema (CSME) and proliferative diabetic retinopathy (PDR)) may occur together or separately. Progression to sight threatening retinopathy may be interrupted by treatment.

In describing transitions between disease states (with the exception of death), the following exceptions were made to the Markovian assumption:

- The initial results from simulation of the original IDDM model [26] showed that the prevalences of BDR and PDR were too high in the first few years after diagnosis. It is extremely rare for patients to get diabetic eye disease before puberty (taken to be at age 12) or within the first two years after diagnosis, and evidence from the Wisconsin study [9] indicates that very few people get PDR or diabetic macular edema (DME) within the first five years after diagnosis. The sampled times from diabetes diagnosis to BDR therefore included an initial period with a low probability of changing state and a later period with a higher probability.

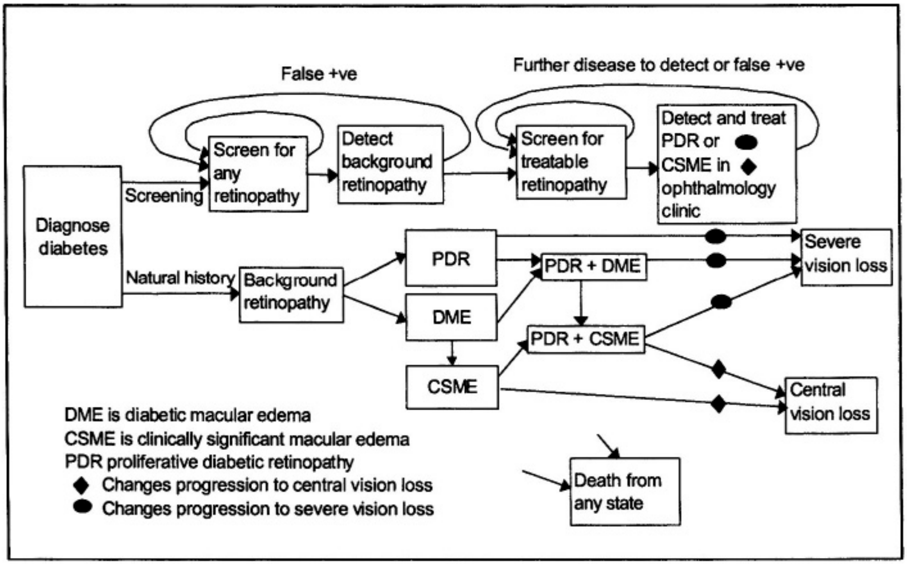
- Time to death depends on a patient's current age and retinopathy status and so as people progress through the stages of retinopathy, their time to death is re-sampled.

Once a patient enters the simulation, screening is started. The simulation models the screening process described in the previous section and shown in outline in Figure 19.1. One attribute of a patient entity is the state of the patient's sight, as determined by the patient's natural history, described above. POST enables screening to take place independently of the natural history, but responds to changes in it.

The effect of treatment is described by re-sampling the time to vision loss using a larger average value in the natural history part of the model. Death causes all of an entity's activities to be terminated.

The models can incorporate an initial screening round in which it is assumed that no one has previously been screened, in which case a large number of prevalent cases will be detected by the first screen. Alternatively, the models can use a warm-up period in order to evaluate screening programs that have been running for a number of years.

Figure 19.1 Flow of patients through the simulation. Patients may get proliferative retinopathy or macular edema and may subsequently acquire both. Background retinopathy, PDR and CSME are only detected when present or when there is a false positive. Treatment takes place if PDR or CMSE is detected but not if a false positive.


False +ve
Further disease to detect or false +ve
Screen for any retinopathy
Detect background retinopathy
Screen for treatable retinopathy
Detect and treat PDR or ● CSME in ◆ ophthalmology clinic
Diagnose diabetes
Screening
Natural history
Background retinopathy
PDR
DME
CSME
PDR + DME
PDR + CSME
Severe vision loss
Central vision loss
Death from any state
DME is diabetic macular edema
CSME is clinically significant macular edema
PDR proliferative diabetic retinopathy
◆ Changes progression to central vision loss
● Changes progression to severe vision loss

19.4 DATA SOURCES

19.4.1 Prevalence, incidence and transitions between states

The data read into the simulation includes:

- both the population breakdown and the prevalence of diabetes, by age and sex and ethnic origin, and the prevalence of eye disease by age and sex;
- the incidence of diabetes by age and sex and ethnic origin and the incidence of eye disease and vision loss;
- mortality by age, sex and by eye disease;
- the characteristics of the screening program and the efficacy of treatment.

Estimates of parameter values were obtained from extensive literature reviews. Many of the details of the data used in the models have been published elsewhere [26-29]. The data, and a description of the models, can also be found on our web site [30]. WESDR was the main source of epidemiological data on retinopathy.

19.4.2 Costs

The costs of screening and treatment were derived from the National Health Service National Screening Committee website [33]. These indicate that screening using a mobile camera costs approximately £20 per patient, with set-up costs averaging about £11 per patient, the cost of a visit to an optometrist, £20, an ophthalmology outpatient visit, £60; and treatment, £180 on each occasion. The quality assurance costs were estimated to be less than £1 per patient. If the set-up costs for the mobile van are spread over five years, then the cost per patient should be around £22. This is similar to James et al. [16] who showed that the cost of a screening visit varies from an average of £19 per visit to £25 per visit, depending on the mode of screening. The outpatient costs for an ophthalmic visit at £60 compare well to Netten and Dennett [34] who calculated the cost of a general outpatient visit to be £55. Screening costs for a diabetologist or a general practitioner are likely to be higher than this but, as screening would only be one part of the clinic activities, it would be difficult to attribute an exact cost to this. Three treatment sessions per eye were allowed for treatment of proliferative retinopathy and 1.5 sessions per eye for macular edema, which reflects typical practice in the UK [35]. In addition to the visits for each treatment session, we assumed one follow-up visit per course of treatment.

19.4.3 Quality-of-life estimates

A QALY is a measure that combines both morbidity and mortality. The measure is based on a utility that is normally between 0 and 1 (where 0 is death and 1 is perfect health). If someone with a condition with a utility of *q* (where q is between 0 and 1) is given a treatment that provides one year of perfect health, then (1-*q*) QALYs are gained. If someone with this condition is given a treatment that delays death by one year but does not change quality of life, then *q* QALYs are gained.

In our previous studies we used sight years saved. QALYs gained have the advantage that they allow results to be compared to studies of other health care programs. QALYs do, however, entail some major assumptions. First, they treat mortality, morbidity and quality of life as commensurate, so that one can be traded off against the other. Second, disagreement exists as to whether the utilities should be derived from surveys of doctors, patients or the public. Third, the methods used in surveys to elicit utilities vary. For example, the Visual Analog Scale (VAS), Standard Gamble (SG), Time Trade Off (TTO), and Person Trade Off (PTO) can all give different values [36].

Many modeling studies are not specific about how the quality adjustments are derived. For example, several studies [15, 22, 37] use 0.69 for the utility of blindness but this choice is not supported by reference to any literature that justifies this number. Javitt and Aiello [20] use 0.45, citing a 1987 consultancy report by Drummond to the National Eye Institute [38]. Torrence and Feeny [39] report a utility of 0.39 for blindness. Brown et al. [40] and Brown [41] suggest that this value may be appropriate for complete blindness with no bilateral light perception. Their research indicates that the utility should vary for the degree of sight loss. For patients with diabetic retinopathy, they used the TTO methodology and found a utility of 0.66 for 20/200 to 20/400 vision and 0.54 for the ability to count fingers or perceive light. Patients who have DME or PDR are, however, unlikely to have perfect vision even with laser treatment. If we use Brown's value of 0.84 for 20/20 to 20/25 vision, then the difference between the lower utility of blindness and of 20/20 or 20/25 vision will be 0.84-0.54 which is 0.30. This is a similar utility difference to that used by many other of the other modelers.

19.4.4 Discounting costs and benefits

In a screening program, the costs (staffing, equipment, stationery and so on) are incurred immediately, whereas the benefits (years of blindness averted) may not be accrued for many years into the future. Discounting of future costs and benefits can help to evaluate this imbalance. Unfortunately there is little consensus as to what discount rate should be used. Gold et al. [42] recommend using 3% for costs and benefits, whereas, in the UK the National Institute for Clinical Excellence [43] recommends using 6% for costs and 1.5% for benefits. These values can produce very different results.

19.5 RESULTS

19.5.1 The model population

The model can be run with an incident cohort population, with a prevalent population or with a prevalent population plus a new incident population each year. It was usually run in the latter mode as this was the most realistic. The prevalent population was assumed to be unscreened. The simulation was run for five years until most of the outstanding patients were screened and the model was in steady state. In order to evaluate the benefits of screening, results were collected from the steady state model.

For the cost-effectiveness analyses, we could not use a steady state analysis because current benefits might relate to past treatment. For these runs, at the start of the data collection, we simulated a cross-section of prevalent patients excluding all those with treatable disease and those who had suffered vision loss and followed the cohort until they had all died. The discounting of costs and benefits was thus able to respond to the delay in realization of benefits.

All the simulation runs were performed for Type 1 and Type 2 diabetes.

19.5.2 Choice of base scenarios

Gold et al. [42] recommend that the modeled scenario should be compared to the current scenario rather than a scenario with no interventions. The current situation differs so much from place to place, however, that in common with most modelers in this field, we have related the modeled scenario to a no-screening scenario.

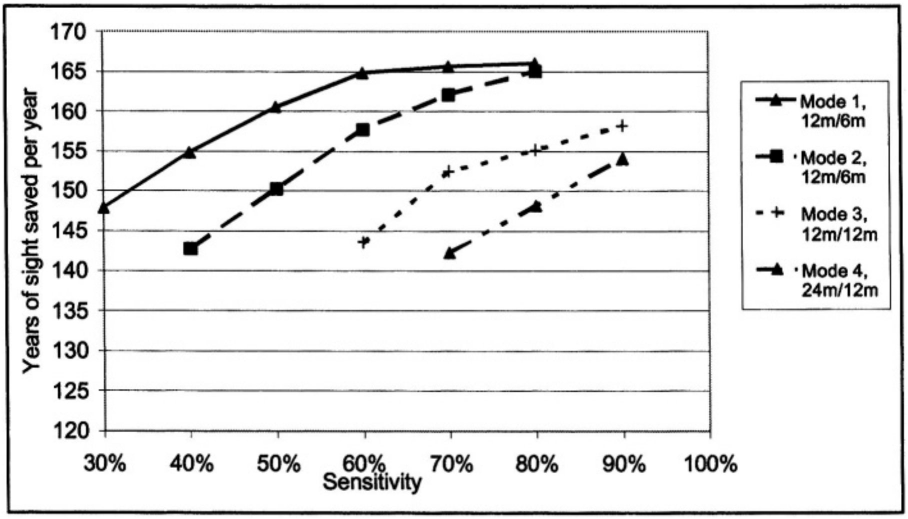
The results were averaged over 500 replications of the simulation in order to reduce the standard deviation of the average years of sight saved to within 1.0% of the estimated mean.

19.5.3 Results from runs

We generated results for Type 1 and Type 2 DM separately. Previously we showed [29] that if the same policies are adopted for both groups, approximately two thirds of the sight years saved are among Type 2 patients and one third are among Type 1 patients. However, screening was more cost effective for Type 1 than Type 2 patients. Below we present the results for Type 1 and Type 2 patients together for the four screening models listed in Section 19.3.3.

In Figure 19.2 we excluded combinations of screening sensitivities and screening intervals which produced results below 85% of the maximum. It shows the effect of different levels of screening sensitivity on years of sight saved. These variations might represent different screening methods and personnel. The more accurate methods might be expected to be more expensive. For screening methods with lower levels of sensitivity it becomes more important to screen more frequently and the advantage conferred by having secondary screening in the ophthalmology clinic increases. Figure 19.3 shows the effect of reduced compliance and reduced sensitivity on years of sight saved. When screening is frequent (i.e. for modes 1 and 2) patient compliance and test sensitivity have a small effect on years of sight saved. The effect is more pronounced when screening is less frequent (i.e., for modes 3 and 4).

Figure 19.2 The effect on years of sight saved of different screening policies and screening test sensitivity for a population of 500,000 individuals. Compliance is assumed fixed at 82%. The intervals are those before and after background retinopathy is detected, respectively.



In this model, screening and treatment for sight problems was assumed to have no effect on health apart from its effect on sight, In practice, screening by a diabetologist or a general practitioner might be expected to be associated with medical services, such as consultations leading to better compliance with insulin, and screening for other complications. The QALYs gained are thus conservative and in direct proportion to years of sight saved. Sight years saved can therefore can be converted into QALYs gained. For example, for a utility of blindness of 0.7 (see Section 19.4.3), the QALY gain is 0.3 per sight year saved. The costs in Figure 19.4 are based on institutional costs only, excluding the costs of blindness. The cost of screening is the cost of using the mobile camera which is, in practice, rather low for the better screening sensitivities, such as 80%, where more expensive equipment might be needed.

Figure 19.4 shows that the first screening policy, in which follow-up screening was in the outpatient clinic, was much more expensive than the other three, in which all the screening took place in the community. Higher discount rates for both costs and benefits gave rise to higher cost-effectiveness ratios. The effect of a high discount rate for costs (6%) and a low rate for benefits (1.5%) was to reduce the cost-effectiveness ratio below the result in which there is no discounting.

19.5.4 Discussion of results

The results show the relative importance of screening sensitivity, patient compliance and screening intervals on the cost-effectiveness of screening for and treating eye disease caused by diabetes. Low values in one of these parameters can be offset, to some extent, by the others. This arises because the lower the sensitivity of screening and the lower the compliance, the more likely it is that disease is missed. Less frequent screening results in a longer time, on average, before a problem is picked up and treatment takes place. When these effects occur together, the problem is compounded.

The results indicate that it is cost effective to screen in the community. Where sensitivity and compliance are high, longer screening intervals appear beneficial but there is little difference in years of sight saved over the range of intervals we have chosen If the sensitivity were lower and the costs remained the same, this difference would be decreased or reversed and shorter time intervals for screening would appear to be relatively more cost effective.

It would be desirable to include the cost of blindness but this is difficult to estimate. An Australian study [44] suggests that the cost of a case of blindness to the welfare state is about $14,656 Australian, i.e. about £6,000.

Figure 19.3 The effect on years of sight saved of changing both sensitivity and compliance for the different policies. The intervals are those before and after background retinopathy is detected, respectively.

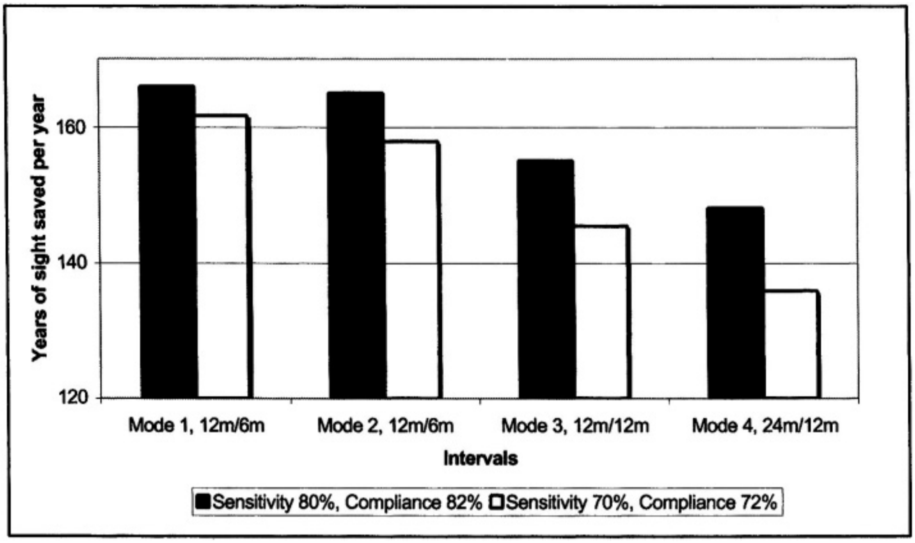
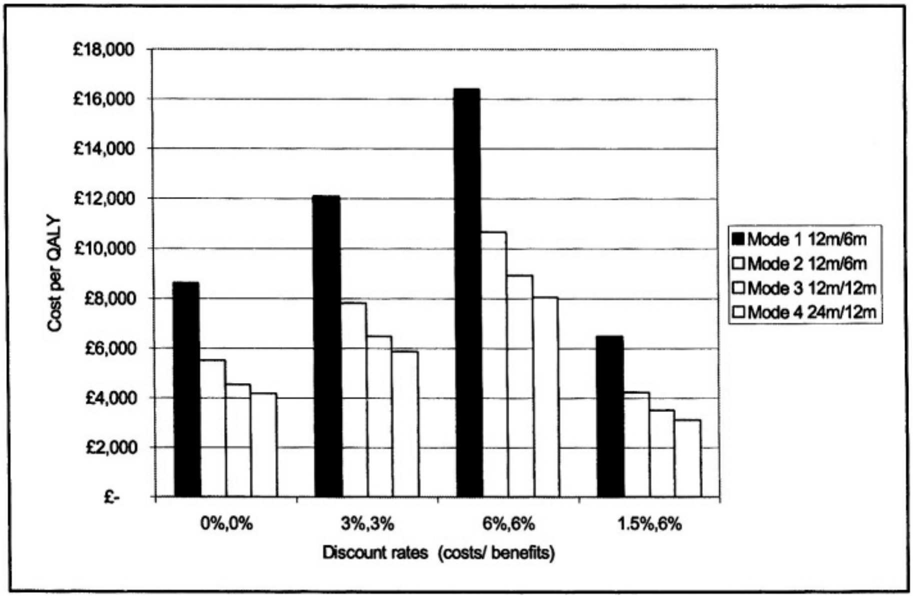


Figure 19.4. Cost-effectiveness of treatment using a utility of 0.7 for blindness



If this cost were included in the analysis then screening would appear cost saving. However, most of this cost is the cost of a disability pension much of which would be available to the elderly anyway.

We have found that it is difficult to follow the reference case costing method recommended by Gold et al. [42]. It has not been possible to find the costs of blindness incurred by the state or individual. Furthermore, we compared screening scenarios with no-screening scenarios rather than opportunistic screening and treatment, because policies varied so much from place to place. From this point of view we have overestimated the benefits of screening. We have, however, developed a robust population model of the progression of diabetic retinopathy, that can evaluate the impact of screening and treatment, including opportunistic treatment, where the data are available.

19.6 AVENUES FOR FURTHER RESEARCH AND CONCLUSIONS

19.6.1 Other complications of diabetes

Eastman et al. [22, 37] and Tomar et al. [14] have published models that describe the different complications of diabetes. Eastman et al. do not consider compound health states; they assume that different complications are independent of each other. This assumption is likely to be unrealistic. Interdependencies are described in Tomar et al.'s model but a very limited number of combinations are allowed. Interdependencies are relatively easy to describe in a discrete event simulation where individuals have attributes that influence their progress through the system. The major problem is to provide the data to drive such a model when the interdependencies are not fully understood. This interdependency will extend to the risk of death, which in our current model is related to the eye condition. A full model of diabetes progression would make it possible to assess methods of controlling blood sugar levels and treatments to prevent the complications of diabetes. It would be possible to assess the benefits of screening for diabetes as well as for any individual complication. Discrete event simulation would be an appropriate technique for doing this.

19.6.2 Costs and QALYs

Further work is needed to assess costs, particularly the costs associated with blindness in older people. There also needs to be some further research and rational discussion about the appropriate values of quality-of-life utilities to be used in these models. The problems in determining QALYs, finding the current opportunistic use of resources and interpreting the results from discounting the costs and benefits are present in many, if not all, modeling studies leading to cost-effectiveness analyses [45].

19.6.3 Other applications

Discrete event simulations of health policy issues follow individuals over long periods of time, possibly a lifetime. The individuals have characteristics that influence their progress through the system and their response to treatments. Simulations of thousands of individuals may take just a few minutes, but multiple iterations are needed in order to get statistically significant results.

Some of the areas in which discrete event simulation has been used are as follows:

- The evaluation of services for patients with end stage renal failure [46];
- Determination of service requirements for patients with AIDS [47];
- Screening for helicobacter pylori to prevent peptic ulcers and gastric cancers [48];
- Interventions for the prevention of coronary heart disease [49];
- Treatment and secondary prevention in coronary heart disease [50].

All of our recent models [46, 48, 50] and the coronary heart disease prevention model from the London School of Hygiene and Tropical Medicine [49] have been developed with the POST simulation software. This software enables an entity to be engaged in more than one activity, or to be present in more that one queue at once. This is a significant advantage because we can describe a patient's natural history of the disease and, independently (or not as we wish) provide screening or interventions.

We have discussed many of the benefits of discrete event simulation. Despite these benefits, this method is not widely used for cost-effectiveness analyses, perhaps because it is perceived to be time consuming, difficult to do and lacking in transparency.

Simulation runs are indeed time consuming, but run times have reduced significantly with the introduction of more powerful computers, and the number of runs can be reduced by the careful use of variance reduction techniques.

There are now many interactive graphical packages that make discrete event simulation transparent and easy to use. One drawback of these packages is that, for any application, an interface is needed to deal with the extensive data requirements. This may need to be programmed or linked to a spreadsheet. We have discussed the advantages of using POST. Each of our applications is “user-friendly” but further research is needed to enable this package to provide a generic data and graphical interface for health policy models.

19.6.4 Conclusion

Discrete event simulation, using POST, enables us to explore the interaction between screening and the natural history of disease. It is possible to model different and interacting disease processes, such as macular edema and proliferative retinopathy. Discrete event simulation is a powerful tool for decision support for policy decisions and, in particular, for cost-effectiveness analyses in a wide variety of health service applications.

Acknowledgements

Chris Canning, a consultant ophthalmologist in the Southampton University Hospitals Trust, and Paul Roderick, a Senior Lecturer at the University of Southampton, have made substantial contributions to this research.

References

[1] Gatling, W., S. Budd, D. Walters, M.A. Mullee, J.R. Goddard, and R.D. Hill (1998). Evidence of an increasing prevalence of diagnosed diabetes mellitus in the Poole area from 1983 to 1996. *Diabetic Medicine,* 15,1015-1021.

[2] Early Treatment Diabetic Retinopathy Study Research Group (1991). Early photocoagulation for diabetic retinopathy: ETDRS report No 92. *Ophthalmology,* 98, 766-785.

[3] Klein, R., B.E.K. Klein, S.E. Moss SE, and K.J. Cruikshanks (1994). The Wisconsin Epidemiologic Study of Diabetic Retinopathy. XIV. Ten-year incidence and progression of diabetic retinopathy. *Archives of Ophthalmology,* 112, 1217-28.

[4] Klein, R., B.E.K. Klein, and S.E. Moss (1984). Visual impairment in diabetes. *Ophthalmology,* 91, 1-9.

[5] Wilson, J.M.G. and O. Junner (1968). The principles and practice of screening for disease. *Public Health Papers,* 34, Geneva: WEW.

[6] Bagga, P., D. Verma, C. Walton, et al. (1998). Survey of diabetic retinopathy screening services in England and Wales. *Diabetic Medicine,* 15, 780-782.

[7] NHS Centre for Reviews and Dissemination (1999). Complications of diabetes: screening for retinopathy, management of foot ulcers. *Effective Health Care*, 5.

[8] Klein, R., S.E. Moss, B.E.K. Klein, M.D. Davies, and D.L. De Mets (1989). The Wisconsin Epidemiologic Study of Diabetic Retinopathy XI. The incidence of macular edema. *Ophthalmology,* 96, 1501-1510.

[9] Klein, R., B.E.K. Klein, S.E. Moss, M.D. Davis, and D.L. De Mets (1989). The Wisconsin Epidemiologic Study of Diabetic Retinopathy IX. Four-year incidence and progression of diabetic retinopathy when age at diagnosis is less than 30 years. *Archives of Ophthalmology,* 107, 237-243.

[10] Diabetic Retinopathy Study Research Group (1981). Photocoagulation treatment of proliferative diabetic retinopathy: clinical application of Diabetic Retinopathy Study (DRS) findings. DRS Report No. 8. *Ophthalmology,* 88, 583-600.

[11] UK Prospective Diabetes Study (UKPDS) Group (1991). UK Prospective Diabetes Study. VII. Study design, progress and performance. *Diabetologia,* 34, 877-890.

[12] Dasbach, E.J., D.G. Fryback, P.A. Newcomb, R. Klein and B.E.K. Klein (1991). Cost-effectiveness of strategies for detecting diabetic retinopathy. *Medical Care,* 29, 20-39.

[13] Bachmann, M. and S. Nelson (1998). Impact of diabetic retinopathy screening on a British district population: case detection and blindness prevention in an evidence-based model. *Journal of Epidemiology and Community Health,* 52, 45-52.

[14] Tomar, R.H., S. Lee, S.-Y. Wu, R. Klein, B.E.K. Klein, S.E. Moss, D.G. Flyback, J.L. Tollios and F. Sainfort (1998). Disease progression and cost of insulin dependent diabetes mellitus: development and application of a simulation model. *Journal of the Society for Health Systems,* 5, 24-37.

[15] Vijan, S., T.P. Hofer, and R.A. Hayward (2000). Cost utility of screening intervals for diabetic retinopathy in patients with Type 2 Diabetes Mellitus. *Journal of the American Medical Association,* 283, 889-896.

[16] James, M., D.A. Turner, D.M. Broadbent, J. Vora, and S.P. Harding (2000). Cost effectiveness of screening for sight threatening diabetic eye disease. *British Medical Journal,* 320, 1627-1631.

[17] Javitt, J.C., J.K. Canner, and A. Sommer (1989). Cost effectiveness of current approaches to the control of retinopathy in Type 1 diabetics. *Ophthalmology*, 96, 255-264.

[18] Javitt, J.C., L.P. Aiello, L.J. Bassi, Y.P. Chiang and J.K. Canner (1991). Detecting and treating retinopathy in patients with Type 1 diabetes mellitus. *Ophthalmology,* 98, 1561-1573.

[19] Javitt, J.C., L.P. Aiello, Y.P. Chiang, F.L. Ferris, J.K. Canner, and S. Greenfield (1994). Preventive eye care in people with diabetes is cost-saving to the federal government. *Diabetes Care,* 17, 909-917.

[20] Javitt, J.C. and L. Aiello (1996). Cost-effectiveness and detecting and treating diabetes. *Annals of Internal Medicine,* 124, 164-169.

[21] Dyck, P.J., K.M. Kratz, J.L. Kames, W.J. Litchy, R. Klein, J.M. Pach, D.M. Wilson, P.C. O'Brien, and L.J. Melton. III (1993). The

prevalence by staged severity of various types of diabetic neuropathy, retinopathy and nephropathy in a population-based cohort: The Rochester Diabetic Neuropathy Study. *Neurology,* 43, 817-824.

[22] Eastman, R.C., J.C. Javitt, W.H. Herman, E.J. Dasbach, A.S. Zbrozeh, F. Dong, et al. (1997). Model of complications of NIDDM. Model construction and assumptions. *Diabetes Care*, 20,725-24.

[23] Javitt, J.C. (2000). How often should patients with diabetes be screened for retinopathy? (letter). *Journal of the American Medical Association,* 284, 437.

[24] Davies, H.T.O. and R. Davies (1995). Simulating health systems: modelling problems and software solutions. *European Journal of Operational Research,* 87, 35-44.

[25] Davies, R. and H.T.O. Davies (1994). Modelling patient flows and resource provision in health systems. *Omega International Journal of Management Science,* 22, 123-131.

[26] Davies, R., P. Sullivan, and C. Canning (1996). Simulation of diabetic eye disease to compare screening policies. *British Journal of Ophthalmology,* 80, 945-950.

[27] Brailsford, S.C., R. Davies, C. Canning, and P. Roderick (1998). Evaluating screening policies for the early detection of retinopathy in patients with non-insulin-dependent diabetes. *Health Care Management Science,* 1, 115-124.

[28] Davies, R., S.C. Brailsford, P. Roderick, C. Canning, and D. Crabbe (2000). Using simulation modeling for evaluating screening services for diabetic retinopathy. *Journal of the Operational Research Society,* 51, 476-484.

[29] Davies, R., P. Roderick, C. Canning, and S.C. Brailsford (2002). The evaluation of screening policies for diabetic retinopathy using simulation. *Diabetic Medicine,* 19(9), 763-771.

[30] Brailsford, S.C. and R. Davies (2001). *Screening for Diabetic Retinopathy.* www.management.soton.ac.uk/retinopathy/

[31] Reenders, K., E. de Noble, H. van den Hoogen, and C. van Weel (1992). Screening for diabetic retinopathy by general practitioners. *Scandinavian Primary Health Care,* 10, 306-309.

[32] Pugh, J.A., J.M. Jacobson, W.A. Van Heuven, et al. (1993). Screening for diabetic retinopathy. The wide angle retinal camera. *Diabetes Care,* 16, 889-895.

[33] NHS NSC Diabetic Retinopathy Screening (2001). *Resources for a National Diabetic Risk-reduction Programme,* www.diabetic-retinopathy.screening.nhs.uk/costings.html.

[34] Netten, A. and J. Dennett (1996). *Unit Costs of Community Care.* Personal Social Services Research Unit, University of Kent.

[35] Bailey, C.C., J.M. Sparrow, R.H. Grey, and H. Cheng (1999). The National Diabetic Retinopathy Laser Treatment Audit. III. Clinical outcomes. *Eye,* 13, 151-159.

[36] Raftery, J. (2000). Methodological limitations of cost effectiveness analysis on health care: implications for decision making and service provision. *Journal of Evaluation in Clinical Practice,* 5, 361-365.

[37] Eastman, R.C., J.C. Javitt, W.H. Herman, et al. (1997). Model of complications of NIDDM II, analysis of the health benefits and cost-effectiveness of treating NIDDM with the goal of normoglycaemia. *Diabetes Care,* 20, 735-744.

[38] Dummond, M. (1987). *Consultant Report to the National Eye Institute.* National Eye Institute, Bethesda, MD.

[39] Torrance, G.W. and D. Feeny (1989). Utilities and quality adjusted life years. *International Journal of Technology Assessment in Health Care,* 2, 559-575.

[40] Brown, M.B., G.C. Brown, S. Sharma, and G. Shah (1999). Utility values and diabetic retinopathy. *American Journal of Ophthalmology,* 128, 324-330.

[41] Brown, G.C. (1999). Vision and quality of life. *Transactions of the American Ophthalmology Society,* 92, 474-511.

[42] Gold, M.R., J.E. Siegel, L.B. Russell, and M. Weinstein, Eds. (1996). *Cost Effectiveness in Health and Medicine.* Oxford University Press, New York,

[43] National Institute for Clinical Excellence (2001), *Guidance for Manufacturers and sponsors.* www.nice.org.uk/Docref.asp?d=16183.

[44] Wright, S.E., J.E. Keeffe, and L.S. Thies (2000). Direct costs of blindness in Australia. *Clinical and Experimental Ophthalmology,* 28, 140-142.

[45] Davies, R., P. Roderick, and J. Raftery (2003). The evaluation of disease prevention and treatment using simulation models. *European Journal of Operational Research,* 150, 53-66.

[46] Davies, R. and P. Roderick (1998). Planning resources for renal services throughout England using simulation. *European Journal of Operational Research,* 105, 285-295.

[47] Brailsford, S.C., A.K. Shahani, R. Basu Roy, and S. Sivapalan (1996). Practical models for the care of HIV and AIDS patients. *International Journal of STD and AIDS,* 7, 91-97.

[48] Davies, R., R. Roderick, D. Crabbe, J. Raftery, P. Patel, and J.R. Goddard (2002). A simulation to evaluate screening for helicobacter pylori infection in the prevention of peptic ulcers and gastric cancers. *Health Care Management Science,* 5, 249-258.

[49] Babad, H., C. Sanderson, B. Naidoo, I. White, and D. Wang (2002). The development of a model for the prevention of coronary heart disease. *Health Care Management Science*, 5, 269-274.

[50] Cooper, K., R. Davies, P. Roderick, D. Chase, and J. Raftery (2002). The development of a simulation model for the treatment of coronary artery disease. *Health Care Management Science,* 5, 259-267.

20 DECISION MAKING FOR BIOTERROR PREPAREDNESS: EXAMPLES FROM SMALLPOX VACCINATION POLICY

Edward H. Kaplan[1] and Lawrence M. Wein[2]

[1] Yale School of Management, and
Department of Epidemiology and Public Health, Yale School of Medicine
Yale University
New Haven, CT 06520

[2] Graduate School of Business
Stanford University
Stanford, CA 94306

SUMMARY

Decision making for bioterror preparedness involves estimating the consequences of different attack scenarios paired with alternative preparedness and response policies, and selecting an appropriate strategy to minimize deaths, disease, and costs to society. Smallpox vaccination policy provides an excellent case study of these concepts in action. We review both the smallpox vaccination policy debate in the United States circa 2002, and the successful use of operations research methods to influence policy in this arena.

KEY WORDS

Bioterrorism, Smallpox, Vaccination policy, Decision analysis

20.1 INTRODUCTION

While the possible terrorist use of biological agents against civilian targets was a subject of concern before September 11, 2001 [1, 2], the deliberate mailings of anthrax-laden letters and resultant infections, deaths and panic that ensued in October of that same year provided a bioterror "proof of concept." Policy makers in the United States and abroad suddenly found themselves faced with numerous decisions regarding bioterror preparedness, and what to do about smallpox quickly topped the list of concerns. Smallpox vaccination policy provides a very interesting case study for those interested in the use of OR/MS methods in confronting bioterror and health issues more generally, especially since OR modeling influentially informed the policymaking process in this instance.

This chapter reports some of the simple yet powerful arguments employed in the smallpox vaccination policy debate. First described are the issues and positions held during this debate, emphasizing the clash between the "bio" and the "terror" in "bioterror." Following this, three simple choice models that were used repeatedly in the arguments are considered. A key common idea across these examples is the importance of focusing on the consequences of decisions made rather than the details of what is believed to be the most likely smallpox scenario, as will become clear shortly. Concluding comments follow.

20.2 VACCINATION POLICY IN THE LARGE AND IN THE SMALL(POX)

The eradication of smallpox surely stands as one of the greatest public health achievements of the last century [3]. However, fear that smallpox could be employed as a weapon of bioterror prompted public health and security officials in the United States and elsewhere to revisit the issue [1, 2, 4]. As of November 2001, the Centers for Disease Control and Prevention (CDC), acting upon the advice of the Advisory Committee on Immunization Practices (ACIP), had developed a smallpox emergency response policy [5]. Absent any analysis other than analogy to the procedures employed by the World Health Organization (WHO) in the global smallpox eradication campaign, the CDC plan called for a ring vaccination response: confirmed smallpox cases would be isolated, their close contacts would be traced and vaccinated, contacts found febrile would be quarantined, while asymptomatic contacts would monitor and report their temperatures to local public health officials. Only if this policy failed to contain a smallpox outbreak would an (unspecified) "broader" vaccination response be initiated.

It did not take long for this policy to come under intense public scrutiny. In an influential article published in the *New England Journal of Medicine,* Dr. William Bicknell, the former Commissioner of Health for the State of Massachusetts, argued the case for voluntary pre-attack vaccination against smallpox [6]. Noting that widespread vaccination would greatly reduce the value (to terrorists) of a smallpox attack, Bicknell further noted the lack of population immunity against smallpox in the United States today (in contrast with the much higher levels of immunity due to prior vaccination campaigns and survival from disease in those countries where the WHO finally eradicated smallpox). Were ring vaccination employed, Bicknell worried that "An epidemic is highly likely to outrun the vaccinators."

In this same issue of the *New England Journal,* Dr. Anthony Fauci, the highly respected Director of the National Institute of Allergy and Infectious Diseases (NIAID), voiced similar concerns [7]. He wrote that "...there is considerable skepticism about the feasibility of this strategy because of the possibility of simultaneous attacks in multiple cities." Fauci also noted the difference in the degree of population immunity against smallpox between the countries of the WHO campaign and the United States today, the logistical difficulties posed by ring vaccination, and the strategic advantage pre-attack vaccination would offer by removing the incentive for a smallpox attack.

However, both Drs. Bicknell and Fauci recognized an important rationale for limiting vaccination, namely, the potential for severe complications, including death, that could result from widespread use of the *vaccinia* anti-orthopox vaccine [8]. While Dr. Bicknell argued that one could proceed with a careful pre-attack vaccination plan that would screen out those at greatest risk of complications [6], Dr. Fauci called for "…an open and public dialogue on the advantages and disadvantages of universal voluntary vaccination, as well as on the smallpox response plan of the CDC" [7].

Such a public dialogue indeed unfurled, with much of the attention focused on a series of ACIP meetings that took place in June 2002. Meanwhile, at the invitation of Dr. Ellis McKenzie from the Fogarty International Center of the National Institutes of Health, the authors (jointly with David Craft) had developed a mathematical model that embedded the logistics of alternative emergency response policies into a smallpox disease transmission model [9]. The model and its attendant results had yet to be published in June, though they had been presented at a small number of public and private scientific meetings.

On June 15, ACIP held a public forum on smallpox vaccination policy at the Institute of Medicine in Washington, DC [10]. Following introductory remarks by Dr. Fauci and others, a review of the clinical features of smallpox, and the historical rationale for the CDC's proposed response policy, the basic results from our model were presented. As detailed in [9], the bottom line was clear: unless faced with a very small attack with an agent of only mild infectiousness, ring vaccination would result in many more deaths – perhaps by a factor of 200 – than would immediate mass vaccination in the area of the attack. Any degree of pre-attack vaccination would make any post-attack policy work better, but absent such pre-attack vaccination, the ring policy would almost surely fail [9, 10 pp. 18-20]. On the basis of this presentation at the Institute of Medicine forum, the first author was invited to brief government officials at the White House Conference Center on July 9.

Also presented at the public forum was Dr. Alan Zelicoff's detailed account of a smallpox outbreak in Aralsk, Kazakhstan in 1971 [10 pp. 20-21, 11]. This outbreak is important to understand, for it occurred among passengers on a research vessel that passed by the former Soviet bioweapons testing site at Vozrozhdeniye Island in the Aral Sea. As Dr. Zelicoff explained, "It is probably the case that smallpox was aerosolized, which answers the age-old question of whether or not smallpox is in fact aerosolizable and infectious in that state."

In spite of these findings, on June 20 ACIP recommended vaccinating only 15,000 first-responders nationwide [12]. However, this recommendation was quickly dismissed. On July 7, the *New York Times* ran a front-page story reporting that the federal Department of Health and Human Services was leaning towards recommending vaccinations for about 500,000 health care and emergency workers [13]. The story also summarized our main findings, Dr. Zelicoff's report, and provided the sense that a shift in thinking regarding the threat of smallpox along with control options was beginning to take hold.

On the afternoon of July 7, the first author received a phone call with an invitation to appear the next morning on the *Today Show*. In a lengthy interview with Katie Couric, the lessons from our analysis regarding the prospects for control of a smallpox attack were discussed. The next day, our results were presented in a three-hour closed-door meeting with representatives from the Office of the Vice President, National Security Council, Office of Homeland Security, Council of Economic Advisors, Health and Human Services, and other government agencies.

This meeting revealed clearly two competing schools of thought – the "bio" and the "terror" in bioterror – that have surfaced repeatedly in the smallpox debate. Adherents to the bio view consider smallpox preparation and response as just another public health problem, and not surprisingly, the bulk of the "bios" can be found in the medical and public health community, as well as agencies such as the Department of Health and Human Services and the Centers for Disease Control. Bios place great weight on historical smallpox outbreaks and control methods, and worry deeply about possible vaccine complications. This latter concern is consistent with the principle of *primum non nocere* (first do no harm), which typifies the beliefs of many health and medical practitioners.

However, those steeped in military strategy and homeland security, economics, public administration and policy making more generally, including some with medical training, expressed the more utilitarian view of getting the greatest good for the greatest number. For example, widespread vaccination against smallpox now could be viewed purely as a matter of strategy by depriving terrorists of an effective weapon. Or, minimizing the time required to bring an outbreak under control could be viewed as an important objective. In considering what to do about a smallpox *bioterror attack,* the historical record is less important than the terrorist imagination.

That these points of view conflict is obvious: to get the greatest good for the greatest number, it might be required to first do *some* harm. To vaccinate widely either before or after an attack, one must face the risk of vaccine complications including death.

Our analysis of emergency response to a smallpox attack appeared in the online *Proceedings of the National Academy of Sciences* on July 12 [9]. On July 14, the first author traveled to Israel, and over the course of two weeks met with leading Israeli officials to discuss smallpox response policy in a series of public and private meetings and presentations [14]. Israel is an interesting contrast to the United States, for essentially the same debate transpired there, but over a much shorter time period.

Our results were presented to CDC staff in Atlanta on August 9. This was a helpful visit, for many had misinterpreted [9] as a call for immediate pre-attack vaccination of the entire United States. While CDC officials continued to disagree with some assumptions and findings of that paper, these disagreements were now much more informed. Indeed, by the end of September, CDC for the first time released a smallpox mass vaccination clinic guide [15], with detailed instructions for clinic operations as well as the estimated number of persons necessary for operating such clinics, with an eye towards rapidly making such clinics operational in the event of a smallpox attack.

Following months of deliberation, on Friday December 13, 2002, President Bush announced his smallpox vaccination policy [16]. During phase 1, 500,000 members of the military were ordered to receive the vaccine; as Commander in Chief, the President was also vaccinated. In addition, 500,000 first responders – those most likely to come into contact with smallpox cases in the event of an attack and those responsible for vaccinating the general public in the event of an emergency – were slated for vaccination. The general public would only receive the vaccine in the event of an actual attack during this phase; indeed, the President stated specifically that neither his family nor his staff would be vaccinated. The second phase of this plan called for the vaccination of roughly 10 million doctors, nurses, police, fire, and other emergency personnel during 2003. Finally, starting in 2004, phase 3 of the policy allowed for the voluntary vaccination of the general public. As with the first phase of this plan, the entire public would be eligible for vaccine in the event of an attack.

Thus, from June through December of 2002, smallpox vaccination policy in the United States evolved from ACIP's recommendation to vaccinate only 15,000 first responders to a much broader plan, and from commitment to a limited ring vaccination response in the event of an attack to preparation for rapid mass vaccination of the affected population in an emergency. Mathematical models played a key role in the arguments for this change in policy. Below, we review some of the simple yet effective modeling arguments that were employed to good effect in the debate over smallpox vaccination policy.

20.3 NATURAL OUTBREAK OR BIOTERROR ATTACK? RING VERSUS MASS VACCINATION

As mentioned previously, ring vaccination was deployed in the World Health Organization's global smallpox eradication program. This policy worked well in areas of *moderate to high immunity,* for controlling *natural outbreaks,* and in populations that had *low mobility* (though even under these circumstances, the value added by ring vaccination has been questioned [17]). It is important to consider each of these points. First, before ring vaccination was deployed to "mop up" the last vestiges of smallpox worldwide, large-scale vaccination campaigns had already increased the degree of immunity in many countries to the point where between 25% and 75% of the population were already immune from smallpox [3]. Moreover, while smallpox killed roughly 30% of its victims, the surviving 70% were effectively immune to reinfection. Such immunity effectively reduces the infectiousness of the virus, since there are fewer persons who can be infected, making it much more difficult for the infection to spread – and much easier for ring vaccination to contain the infection.

Second, the outbreaks in question were all *natural.* One can think of a natural outbreak as infections as yet undetected by prior control efforts. Typically such "leftover" infections numbered in the single digits. By contrast, a bioterror *attack* would involve the deliberate infection of as many persons as possible. The potential for an aerosol attack makes an initial outbreak of hundreds or even thousands of infections feasible (e.g. by deployment in a congested area such as a train station, airport, sporting event, concert etc.), if not highly likely. Thus in preparing for response to a smallpox *bioterror attack,* it is important to decide the *outer level of risk* against which one wishes to defend. It makes little sense to focus only on policies that work for limited outbreaks of the form ring vaccination was designed to contain.

Third, smallpox was largely endemic in the rural areas of countries where sanitation and other health conditions were generally poor. Person-to-person contact in such immobile populations is thus somewhat limited. For example, in villages that are spaced 30 miles apart, the imposition of a 20-mile travel quarantine seems sensible – it could prevent disease from leaving one village and entering another. Incidentally, the November 2001 CDC emergency response plan stated that asymptomatic contacts of smallpox cases would also be restricted to traveling no further than 20 miles from their place of residence [5, Guide C, p. C-7]; one can only marvel at the presumed effectiveness of such a plan when applied to New York or any large American (or other) city!

In [9], the view taken was deliberately rather different than that promulgated by ACIP and CDC. Rather than consider only small outbreaks, a large attack that infects 1,000 persons in a city of 10 million persons was considered as a base case, though several parameters were varied over wide ranges in sensitivity analyses. For example, the number of secondary infections transmitted by each infected person early in the outbreak (the so-called reproductive number denoted by R_0) was assigned a value of 3 in this base case (at the lower end of the range estimated recently in [18] and thus favorable to ring vaccination), though this too was varied from 1 through 20 in sensitivity analyses.

Others have criticized such assumptions as overly pessimistic, preferring instead to focus on more likely scenarios. For example, Halloran *et al* state that, "...we explored what we consider to be the most likely method of attack, namely, a few infected individuals moving through the community" [19]. In evaluating the choice between ring versus mass vaccination, then, a question of modeling strategy emerges – should one focus on the "most likely" scenario, a "worst case" scenario, or some combination of these or other possibilities? This choice is important, for without conducting any analysis, it seems clear that ring vaccination is optimal for small outbreaks with low levels of infectiousness while mass vaccination is optimal for large outbreaks with an infectious agent.

Consider the following simple argument: It is not sensible to focus on the "most likely method of attack" and decide upon a response policy based solely on such a scenario, even if one believes that, should an attack occur at all, the odds are overwhelming that only a small attack would transpire. Rather, one should consider the *consequences* of different response policies when the attack scenario itself is uncertain. The idea is not to guess what will happen and then pretend that is what will transpire. The idea is to choose a robust response policy that will deliver good results *no matter what.*

To fix ideas, consider two possible attack scenarios, which we refer to as Big (e.g., aerosol attack with highly infectious virus) and Small (e.g., a single infected terrorist trying to spread smallpox by mingling in the population). Also, consider two possible responses – ring (or *traced*) vaccination (TV) and mass vaccination (MV). Using the models reported in [9], one can estimate the expected number of deaths that would occur for each of the four response strategy/attack scenario combinations. For a given strategy/scenario pair, let $d(\text{Stategy}|\text{Scenario})$ denote the expected number of deaths that would result (where Strategy $\in$ {TV,MV} and Scenario $\in$ {Big, Small}).

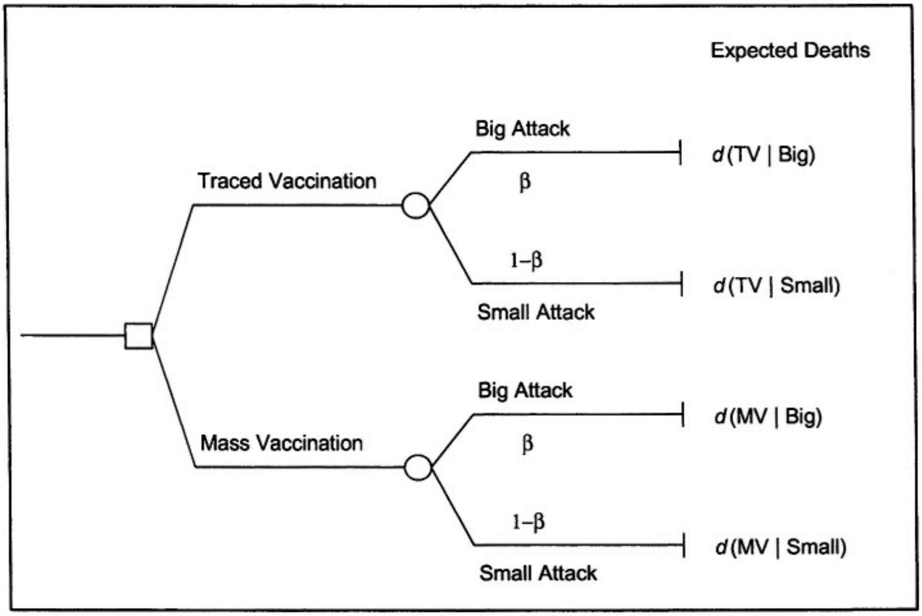
The decision problem is to choose a response strategy – TV or MV – once an attack has been detected. However, at the time one would face this decision, it would not be known whether the attack was Big or Small. Letting β denote the conditional probability that that attack is Big (given that an attack has occurred), the choice problem is represented by the decision tree shown in Figure 20.1.

Presuming that the objective is to minimize the expected number of deaths, the optimal decision is to choose MV providing

$$\beta\, d(\text{MV} \mid \text{Big}) + (1\text{-}\beta)\, d(\text{MV} \mid \text{Small})$$
$$<$$
$$\beta\, d(\text{TV} \mid \text{Big}) + (1 - \beta)\, d(\text{TV} \mid \text{Small})$$

which is equivalent to selecting MV if

Figure 20.1 Decision tree: choosing a response strategy



$$\beta > \frac{d(\mathrm{MV} \mid \mathrm{Small}) - d(\mathrm{TV} \mid \mathrm{Small})}{\left(d(\mathrm{MV} \mid \mathrm{Small}) - d(\mathrm{TV} \mid \mathrm{Small})\right) + \left(d(\mathrm{TV} \mid \mathrm{Big}) - d(\mathrm{MV} \mid \mathrm{Big})\right)}$$

As a numerical example, let a Big attack be specified by 1,000 initial infections with $R_0 = 3$, and a small attack correspond to a single initial infection with $R_0 = 1.2$. Then using the model in [9] (with all other parameters set at the base case values reported there), one obtains the following "payoff matrix" for the number of deaths that would result from considering all Strategy/Scenario pairs:

Expected Deaths

	Big Attack	Small Attack
Choose TV	110,000	2.3
Choose MV	560	10.4

From these figures, it is apparent that the death-minimizing choice is to respond with MV if the probability of a Big Attack exceeds

$$\frac{10.4 - 2.3}{(10.4 - 2.3) + (110{,}000 - 560)} = 7.4 \times 10^{-5}$$

This means that even if one is 99.99% certain that the attack is Small, it is still optimal (to a death minimizer) to respond with mass vaccination. Worded differently, even though one would almost surely employ TV with perfect information regarding the size of the attack, absent such information one should respond with MV.

Though simple, this example makes a very powerful point. It is a mistake to first envision the most likely mode of attack (as argued in [19]), and then design a response strategy that is optimal for such an attack scenario. As argued in [9], the consequences of choosing the wrong response policy are highly asymmetric (as is clear from the payoff matrix shown above). In the face of a small attack, choosing MV over TV leads to only a few expected incremental deaths (8.1 in the example above). In the event of a large attack, however, responding with TV would result in such a disaster that it is sensible to avoid this possibility altogether by responding with MV in the event of an attack.

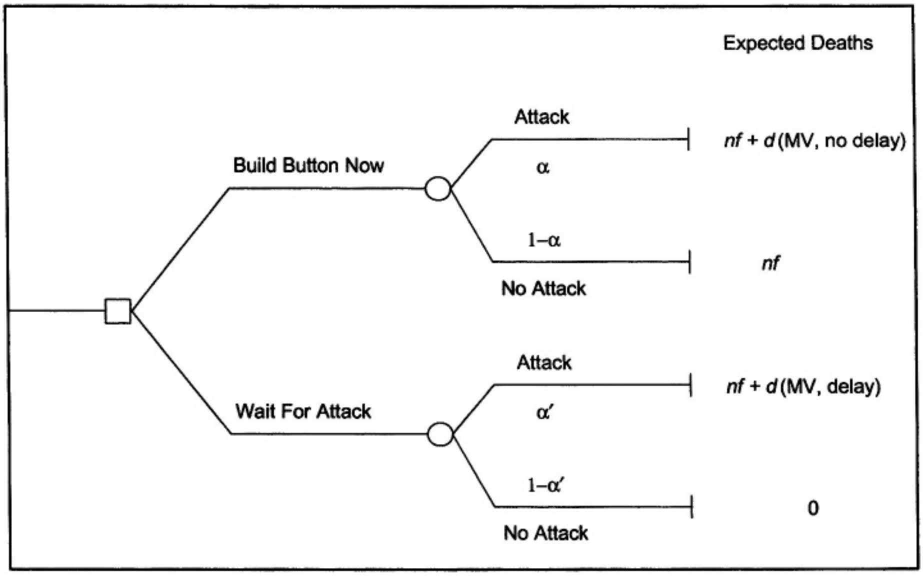
20.4 BUILD THE BUTTON NOW: VACCINATING THE VACCINATORS

The previous section suggested that mass vaccination is the expected death minimizing response in the event of a smallpox attack. However, in order to implement mass vaccination after an attack, those charged with vaccinating the public (as well as those responsible for receiving and treating initial smallpox cases) must themselves be immunized. Should one wait until an attack occurs before "vaccinating the vaccinators," or instead should one "build the button now?" That is, should one vaccinate a sufficient number of vaccinators now so that, in the event of an attack, localized mass vaccination of the population in the area of the attack could begin immediately?

In [9], it is argued that response delay is equivalent to increasing the size of an attack, since infections would continue to spread before actions to control the epidemic could be launched, increasing the total number of persons infected at the time response operations begin. However, as discussed earlier, vaccination against smallpox is not risk free. Thus, unlike the decision regarding which response strategy to employ upon detecting an attack, choosing how many vaccinators (and other first responders) to vaccinate now must depend upon the probability of an attack.

A simple version of this decision problem is illustrated in Figure 20.2. If one opts to "build the button now" and vaccinate n first responders, a fraction f of whom are expected to die from the vaccine itself, then one should expect nf deaths among first responders in the absence of an attack.

Figure 20.2 Decision tree: build button now or wait for attack



However, if an attack does occur (which it will with probability α), then in addition one would expect d(MV, no delay) deaths to occur in the population. If instead one opts to wait for an attack before "vaccinating the vaccinators," then there will be no deaths in the absence of an attack, but $nf + d$(MV, delay) if an attack does occur (which will occur with probability $\alpha' > \alpha$, assuming that an attack would be *less* likely given evidence of serious preparation).

Again choosing death minimization as the objective, it is sensible to vaccinate n vaccinators now if

$$\alpha > \frac{nf}{nf + \left(\frac{\alpha'}{\alpha} d(\text{MV, delay}) - d(\text{MV, no delay})\right)}$$

Since $\alpha' > \alpha$, a sufficient condition to build the button now is given by equating α' to α in the expression above, which means preparing now if

$$\alpha > \frac{nf}{nf + \left(d(\text{MV, delay}) - d(\text{MV, no delay})\right)}.$$

As an illustration, in [9] there were 5,000 full time vaccinators, which would translate to 15,000 vaccinators each working eight-hour shifts. Presuming a Big Attack, that mass vaccination would be employed in such an event, and that failing to prepare in advance would contribute an incremental one-day of response delay to the base case studied in [9], d(MV, no delay) = 560 deaths while d(MV, delay) = 653 deaths. Given that vaccination itself carries a 1 in 1 million risk of death, with these figures it makes sense to build the button now if the probability of an attack when such preparations are undertaken exceeds

$$\frac{15{,}000\times10^{-6}}{15{,}000\times10^{-6}+(653-560)}=1.6\times10^{-4}.$$

To ascertain whether this is a large or small threshold for the unknown probability of attack, it is important to note that the correct time scale to consider is the duration of time over which vaccination against smallpox remains effective. This is on the order of five to 10 years, so if you believe the likelihood of an attack over a 5 to 10 year period from now exceeds 1.6 in 10,000, it is prudent to prepare now.

Israel's decision-making can be contrasted with US policy in light of the results presented above. While US officials continue debating what to do about smallpox, Israel quickly (i.e. over a period of several weeks) decided to "build their button now." Approximately 15,000 Israeli first-responders have already been vaccinated, while a plan for locating clinics and rapidly vaccinating the Israeli public has already been determined [20]. Israel's policy of "vaccinating the vaccinators" stands in contrast to preparations in the United States [21].

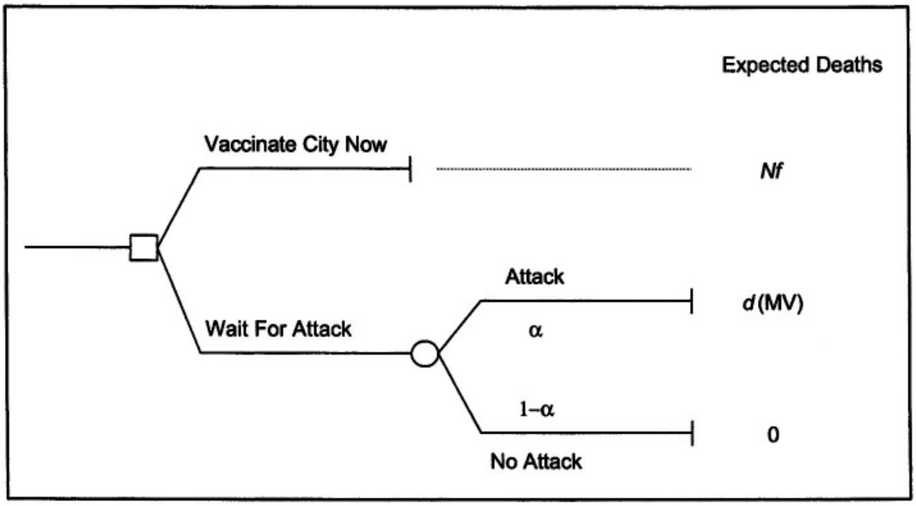
In the United States, the CDC has released a vaccination clinic guide [15]. Extrapolating from this guide suggests that to staff a sufficient number of vaccination clinics to vaccinate the entire US population within 10 days of an outbreak requires the vaccination now of approximately 1.25 million doctors, ambulance drivers, police officers etc. To reach the population within 5 days would double this requirement to roughly 2.5 million. Given these simple back-of-the-envelope calculations, it remains difficult to understand why the government is currently calling only for the vaccination of 510,000 hospital workers.

20.5 WHY WAIT? PRE-VERSUS POST-ATTACK VACCINATION

In view of the benefits that accrue from preparing in advance for a smallpox attack, many have argued that perhaps we should "go all the way" and vaccinate the entire population now. Whether it is sensible to do so again depends upon the vaccine fatality rate, the risk of attack, and the consequences of an attack should one occur. As discussed earlier, the consequences of an attack further depend upon the particular response policy employed, thus whether to vaccinate the entire population prior to an attack depends in a crucial way on the policy one would employ to control an epidemic post-attack.

This decision problem is shown in Figure 20.3. If the entire population of N persons were vaccinated, one would expect Nf deaths as a result. If instead one opts to only vaccinate post-attack, assuming that a mass vaccination response is employed, one would expect d(MV) deaths. An attack will occur with probability α if no vaccination occurs pre-attack. We ignore vaccine fatalities among first-responders as these would be negligible relative to vaccine fatalities in the general population, let alone deaths in the event of a smallpox attack. We also assume a perfect vaccine and 100% vaccination coverage; departures from both of these assumptions could easily be incorporated without changing the basic result.

Figure 20.3 Decision tree: pre- versus post-attack vaccination



To a death minimizer, it makes sense to vaccinate the entire population pre-attack if

$$Nf < \alpha\, d(\mathrm{MV})$$

which again yields a threshold on the attack probability – it is sensible to act now if

$$\alpha > \frac{Nf}{d(\mathrm{MV})}.$$

Consider again the base case of a Big Attack from [9], where N= 10 million, f = 1 per million, and d(MV) = 560. With these parameters, it would only make sense to vaccinate the entire population now if the probability of attack exceeds 10/560 or about 1.8%.

While some would consider this to be a small probability upon recalling that it represents the likelihood of an attack over a 5 to 10 year time horizon, note that it is two orders of magnitude higher than the attack threshold justifying the immediate vaccination of first-responders. Also, note that if traced vaccination were to be employed in response to an outbreak, then the threshold would drop to 1/11,000 due to the expected 110,000 deaths that would result were TV invoked post-attack.

This simple model suggests an explanation for much of the controversy in the smallpox debate. Whatever post-attack strategy is employed, if one is highly confident that the resulting outbreak can be quickly controlled, then there is little pressure to vaccinate the population now. Alternatively, if one harbors doubts regarding the ability to contain a smallpox attack, pre-attack vaccination becomes a more attractive option. Clearly, citizens differ in their assessments of government's ability to provide protection from smallpox post-attack. Citizens also differ in their views regarding the likelihood that such an attack could occur. Given such heterogeneity in views regarding risks and consequences, it is hardly surprising that many seek access to smallpox vaccine now, while others prefer to wait.

While the arguments above suggest that it is not sensible to vaccinate the entire country if the risk of an attack remains below about 1.8%, this model assumes a common valuation of risks and benefits. However, in a democracy such as the United States, people assume different risks and benefits all the time. While there is roughly a 1 per million chance of death from a smallpox vaccination, those of us who drive face a 145 per million risk of death each year from road accidents. Some people smoke, some drink and use illicit drugs, and some jump out of airplanes for sport. As Dr. Bicknell would argue, why shouldn't individuals be afforded a choice in this instance as well [6], especially since as previously noted, any amount of pre-attack vaccination makes any post-attack policy work better? This position presupposes that individuals are truly knowledgeable about the risks of being vaccinated or not (or can be made aware of these risks via communication of the facts), are capable of rational decision making, and thus the only differences among individuals are their preferences/beliefs regarding risks and benefits.

20.6 CONCLUSIONS

Smallpox was eliminated from the planet more than 20 years ago, but it could return as a weapon of bioterror. Barring an accident at one of the two known smallpox repositories (in Atlanta and Novosibrisk, Siberia), the appearance of smallpox anywhere in the world must be considered the result of a deliberate release. While the WHO's eradication campaign provided a wealth of data pertaining to the spread and control of natural outbreaks, there are simply no data at all that describe how smallpox would spread as the result of a bioterror attack. Some have assumed that deliberate outbreaks would mimic their natural counterparts, insinuating that methods believed successful in the past would surely work again in the future. By contrast, the arguments of this chapter do not suggest that the way to prepare for a disaster is to assume that it cannot occur (nor for that matter do they suggest that the way to evaluate a policy is to assume that it works, as ACIP did in recommending ring vaccination in response to a smallpox attack). Rather, simply allowing for the possibility that a large bioterror attack could occur greatly changes the optimal preparedness and response strategy, even if the probability assigned to such an attack remains very small. The consequences of decisions matter more than the details of the "most likely scenario" in considering what to do about smallpox and other bioterror agents. Simple models have served well to make this argument in the ongoing smallpox debate.

In spite of such arguments and the President's policy decision of December 2002, one December later, the United States is still unprepared for a smallpox bioterror attack. Nationwide, fewer than 40,000 first responders have volunteered to receive the smallpox vaccine as of December 2003, far short of the 10.5 million that were to have been immunized by the end of Phase II. As detailed elsewhere [22], there are numerous administrative, political, and public health reasons for this state of affairs. Whether or not the pre-event smallpox vaccination program will recover remains to be seen.

Acknowledgments

E.H.K. was supported in part by the Center for Interdisciplinary Research on AIDS via Grant MH/DA56826 from the US National Institutes of Mental Health and Drug Abuse.

References

[1] Miller, J., S. Engelberg, and W. Broad (2001). *Germs.* Simon and Schuster, New York.

[2] Alibek, K. (1999). *Biohazard.* Random House, New York.

[3] Fenner, F., D.A. Henderson, I. Arita, Z. Jezek, and I.D. Ladnyi. (1988). *Smallpox and its Eradication.* World Health Organization, Geneva.

[4] Tucker, J.B. (2001). *Scourge: The Once and Future Threat of Smallpox.* Atlantic Monthly Press, New York.

[5] Centers for Disease Control and Prevention (2001). *CDC Interim Smallpox Response Plan and Guidelines, Draft 2.0.* Centers for Disease Control and Prevention, Atlanta, GA.

[6] Bicknell, W.J. (2002). The case for voluntary smallpox vaccination. *New England Journal of Medicine,* 346, 1323-1325.

[7] Fauci, A.S. (2002). Smallpox vaccination policy – the need for dialogue. *New England Journal of Medicine,* 346, 1319-1320.

[8] Neff, J.M., J.M. Lane, V.A. Fulginiti, and D.A. Henderson (2002). Contact vaccinia – transmission of vaccinia from smallpox vaccination. *Journal of the American Medical Association,* 288, 1901-1905.

[9] Kaplan, E.H., D.L. Craft, and L.M. Wein (2002). Emergency response to a smallpox attack: the case for mass vaccination. *Proceedings of the National Academy of Sciences of the USA,* 99, 10395-10440.

[10] Institute of Medicine (2002). *Scientific and Policy Considerations in Developing Smallpox Vaccination Options.* National Academies Press, Washington, DC.

[11] Zelicoff, A.P. (2002). An epidemiological analysis of the 1971 smallpox outbreak in Aralsk, Kazakhstan. In J.B. Tucker and R.A. Zilinskas (Eds.) *The 1971 smallpox epidemic in Aralsk, Kazakhstan, and the Soviet Biological Warfare Program.* Occasional Paper No. 9, Monterey Institute of International Studies, Center for Nonproliferation Studies, Washington, DC, 12-21.

[12] Altaian, L.K. (2002). Smallpox proposal raises ethical issues. *New York Times,* June 22, A9.

[13] Broad, W.J. (2002). U.S. to vaccinate 500,000 workers against smallpox. *New York Times,* July 7, 1.

[14] Siegel-Itzkovitch, J. (2002). Should only infected people be vaccinated? *Jerusalem Post,* July 28, 9.

[15] Centers for Disease Control and Prevention (2002). *Smallpox vaccination clinic guide.* Centers for Disease Control and Prevention, Atlanta, GA.

[16] Stevenson, R.W. and S.G. Stolberg (2002). Bush lays out plan on smallpox shots; military is first. *New York Times,* December 14, A1.

[17] Kaplan, E.H. and L.M. Wein (2002). Smallpox eradication in west and central Africa: surveillance-containment or herd immunity? *Epidemiology,* 14, 1-4.

[18] Gani, R. and S. Leach (2001). Transmission potential of smallpox in contemporary populations. *Nature,* 414, 748-751.

[19] Halloran, M.E., I.M. Longini Jr., A. Nizam, and Y. Yang (2002). Containing bioterrorist smallpox. *Science,* 298, 1428-1432.

[20] Sa'ar, R. (2002). Ministry presents smallpox crisis plan. *Ha'aretz,* October 18, 1.

[21] National Public Radio (2002). Profile: smallpox preparedness of Israel and the U.S. *All Things Considered,* November 1, http://www.npr.org/programs/atc/transcripts/2002/nov/021101 .northam.html, Accessed December 7, 2002.

[22] Bicknell, W.J. and K.D. Bloem. *Smallpox and bioterrorism: why the plan to protect the nation is stalled and what to do.* Cato Institute: Briefing Paper #85, September 5, 2003.

21 MODELS FOR KIDNEY ALLOCATION

Stefanos A. Zenios

Graduate School of Business
Stanford University
Stanford, CA 94306

SUMMARY

The continued shortage of organs implies that the organ allocation policy determines who lives and who dies. This creates one of medicine's most vexing dilemmas, and the crux of this dilemma is the tradeoff between clinical efficiency and equity. This chapter describes OR models that have been used to study the problem and the related tradeoffs. A taxonomy of the literature is developed, and a description of the key analytical and computational models is provided. Directions for future research are also presented.

KEY WORDS

Kidney allocation, Equity, Efficiency, Queueing models, Fluid models, Simulation, Evidence-based medicine

21.1 INTRODUCTION

Kidney transplantation is the preferred treatment for patients suffering from chronic renal insufficiency (CRI), also known as chronic kidney failure. However, the supply of cadaveric kidneys for transplantation has failed to meet the ever-increasing demand. In 2000, 22,271 new candidates joined the kidney transplant waiting list but only 9,278 transplantations were performed. At the end of the same year, 47,873 patients were on the waiting list, and its size has been growing steadily [1]. The continued shortage of organs creates several challenges related to their allocation and distribution that can be addressed using carefully crafted OR models. This chapter provides the relevant institutional background about the kidney allocation problem and then presents a review of the OR literature on kidney allocation models. It concludes with a discussion of fruitful avenues for future research.

21.2 THE KIDNEY ALLOCATION SYSTEM

The origins of the kidney allocation system employed in the US are traced to the National Organ Transplant Act (NOTA) enacted by the US congress in 1984. The Act established a national organ sharing system, the United Network of Organ Sharing (UNOS), whose purpose is to maintain a national transplant waiting list and to coordinate the activities of the local agencies that procure the organs for transplantation. In broad terms, the system is organized as follows: There are 69 regional Organ Procurement Organizations (OPO) that procure donated organs in their region and coordinate their transplantation to patients living in the same region; such patients are called *local* patients. These OPOs are organized into 11 broader geographic regions. Organs are procured by the OPO who then uses an allocation algorithm approved by UNOS to prioritize local transplant candidates. If an appropriate candidate is not found within the OPO, then the search is broadened to include the whole region in which the OPO belongs, and if it fails in that stage, the search becomes national.

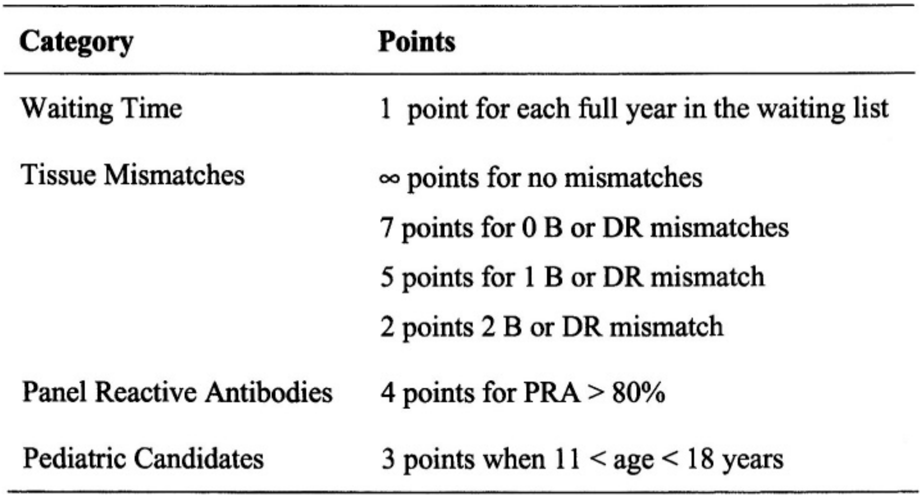
The basic allocation algorithm has evolved over the years to reflect the expanding nature of medical knowledge but its key ingredients have remained constant. First, organs must be transplanted to patients who are blood-compatible but not necessarily rhesus-type compatible. Second, organs should not be transplanted to presensitized patients; these are patients who may exhibit an immune response to the proteins in the donor's organ. An immunological test referred to as the presensitization test must be performed in order to determine the risk of such a reaction. Only patients who test negative (also known as *negative-crossmatch or non-presensitized patients*) can receive a transplant [2-4]. Third, patients must be in reasonably good physical condition (good relative to the CRI standards) and must be psychologically ready for the operation and likely to comply with the post-operation instructions of the surgeon. Fourth, it is desirable for the so-called tissue type of the donor and tissue type of the patient to match. The tissue type, also known as HLA type, is a combination of six proteins: Two of type A, two of type B, and two of type DR. Empirical and clinical evidence show that when the donor and recipient share all six proteins in common (zero mismatches), the risk of graft rejection is minimized (graft is the medical term for the transplanted kidney). The risk increases with the number of mismatches, and so allocating an organ to the recipient with the smallest number of tissue type *mismatches* reduces the chance of graft failure; the number of mismatches is the number of donor HLA proteins that are absent in the recipient.

The exact algorithm used is based on the UNOS point system but individual OPOs may request an exemption and implement their own variation. In the point system, each transplant candidate receives a number of priority points based on the total number of tissue matches. To compensate candidates with rare tissue types, the policy also awards points based on the candidate's waiting time and rank on the waiting list; consequently, candidates do not stay on the waiting list indefinitely. Finally, the system allocates priority points to candidates with high *panel reactive antibodies* (pra); the pra of a candidate is an estimate for the probability that the candidate will crossmatch positive with a randomly selected donor. The pra points ensure that a golden (but rare) opportunity of a negative crossmatch will not be missed by those candidates. The details are described in Table 21.1. Once candidates are prioritized, organs are allocated according to the following sequence: First, the organ is offered to an identical blood-type zero antigen mismatched local patient, then regionally and then nationally. Then it is offered to a blood-type compatible zero-antigen mismatched patient using the same geographic hierarchy. Finally, the organ is offered to all other blood-type compatible candidates ranked according to their total number of points. The algorithm also provides for exchanges between OPOs that are forced to share zero antigen mismatched kidneys and it also specifies in unambiguous terms the criteria used to register patients on the transplant waiting list in order to avoid abuses in waiting time points.

A recent modification in this system has been the introduction of a waiting list for so-called "expanded-criteria donors." These are older donors who typically provide organs of marginal quality. Even though these organs may not be appropriate for all transplant candidates, some candidates, together with their physicians, may find them appealing. Hence, patients can now declare their willingness to accept these marginal kidneys. They then join the corresponding waiting list where the allocation is done according to First-

Come First-Transplanted. These patients remain eligible for the so-called "standard" kidneys and their enrollment in the waiting list for "marginal" kidneys does not affect their standard priority in any way.

Table 21.1 The current UNOS point system*



Category	Points
Waiting Time	1 point for each full year in the waiting list
Tissue Mismatches	∞ points for no mismatches 7 points for 0 B or DR mismatches 5 points for 1 B or DR mismatch 2 points 2 B or DR mismatch
Panel Reactive Antibodies	4 points for PRA > 80%
Pediatric Candidates	3 points when 11 < age < 18 years

* UNOS [1] provides a detailed description

While the development of the current system has evolved over several years, important aspects of it remain controversial. In particular, despite repeated efforts to achieve equity in access to transplantation, several demographic groups and patients in a few geographic regions wait much longer than the national average. The problem is particularly acute for African Americans who are 46% less likely to receive a transplant than Caucasians. Data on transplantations performed between 1988 and 1992 show that 30.3% of all African American transplant candidates receive a transplant within 5 years from the onset of CRI while the corresponding percentage for Caucasians is 56.7% [5]. While several hypotheses have been proposed for this observation, the most appealing one is that "good" tissue matches are less likely to occur for African Americans because they have a more diverse genetic makeup and because there is an imbalance in the supply of organs from African American donors. Specifically, while African Americans make up only 10% of all donors, they constitute 30% of the CRI population. Because the tissue type of a donor is unlikely to match the type of a recipient of a different race, African Americans are much less likely than Caucasians to receive a large number of tissue match points.

In this environment of acute organ shortage, a systematic modeling-based analysis of the allocation system can better clarify the main tradeoffs and can enable more informed decision-making. OR models have already contributed to the debate. The remainder of this chapter will describe the key contributions and will outline future research opportunities.

21.3 TAXONOMY OF THE LITERATURE

OR models can be used to illuminate the decision-making behavior of the different parties. In the context of the allocation problem, the relevant parties are the central planner (UNOS) who determines the allocation policy, and the individual patients who may wish to determine whether to accept an organ. To study the choices of these parties one may employ analytical models that can be used to develop "optimal" allocation policies, or analytical models that derive closed-form expressions for performance evaluation, or simulation-based models that provide performance evaluation. In each case, the critical issue that needs to be resolved is the determination of the relevant performance measures that will be used either for optimization or for performance-evaluation. Because of the underlying complexity of the problem, a credible analytical model must be validated using a computer simulation model. Table 21.2 categorizes the existing literature; this includes literature that is directly motivated by the kidney allocation problem as well as more general literature that is deemed relevant.

The table is not meant to be exhaustive. Rather, it highlights the majority of the relevant papers in each area. References in these papers provide a more comprehensive overview of the literature. Gaps in the table indicate lack of relevant papers.

In the remainder of this chapter, we will describe the main models in each category, starting with models for patient decision making.

21.4 MODELS FOR PATIENT DECISION MAKING

In the current environment of acute organ shortage, one would expect that very few patients would have any reason to refuse a kidney offered by UNOS. However, approximately 45% of all organs are refused by the first-offered candidate [1]. These refusals are typically made by the transplant surgeons and they reflect their own experience and beliefs about the type of organs best suited to their patients [1].

David and Yechiali [6] present a stylized model that captures such decision making. In their model, the patient (or the surgeon) receives offers in discrete $0 = t_0 < t_1 < t_2 < .....$ epochs. In epoch t_j an organ offer X_j is

Table 21.2 Taxonomy of modeling papers in kidney allocation

Model	Decisions	Optimization	Performance Evaluation
Analytical	Patient	David and Yechiali [6]	
	UNOS	Derman, Lieberman, and Ross [9] Righter [11] David and Yechiali [12] Zenios, Chertow, and Wein [13]	Zenios [19]
Computational	Patient	Ahn and Hornberger [7]	
	UNOS		Opelz and Wujciak [17] Pritsker [16] Zenios, Chertow, and Wein [16] Zenios, Wein, and Chertow [13] Howard [18] Votruba [15]

received. The sequence of organ offers consists of independent identically distributed random variables with probability distribution function $F(x) = P(X \leq x)$. The decision at time t_j is whether to accept or reject the offer. If the offer is accepted, a reward $\beta(t_j)X_j$ is gained. The function $\beta(t)$ is a non-increasing discount function. Further, the process may terminate by itself because of patient death.

The authors solve the optimal stopping time problem under a variety of assumptions and they obtain insightful expressions. To illustrate, consider the case where the discount function is $\beta(t) = e^{-\beta t}$, the death rate is r (that is, patient lifetime is exponential with rate r), and offers arrive according to a time-homogeneous Poisson process with intensity λ. Then, the optimal stopping time is to accept offers with reward that exceeds a threshold γ obtained by the unique solution to

$$\frac{\gamma}{P(X \geq \gamma)} = \frac{\lambda}{\beta + r}. \tag{1}$$

This expression implies that the decision threshold γ increases as the offer arrival rate increases, but it decreases as the discount rate or as the death rate increases. Hence, patients are more selective when they expect frequent offers, but they becomes less selective when they are "impatient" or when their mortality rate increases.

Ahn and Hornberger [7] extend this basic model to improve its clinical relevance. Their analysis assumes that the patient can be in one of five states in each period: alive on dialysis awaiting a transplant; not eligible for transplant; received a functioning living transplant; transplant failed; and death. They assume that monthly transitions between states follow a Markov Chain and patients in the "alive on dialysis awaiting a transplant" state specify a threshold for the minimum acceptable kidney offer. Using historical data, the authors develop and validate a clinically relevant model and then examine how the decision threshold can vary with the patient's preferences. An important novelty of their model is that it considers explicitly the utility of the patient in each state, commonly referred to as the patient's quality-of-life score; see Gold et al. [8]. Their results show that a patient's decision threshold changes with the patient's perception about quality of life in different health states. The authors propose an allocation system with increased patient involvement based on the rationale that such a system will reduce inefficiencies and improve the utilization of marginal organs.

21.5 MODELS FOR ORGAN ALLOCATION

One of the most fundamental models for organ allocation is the sequential stochastic assignment model developed in Derman, Lieberman, and Ross [9]. This is a discrete-time model in which *n* transplant candidates receive n organs that arrive sequentially, one in each period; the original paper used the language "jobs" and "workers" and did not provide the organ allocation motivation since the transplant problem had not been identified in 1972. Associated with each candidate is a value $r_1 \geq r_2 \geq \ldots \geq r_n$, Also, each arriving organ has a value, X, which is a non-negative random variable with known distribution. When the organ arrives, its value is observed and then a decision is made whether to reject it or to assign it to some candidate. If an organ of value x is assigned to a candidate with value r_i, then the reward is $r_i x$. The objective is to determine the allocation rule that maximizes total expected reward. The authors provide a complete characterization of the exact optimal policy.

In a followup paper, Albright and Derman [10] provide a closed-form solution for the policy that is asymptotically optimal as $n \to \infty$. Their main result identifies a set of thresholds $c_1 \geq c_2 \geq \ldots \geq c_n$ such that an organ is offered to the candidate with value r_1 if the organ's value exceeds c_1; otherwise it is offered to the candidate with value r_2 if the organ's value is between c_1 and c_2, and so on. Thus organs of higher quality are assigned to candidates most likely to enjoy the benefits from such elevated quality. The critical thresholds are obtained from the solution to the following critical fractile equations

$$P(X \geq c_i) = \frac{i}{n}; \text{for } i = 1, \ldots, n. \tag{2}$$

This model provides a very stylized description of the basic organ allocation model as the model was not developed with the organ allocation problem in mind.

Righter [11] makes an important step towards a more relevant generalization. In her model, each candidate has a "random deadline" that reflects death, and the candidate values may change according to an underlying Markov chain that captures the dynamics of CRI. It is shown that under this expanded set of dynamics, a threshold policy remains optimal but the thresholds depend on the state of the underlying Markov chain. Structural properties for the thresholds are obtained and conditions are provided in which the thresholds are monotone in the states of the Markov chain. These results suggest that threshold policies are robust and are an attractive candidate for a practical implementation. However, the utilitarian framework utilized in the analysis makes such threshold policies potentially unappealing because they would violate the requirement that organs are allocated both efficiently and equitably.

Another important step was made by David and Yechiali [12] whose main innovation was to consider a reward structure that reflects the tissue matching reality of organ transplants. In their model, the reward is R when the candidate and organ match, and $r(R > r)$ if they do not match. Each candidate has a known attribute and each organ also has an attribute that becomes known when the organ becomes available. A match exists when the attribute of the organ and the attribute of the candidate are identical; otherwise it is a mismatch. Their analysis focuses on the so-called "intuitive" policy in which a match is always assigned, whereas a mismatch is assigned only if the number of candidates exceeds the number of offers and it is then assigned to the candidate with the rarest attributes. The authors pursue an extensive investigation of different special cases, identify cases in which the "intuitive" policy is optimal, and in all other cases they refine the policy to achieve optimality.

The sequential stochastic assignment model and all its extensions provide an elegant and stylized description of the basic problem that is analytically tractable. However, such models provide a very limited and "artificial" representation of the transplant reality and only focus on a utilitarian perspective of maximizing clinical efficiency. To overcome these limitations, Zenios, Chertow, and Wein [13] develop a fluid-based model that attempts to provide a more clinically relevant description of the problem.

Their model is a continuous time, continuous space deterministic model in which the CRI population is divided into K distinct categories, or classes, based on demographic (age, gender, race), immunological (blood type, tissue type, pra) and physiological characteristics (height, weight). The donor population is also divided into J classes, based again on demographic, immunological, and physiological characteristics. Without loss of generality, patients of class $k = 1,\ldots,K_W$ are registered on the waiting list and patients of class $k = K_W + 1,\ldots,K$ have a functioning graft.

The state of the system at time t is described by the K-dimensional column vector $x(t) = (x_1(t),\ldots,x_K(t))'$ (where primes denote transposes), which gives the number of patients in each class. Transplant candidates of class $k = 1,\ldots,K_W$ join the waiting list at rate λ_k^+ per unit time. These patients depart from the waiting list via death, which occurs at rate μ_k per unit time for class k patients, or organ transplantation. Organs of class $j = 1,\ldots,J$ arrive at rate λ_j^- per unit time; the demand-to-supply ratio $\rho = \sum_{s=1}^{K_W} \lambda_s^+ / \sum_{j=1}^{J} \lambda_j^-$ greater than one. A fraction $u_{jk}(t)$ of class j organs is allocated to transplant candidates of class k; thus, $u_{jk}(t)$ is a control variable and $u_{jk}(t) = \lambda_j^- u_{jk}(t)$ is the instantaneous transplantation rate of class j kidneys into class k candidates.

When a class $j = 1,...,J$ kidney is transplanted into a class $k=1,\dots,K_W$ candidate, the class k candidate leaves the waiting list and becomes a patient of class $c(k,j)\in\{K_W+1,\dots,K\}$. Patients of class $c(k,j)$ depart this class at rate $\mu_{c(k,j)}$ per unit time; a fraction $q_{c(k,j)}\in[0,1)$ of these patients experience graft failure and rejoin the waiting list as patients of class k, and the remaining fraction exit the system due to death.

The dynamics of the system are described by the ordinary linear differential equations

$$\frac{d}{dt}x_k(t)=\lambda_k^+-\mu_k x_k(t)-\sum_{j=1}^{J}u_{jk}(t)+\sum_{j=1}^{J}q_{c(k,j)}\mu_{c(k,j)}x_{c(k,j)}(t);k=1,\dots,K_W \tag{3}$$

and

$$\frac{d}{dt}x_k(t)=\sum_{j=1}^{J}\sum_{i=1}^{K_W}u_{ji}(t)1_{\{c(i,j)=k\}}-\mu_k x_k(t);k=K_W+1,\dots,K. \tag{4}$$

The objective function for this system contains two criteria. The first criterion is clinical efficiency measured using quality-adjusted life years (QALY); the reader is referred to Gold et al. [8] for a detailed introduction to QALY and their applications to societal decision making. In this measure, patients in class k have a *quality of life* (QOL) score h_k and the total QALY over a finite time horizon T is the total number of life years multiplied by the QOL scores,

$$\int_0^T\sum_{k=1}^{K}h_k x_k(t)dt; \tag{5}$$

discounting can also be incorporated into the objective if desired.

The second criterion is equity which is measured using the variance in waiting time across different patient classes. The rationale is that in an equitable allocation policy there will be little differences in average waiting time across different patient demographic groups. Then it can be argued that the following measure, referred to as *waiting time inequity,* is a tractable proxy for the desired variance measure:

$$\frac{1}{2}\int_0^T \sum_{k=1}^{K_W}\sum_{i=1}^{K_W} \lambda_k(t,u(t))\lambda_i(t,u(t))\left(\frac{x_k(t)}{\lambda_k(t,u(t))}-\frac{x_i(t)}{\lambda_i(t,u(t))}\right)^2 dt, \qquad (6)$$

where $\lambda(t,u(t)) = \lambda_1(t,u(t)),\ldots,\lambda_{K_W}(t,u(t))\;)$ denotes the instantaneous arrival rate into class k under allocation policy $u(t)$, and is given by

$$\lambda_k(t,u(t)) = \lambda_k^+ + \sum_{j=1}^{J} q_{c(k,j)}\mu_{c(k,j)}x_{c(k,j)}(t) \quad \text{for } k = 1,\ldots,K_W \qquad (7)$$

Because the waiting time may be confounded by each patient's underlying mortality, it is also suggested that constraints are imposed on differences in likelihood to transplantation.

The authors use optimal control theory to obtain a dynamic heuristic policy that maximizes a weighted combination of (5) and (6) subject to (3)-(4). The policy is a dynamic index: At time t an index $G_{jk}(t)$ is computed for each possible combination of patient class k and organ class j. Organs of class j are assigned to the candidate with the highest index. The index has three components: an efficiency component, an equity component, and a subsidy. The efficiency component gives the expected increase in candidate *k's* QALY if she receives an organ j. The equity component is based on actual waiting times and attempts to bring the waiting times for all classes close to their waiting times under the equity standard of First-Come First-Transplanted (FCFT). The subsidy attempts to elevate the likelihood of transplantation of those patient classes whose high mortality rate prevents them from accumulating a sufficiently strong equity component. The policy is appealing because it can be formulated as a point system similar to the one utilized by UNOS. On the other hand, it considers a much broader spectrum of variables and criteria than the policy utilized by UNOS, and hence its practical implementation is challenging.

In the same paper, the authors develop a simulation model for the organ allocation system and compare the performance of their policy to competing policies such as the UNOS policy (referred to below as simply "UNOS") and First-Come First-Transplanted. The analysis suggests that the mathematically derived policies dominate UNOS in all dimensions considered, including clinical efficiency and various alternative measures of equity. Further, the QALY difference between the optimal efficiency-based policy and UNOS is comparable to the QALY gains attributed to immunosuppressive drugs; these drugs, which were introduced in the 1980's, revolutionized the field of transplantation by substantially increasing the transplant success rate. More surprisingly, FCFT has almost the same QALY outcomes as UNOS but is much more equitable. Thus, UNOS is dominated both in the efficiency and in the equity dimension. A more in-depth simulation-based comparison of the different policies is provided in Zenios, Wein, and Chertow [14].

Votruba [15] adopts a similar approach to study the problem of organ allocation. Using a simple static model, the author arrives at the efficiency-based policy developed in Zenios, Chertow, and Wein [13]. The author tests the policy using a simulation model and compares its performance to those of competing policies such as UNOS, FCFT, and SIRO (serve in random order). Using more recent data than Zenios, Chertow, and Wein [13] the author reaches the same conclusion that UNOS is dominated in the equity dimension by both FCFT and SIRO without a commensurate difference in QALY, and is dominated in the efficiency dimension by the efficiency-based policy.

UNOS has compared different allocation policies using computer simulation models. The UNOS Kidney Allocation Model was developed by the Pritsker corporation for UNOS with input from the scientific community; see Pritsker [16]. The model was developed and validated over a period of two years and it is now used routinely to compare and test different policies before then-implementation. In the view of the author, this model represents one of the most impressive efforts to utilize OR models in health care delivery. The early work by Opelz and Wujciak [17] is also worth mentioning as it has provided a proof-of-concept for the relevance of simulation models in organ allocation. More recently, Howard [18] used a simulation model to compare different policies for liver allocation.

21.6 ANALYTICAL PERFORMANCE EVALUATION MODELS

While most of the existing models in organ allocation focus on the development and simulation-based evaluation of alternative organ allocation policies, there is also a need for analytical performance evaluation models. Such models provide closed-form expressions for the key performance measures. There are two important categories of performance measures: clinical efficiency measures such as QALYs, probability of graft survival one and five years after transplantation, and average graft survival; and access measures such as waiting time, waiting time until transplantation, and likelihood of transplantation. Much has been said in the literature about the effect of the organ allocation systems on these measures. It is well recognized that not all patients have equal access to transplantation: in particular, African Americans have longer waiting times. Simple analytical models can shed light to these empirical observations.

Zenios [19] develops a simple queueing model for the transplant waiting list. The model assumes there are K classes of transplant candidates who arrive according to independent Poisson processes with rate $\lambda_k^+, k=1,\ldots,K,$ and J classes of organs that arrive according to independent Poisson processes with rate $\hat{\lambda}_j^-, j=1,\ldots,J.$ The organs are allocated to transplant candidates and force their departure from the waiting list. In addition, patients of each class $k = 1,\ldots,K$ renege from the system due to death after an exponentially distributed amount of time with rate μ_k. The allocation policy takes the form of a static randomized policy. In particular, $\nu = (\nu_{jk})$ is the fraction of class j organs that are allocated to patients of class k. Candidates of the same class are allocated organs on a first-come first-transplanted (FCFT) basis.

The paper develops asymptotic (as the patient and organ arrival rates become arbitrarily large) closed-form approximations for the steady-state waiting time W_k for each patient class k, for the steady-state waiting time given transplantation W_k^T, and for the steady-state likelihood of transplantation L_k. These expressions are as follows (where $\lambda_k^- = \sum_{j=1}^K \hat{\lambda}_j^- \nu_{jk}; k=1,\ldots,K)$:

$$W_K = \frac{1}{\mu_k}(1-\frac{\lambda_k^-}{\lambda_k^+}), \tag{8}$$

$$W_k^T = \frac{1}{\mu_k}\ln\left(\frac{\lambda_k^+}{\lambda_k^-}\right), \tag{9}$$

$$L_k = \frac{\lambda_k^-}{\lambda_k^+}. \tag{10}$$

These expressions reveal that the waiting time for each patient class is determined by that class's supply-to-demand ratio $\frac{\lambda_k^-}{\lambda_k^+}$ and by its mortality rate μ_k : A long waiting time may indicate reduced access to transplantation or may suggest a lower-than-average mortality rate. In the context of historical data on transplantation, this suggests that the long waiting times experienced by African American transplant candidates are caused both because of reduced access to transplantation as measured by the supply-to-demand ratio and because of historically lower-than-average mortality on the waiting list [20].

21.7 DIRECTIONS FOR FUTURE RESEARCH

Kidney transplantation is one of modern medicine's major success stories, but has also created one of its major dilemmas: how to allocate donated organs efficiently and equitably. The tremendous success of dialysis as a maintenance option for CRI and of transplantation as the preferred treatment option have created an environment of continued organ shortage. In this environment, the statement "there is no free lunch" takes a new meaning: every change in the allocation policy is likely to create winners and losers. Rigorous models such as the ones developed by operations researchers can prove invaluable in clarifying the tradeoffs. Carefully crafted analytical models can be used to identify policies that balance the key tradeoffs between maximizing clinical efficiency and promoting equity, while simulation models can be utilized to test the performance of these policies and to compare them to existing strategies.

While much work has been done in model development, several aspects of the kidney allocation problem remain understudied. First, a mathematical definition of equity or fairness remains elusive. A systematic study of the issues and the development of an axiomatic decision theoretic framework for equity is a task worth undertaking. Second, despite the continued shortage of organs, more than 10% of all donated organs are eventually discarded. Ways must be found to utilize these organs more efficiently. The recent proposal for a "marginal donors" waiting list appears to be the first step in the right direction. OR models can be used to identify the main causes for this wastage and to identify remedies. Su and Zenios [21,22] and Howard [23] represent two recent attempts to study this problem.

This chapter summarizes OR models applied to the kidney allocation problem. These models can also be extended to other organ allocation settings with minor modifications. Specifically, beyond tissue matching that is an idiosyncratic feature of kidney transplantation, all other aspects of the kidney allocation problem captured in these models are universally relevant. In particular, the efficiency-equity tradeoff prevails in all transplant waiting systems for all organs in most countries.

The key problem in organ transplantation is the continued shortage of donated organs. Two proposals that are likely to solve the problem have attracted attention recently. The first proposal is to create a market for organs in which UNOS (or some other authority) is the single buyer of organs and the sellers will be the estate of the deceased [24]. The second proposal is to create a mutual insurance pool. Participants in the pool commit to donate their organs upon death and in exchange they obtain access to the organs donated by other participants in the pool. While both proposals are controversial, they are appealing because they are likely to alleviate the shortage, and they are now debated actively [25]. However, implementation of these proposals will be associated with significant operational and logistic hurdles that can be addressed using OR models. Therefore, one hopes that the future will see more OR research done in this area.

References

[1] United Network for Organ Sharing (2002). *UNOS Statistical Reports.* http://www.unos.org/frame_Default.asp?Category=Data.

[2] Allen, R.D.M. and J.R. Chapman (1994). *A Manual of Renal Transplantation.* Edward Arnold, London.

[3] Ghjertson, D.W., P.I. Terasaki, B.S. Takemoto, and M.R. Mickey (1991). National allocation of cadaveric kidneys by HLA matching. *New England Journal of Medicine,* 324, 1032-1036.

[4] Chertow, G., S.L. Milford, H.S. MacKenzie, and B.M. Brenner (1996). Antigen-independent determinants of cadaveric kidney transplant failure. *Journal of the American Medical Association,* 276, 1732-1736.

[5] Eggers, P.W. (1995). Racial differences in access to kidney transplantation. *Health Care Financing Review,* 17, 89-103.

[6] David, I. and U. Yechiali (1985). A time-dependent stopping problem with application to live organ transplant. *Operations Research,* 33, 491-504.

[7] Ahn, J.H. and J.C. Hornberger (1996). Involving patients in the cadaveric kidney transplant allocation process: A decision-theoretic perspective. *Management Science,* 42, 629-641.

[8] Gold, M.R., J.E. Siegel, L.B. Russell, and M.C. Weinstein, Eds. (1996). *Cost-Effectiveness in Health and Medicine.* Oxford University Press, New York.

[9] Derman, C., G.J. Lieberman, and S.M. Ross (1972). A sequential stochastic assignment problem. *Management Science,* 18, 349-355.

[10] Albright, S.C. and C. Derman (1972). Asymptotic optimal policies for the stochastic sequential assignment problem. *Management Science,* 19, 46-51.

[11] Righter, R.L. (1988). A resource allocation problem in a random environment. *Operations Research,* 37, 329-338.

[12] David, I. and U. Yechiali (1995). One-attribute sequential assignment match processes in discrete time. *Operations Research,* 43, 879-884.

[13] Zenios, S.A., G.M. Chertow, and L.M. Wein (2000). Dynamic allocation of kidneys to candidates on the transplant waiting list. *Operations Research,* 48, 549-569.

[14] Zenios, S.A., L.M. Wein, and G.M. Chertow (1999). Evidence-based organ allocation. *American Journal of Medicine,* 107, 52-61.

[15] Votruba, M. (2002). Efficiency-Equity Tradeoffs in the Allocation of Cadaveric Kidneys. Working Paper, Princeton University, Princeton, NJ.

[16] Pritsker, A.B. (1998). Life and death decisions: Organ transplantation allocation policy analysis. *OR/MS Today,* 25, 22-28.

[17] Opelz, G. and T. Wujciak (1995). Cadaveric kidneys should be allocated according to the HLA match. *Transplantation Proceedings,* 27, 93-99.

[18] Howard, D.H. (2001). Dynamic analysis of liver allocation policies. *Medical Decision Making,* 21, 257-266.

[19] Zenios S.A. (1999). Modeling the transplant waiting list: A queueing model with reneging. *Queueing Systems: Theory and Applications,* 31, 239-251.

[20] U.S. Renal Data System (2002). *Annual Data Report: Atlas of End-Stage Renal Disease in the United States.* National Institutes of Health, National Institute of Diabetes and Digestive and Kidney Diseases, Bethesda, MD.

[21] Su, X. and S.A. Zenios (2002). Patient Choice in Kidney Transplantation: A Sequential Stochastic Assignment Model. Working Paper, Graduate School of Business, Stanford University, Stanford, CA.

[22] Su, X. and S.A. Zenios (2002). Optimal Control of a Medical Waiting System with Autonomous Patients. Working Paper, Graduate School of Business, Stanford University, Stanford, CA.

[23] Howard, D.H. (2000). Endogenous Determinants of Organ Supply. Unpublished Manuscript.

[24] Kaserman, D.L. and A.H. Barnett (2002). *The U.S. Organ Procurement System: A Prescription for Reform.* The AEI Press, Washington, DC.

[25] Kristof, N.D. (2002). Psst! Sell your kidney? *New York Times,* November 12, A27.

22 PLANNING IN THE FACE OF POLITICS: RESHAPING CHILDREN'S HEALTH SERVICES IN INNER LONDON

Mike Cushman[1] and Jonathan Rosenhead[2]

[1] Department of Information Systems
London School of Economics and Political Science
London WC2A 2AE, United Kingdom

[2] Department of Operational Research
London School of Economics and Political Science
London WC2A 2AE, United Kingdom

SUMMARY

There was a broad measure of convergence among health care professionals in a central London Health Authority that changes in patterns of care delivery and specialist staffing required a reduction in the number of inpatient units, a substitution of ambulatory care units, and an extension of community care provision. Strategic Choice was used in a series of workshops with intervening analysis to convert this 'in principle' agreement into a specific proposal that achieved consensus among stakeholders. This process is analysed in terms of the opportunities provided by sequential workshops and the difficulties presented by inter-organizational working and absent stakeholders.

KEY WORDS

Strategic Choice Approach, Problem structuring methods, Facilitated change workshops, Children's health services, Actor network theory

22.1 INTRODUCTION

This chapter describes an engagement in which the Strategic Choice Approach was used with multiple stakeholders to redesign children's health care provision for an inner city area covering two boroughs with a total population of some 375,000 people. Section 22.2 describes the genesis of the work, and the relevant context. In Section 22.3 the intervention itself is described. Section 22.4 covers the process of reporting outputs to the work's sponsors, and a summary of feedback received from key actors one year later. A concluding section considers some lessons that can be learned from the experience.

22.2 BACKGROUND

Planning children's health services for an area with a substantial population has some features in common with other medical specialties, and others that are distinct. Common features, in the United Kingdom (UK), include:

- the necessary participation of a number of service providers, as well as an agency charged with representing the health needs of the population – at the time of the study on which this chapter is based, the relevant Health Authority;

- a degree of interaction of the services under consideration with other specialties, with primary care provision and also with teaching arrangements;

- uncertainties as to current utilisation patterns, future tendencies, decisions in related areas and political priorities;

- the involvement of the public in the ratification of any proposals.

The principal distinctive feature of planning for children's health care provision is that the specialty is based not on the health condition of its patients, but on their age. There are evident reasons for this, including the need for segregated treatment environments for such vulnerable and impressionable patients. However, it does mean that children's health care has to deal with patients who, if they were adults, would fall within a wide range of different specialties. Among the resulting complications are the provision of Children's Accident and Emergency facilities, and the need to ensure continuity of care from children's through to adult services for those with life-long conditions (perhaps taking in along the way treatment in an adolescent unit).

22.2.1 The nature of the problem

This chapter reports and comments on a study to effect a significant re-orientation of children's health services in the inner London district of Camden and Islington. However, the conditions that provoked the review are by no means limited to those geographic boundaries.

Patterns of childhood illness and clinical practice have both been changing, resulting in a striking and sustained move away from hospitalisation. For example, over the last 20-30 years there has been a marked reduction in acute illnesses such as serious infections that used to be a common reason for admission to hospital. Improvements in practice and new treatments and technologies mean that many conditions can now be treated on a day-case basis [1]. Avoiding the trauma of inpatient stays is an undoubted benefit for the children concerned and their families — and now fewer children are being admitted, and for shorter stays. Furthermore, more children with chronic illnesses are surviving, and they and their families need to be supported at home, necessitating more services in the community.

The downside of all these advances is that historically located services need re-orientation to meet the new situation. There has been an increase in the need for day-case, outpatient, community-based and home care services. And there has been a reduction in use of hospital beds. In Camden and Islington it was policy that ill children should whenever safely possible be treated at home, in familiar surroundings and close to family and friends [1].

There were other related pressures on the Health Authority to reduce the concentration of resources in hospital-based pediatric services. Guidance from professional bodies, and the corresponding quality standards, are based on there being a 'critical mass' at any children's unit, in terms of both quantity and mix of clinical cases. The three secondary inpatient units in Camden and Islington were each well below an appropriate level. These caseloads were regarded by local pediatricians as inadequate to provide a satisfactory clinical experience to all trainees. The available child inpatient cases were being spread over too many units. Furthermore, the number of local pediatric training posts was due to be reduced in line with national medical manpower plans. Each of the three units was expected to lose one specialist registrar per year for three years. This put in question the ability of the existing units to cover adequately the full spectrum of children's conditions. To compound these difficulties, the reduction in hours worked by junior doctors had cut the amount of service provided by doctors in training.

It was against this background that the Camden and Islington Health Authority (C&IHA) encouraged the Director of Public Health to use her

1999 Annual Report [1] to initiate a process, involving the main providers of children's health services, to secure as much agreement as possible on how these difficulties should be addressed. By September 2000 this process, building on discussions over a period of years, had produced a strong consensus among those most involved that a reduction in the number of secondary inpatient units was both desirable and necessary.

A discussion document was drafted by a broad-based Children's Strategy Working Group, and published under the imprint of 12 local health care agencies [2]; a summary leaflet was then circulated widely as part of a three-month discussion process. A wide range of stakeholder groups were involved in different ways. Presentations were made to Trust and Primary Care Group Boards and to local government sub-committees. A conference was called for relevant voluntary organizations and a series of workshops was held with health care professionals (in children's emergency services, pediatric surgery, tertiary pediatrics, neonatal intensive care and maternity, primary care and community pediatrics, and workforce education and training) Each of these looked at the strengths and weaknesses of current arrangements, and considered how the service might provide a better fit for the future.

The discussion document stated that "we believe that there should be only one secondary pediatric and surgical inpatient unit" [2]. However, this was an input to the discussion process rather than a committed position. What was certainly not agreed was where closures would fall, the specification of the non-inpatient services which would complement the remaining unit(s), and where these services would be located. Indeed the entire September to November discussion process was conducted 'in the abstract', not linking any element of the possible new service to any particular real location. It is probable that without this self-denying ordinance, constructive discussion of principles would have proved impossible.

It was in the resolution of these issues, and their ramifications for other parts of the local health service, that the authors were invited to be involved. It was realised by all those involved that agreement in principle to closures was one thing, and agreement in practice was quite another. Building on the outputs of the discussion period there would need to be "a process of synthesis, consolidation, analysis and negotiation ... to draw together the various strands towards producing a set of concrete options for change and criteria for discriminating between them" [3]. A consensual outcome where major institutional interests were involved was likely to be hard to reach, yet without agreement among the institutions a suspicious public would be still more likely to react vigorously to any talk of closures and to exercise an effective veto.

This touches on a key aspect of the decision process. Health services issues around the world tend to be politically sensitive (see, for example, [4]). This is perhaps particularly true of Britain, where the National Health Service as an institution is highly regarded by the public. The legacy of bed and hospital closures by the Thatcher and successor administrations in the 1980s and 1990s had led to considerable popular mobilisation. There was a widespread presumption that proposals for closures were cost-driven, and possibly a preparation for privatisation.

22.2.2 Organizing the problem structuring workshop process

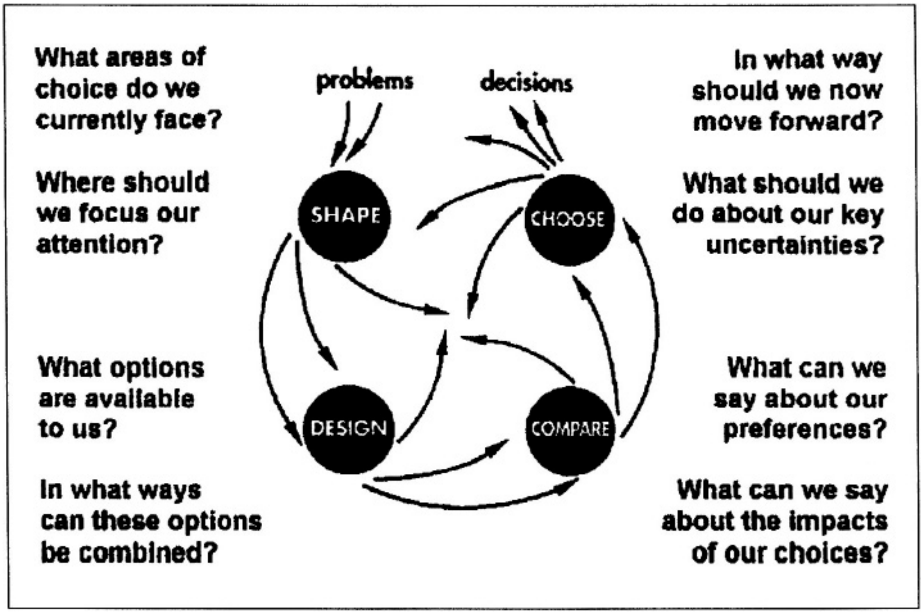
It was evident to the Director of Public Health and her team that the situation with which they were grappling was characterised by high levels of complexity and uncertainty, compounded by the need for the involvement, participation and commitment of multiple stakeholders. Her enquiries as to methods which could be effective in helping in such situations led her to consult with the authors about the possible use of problem structuring methods.

Initial meetings and documentation led us to propose the Strategic Choice Approach (SCA) [5] as, in principle, the most appropriate method to employ, though we wished to retain the freedom to adopt other methods depending on the evolution of the engagement. In particular, if the question of sequential implementation of service configurations were to be reached (which would be well down the road) then robustness analysis [6] was thought likely to be a valuable complement. (In fact, this option did not materialise.) A very rough scoping of the exercise led to a proposal for three whole-day workshops at monthly intervals. In the event this time-scale proved appropriate.

SCA is a participative method for working with groups facing a joint problem situation characterised by complexity and uncertainty which requires strategic thinking. Either complexity or uncertainty can undermine the clarity of thought and understanding necessary for confidence in decision-making. Together they can be a lethal combination. SCA offers a format and a procedural framework for eliciting information from the group and its members, and then iteratively growing a picture of the interacting issues by further elicitation and structuring. This shared representation, and the tools used to develop it and to explore its implications, enables the group to establish what commitments can and should be made, and in what areas additional information is needed to better inform decisions. (For an accessible introduction to SCA, see [7].)

The SCA process rests on four modes of decision making, as shown in Figure 22.1. Though an engagement using SCA will generally start with *shaping* and conclude with *choosing,* there is no necessarily linear path through the modes. Understanding gained, or obstacles encountered, may indicate the advantage of returning to an earlier stage of analysis for reformulation. In each mode there are tools to assist in the elicitation and structuring of information. These tools are low-tech and capable of being understood intuitively by lay participants.

Figure 22.1 The Strategic Choice Approach


What areas of choice do we currently face?
Where should we focus our attention?
What options are available to us?
In what ways can these options be combined?
problems
decisions
SHAPE
CHOOSE
DESIGN
COMPARE
In what way should we now move forward?
What should we do about our key uncertainties?
What can we say about our preferences?
What can we say about the impacts of our choices?

The key concept in SCA is the 'decision area'. In *shaping* mode, the group identifies the set of interconnecting decision areas that constitutes their problem, and prioritizes a manageable number of them as a 'problem focus'. In *designing* mode, the options for choice in each decision area in the problem focus are identified, and incompatible combinations are weeded out to establish a list of feasible 'decision schemes', each of which consists of one option from each decision area. In *comparing* mode, decision schemes are rated against each other on a range of criteria generated by the group. This may eliminate some schemes, or even identify one which is clearly preferred. More commonly it highlights what uncertainties obstruct commitment to any scheme in its entirety. In *choosing* mode, this information is consolidated into a 'progress package' consisting of agreed commitments, explorations to reduce key uncertainties, and contingency plans.

Various aspects of the problem of re-designing Camden and Islington's children's health services seemed to make Strategic Choice a good choice. There were evidently a considerable number of inter-linked decision areas, each with alternative options – for example, numbers of units of different kinds, and locations of those units. And there were numerous and varied uncertainties.

Maternity provides a double example of the inherent uncertainties. There was a link between maternity and pediatric provision, through the practice of providing shared staff rotas between neonatal intensive care units and pediatrics in smaller units. But it was unclear whether this was a firm constraint. Furthermore, there was an ongoing review of neonatal services provision across London. As the maternity review was being conducted by a different sponsor, C&IHA could not require that it be postponed and was unwilling to delay the pediatric review for an unspecified period to allow the maternity review to be completed. In any case our problem was more than complicated enough, and to merge it with another one of comparable intricacy would render it still more intractable.

Another factor speaking in favor of SCA was the political sensitivity of the issues under consideration. Unusually among analytic methods, SCA can incorporate political factors, or the unpredictable reception of proposals in the wider world, by representing them as uncertainties. It has already been mentioned that health service changes or shortcomings, and closure proposals in particular, are capable of generating quite intense political disturbances. Both the population affected and their political representatives take these matters very seriously. The high-profile organizational participants and central London location of this study guaranteed that it would not be the exception to this particular rule. Indeed there were aggravating factors. A long-serving Camden Member of Parliament had only recently ended a well-regarded tenure as Secretary of State for Health; and the Hampstead and Highgate Express, covering much of Camden, was regularly garlanded as among the highest quality local papers in the country. It was an effective campaigner, and had a strong readership among the concentration of national movers, shakers and opinion formers living locally. Therefore the capacity of SCA to accommodate the political dimension of the issues under discussion was a valuable bonus.

It was agreed that SCA should be the method for use at the workshops. Membership of the 'core group' to take part in these workshops was given a great deal of detailed consideration. The participants needed, between them, to represent both the main stakeholder institutions, and the principal relevant professional and disciplinary groupings. As consultants, we explained the importance of keeping the group size small to facilitate constructive conversation, and argued for ten members as the upper limit. However, our C&IHA collaborators came to the conclusion that a feasible design could not be achieved within that constraint, and we accommodated ourselves to a group size of 12. Ten were members of the Children's Strategy Working Group and so were broadly familiar with the issues that would need to be addressed, and in addition there were representatives of the two local Community Health Councils who had been closely monitoring the process on behalf of users.

To make this account of the workshop process understandable, it is necessary to sketch in the roles of the key institutional players in relation to children's health services in Camden and Islington. We should start with C&IHA itself, as the Health Authority has a particular role in the UK health service which is not precisely replicated elsewhere.

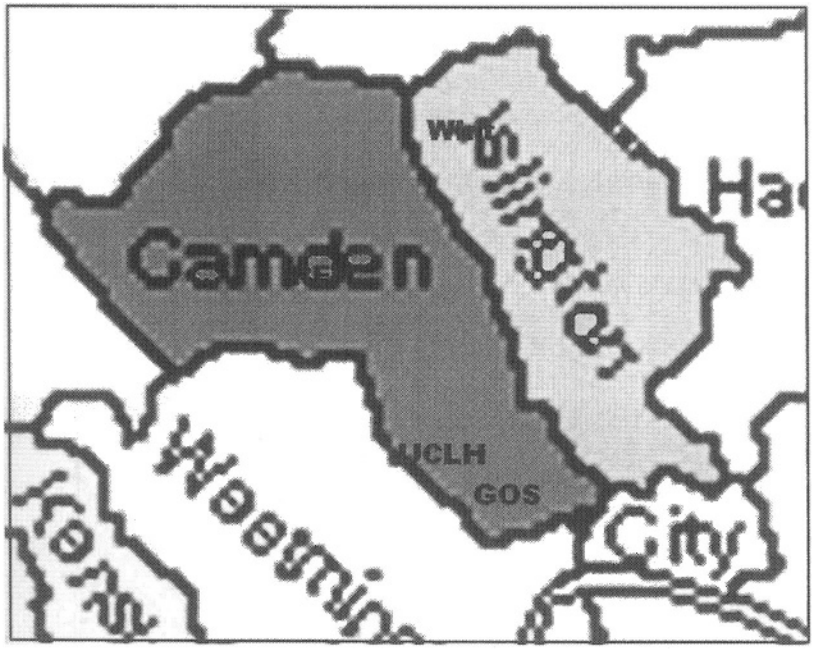
At the time of this work, any Health Authority was responsible for ensuring that the health needs of the population in its area were met, and it received from central government the bulk of the funds made available for this purpose. (Changes in these arrangements have occurred since the project was carried out.) However, the actual delivery of health services was and is provided by a number of autonomous health care trusts, comprising a variety of types of hospitals and hospital groupings, as well as trusts dedicated to the provision of community-based or specialised services. The bulk of patients are treated by trusts in or geographically close to the Health Authority area in which they live.

C&IHA's area consisted of the two inner London local government boroughs of Camden and Islington with a total population of about 375,000 people, of which some 65,000 are under 16 (a lower proportion than the inner London or UK average). The maximum East-West distance is about 10 km, and the North-South span is around 6 km (see Figure 22.2).

Both boroughs, but in particular Islington, are characterised by areas of intense deprivation and large public housing estates. The population is both ethnically and linguistically highly diverse. Both boroughs, but in particular Camden, have very affluent districts. This mixture of rich and poor is characteristic of many London boroughs.

The institutions represented at the workshops, in addition to C&IHA itself, were:

Figure 22.2 Map of Camden and Islington showing locations of main hospitals



<u>**Camden and Islington Community Health Services NHS Trust**</u> The trust covers the same area as the Health Authority and was responsible for all community-based services including health visitors, children's health clinics, district nurses, midwives and domiciliary care. It has major responsibilities for preventative medicine as well as for medical care.

<u>**Great Ormond Street Hospital**</u> Great Ormond Street is regarded as the pre-eminent children's hospital in the UK, with an international reputation for both care and research. Located near the southern tip of Camden, it currently took only tertiary patients. However, it was expressing interest in developing a secondary inpatient pediatric service, which would require the acquisition of additional premises. Great Ormond Street is relatively close to University College Hospital, with whom it has been developing cooperative arrangements.

<u>**University College London Hospitals (UCLH)**</u> UCLH was formed from the merger of two major teaching hospitals, University College and Middlesex. Their services (and those of a number of other units) were shortly to be brought together in a major new hospital building currently under construction through a public finance initiative (PFI) arrangement. UCLH is certainly one of the most prestigious and powerful teaching and

research hospitals in the country with an international reputation. It is well located for a range of public transport services.

Royal Free Hospital The Royal Free is another distinguished London teaching hospital, located rather to the north of the borough of Camden. It is linked to UCLH through a joint medical school. Far more than UCLH it draws patients not only from Camden and Islington, but also from Barnet, Enfield and Haringey, the Health Authority (just being merged out of two predecessor authorities) immediately to the north of Camden and Islington.

Whittington Hospital The Whittington is a large hospital located at the northern edge of Islington. It functions broadly as a district general hospital, and services an area spanning across the boundary into Barnet, Enfield and Haringey with a high population and population density.

Moorfields Eye Hospital The country's leading specialist eye hospital, Moorfields provides a mix of secondary and tertiary care. It was directly, if somewhat tangentially, involved in the redesign of pediatric services through its inpatient provision for children.

Community Health Councils Community Health Councils (CHCs) were bodies charged with representing the interests of the public in their areas. (They have since been abolished by the national government.) Both Camden and Islington CHCs were represented at the workshops by their Chief Officers and/or Chairs.

Primary Care Groups were represented by a long established and well-respected local general practitioner (GP) who also voiced the viewpoint of GPs.

At the Workshops, senior representatives of these groups were confronted with a problem which could be summarised as:

- Which of Great Ormond Street, Royal Free, UCLH and Whittington should have secondary inpatient units?

- How many non-inpatient 'ambulatory care centers' should complement these, and where should they be located?

- How and to what extent should community services be strengthened?

22.3 THE WORKSHOP PROCESS

Our preparation for the workshop not only consisted of discussion with the Director of Public Health and her team. One of us attended a meeting of the

Children's Strategy Working Group which occurred between the commissioning and the start of our project; and we both read the distributed discussion document, reports from consultation meetings and relevant background materials. From these we distilled what seemed to us to be some principal areas of choice in the situation, and pre-prepared a set of a dozen or so ovals (large oval 'Post-it' notes with particularly convenient adhesive properties) each conveying concisely one of these candidate *decision areas.*

22.3.1 Workshop 1, January

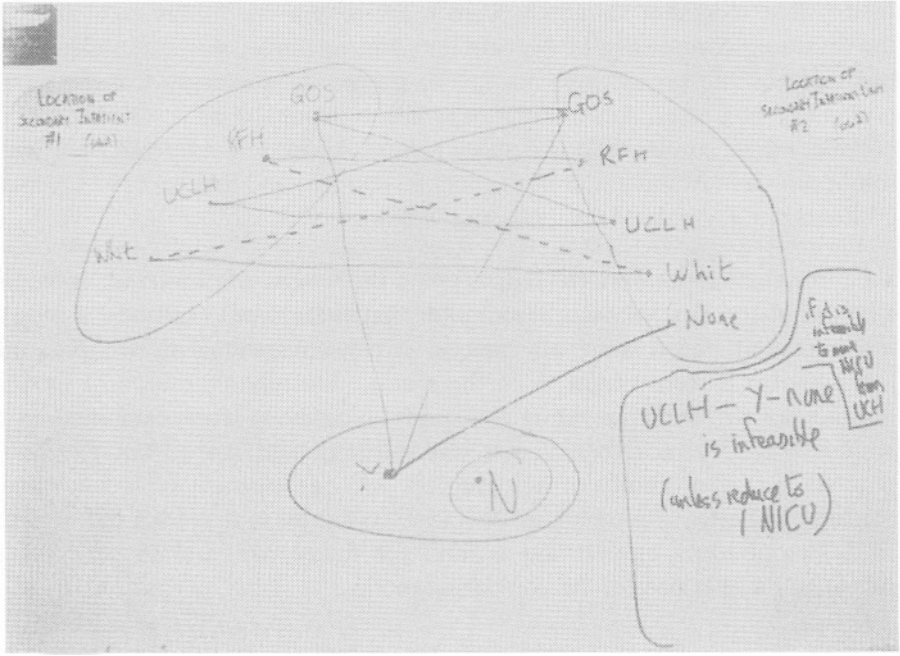
These ovals formed the starting point of the first all-day Workshop, held in January 2001 at a well-appointed location away from any of the participants' places of work. They were displayed on an end wall papered with A1 flip charts. An initial discussion confirmed that these were broadly the issues that mattered, though the group made some alterations to the way they were formulated. This discussion also demonstrated the interconnection of the issues; only a very few of them could be set aside as separable or secondary. When one decision area was raised, other factors were at once identified as needing to be taken into consideration with it, and these led on to others in a similar fashion. It seemed both that no decision could be taken in isolation, but that the ensemble of decisions was too complex to be comprehended simultaneously.

This experience, demonstrating in effect the need for some analytic assistance, provided a persuasive motivation for the use of Strategic Choice. The first pass through the approach addressed the remaining decision areas, and the group was asked to agree preferably no more than three of them as an initial, priority, *problem focus.* The remaining ovals were 'parked' to one side, and discussion was centered on:

- the location of the first inpatient unit;
- location of the second such unit (if any);
- whether neonatal intensive care units must always be co-located with inpatient pediatrics.

This last question was a technical one, based on the argued need to share staff rotas for the two activities to provide sufficient coverage 24 hours per day, seven days per week. It had provoked lively debate in the initial discussion, but of course had not then been resolved as the argument cascaded on. The provisional reduction in complexity provided by the problem focus allowed the matter to be resolved decisively. An *option graph* was developed of the problem focus, looking at the options available within each of the three decision areas, and identifying which combinations were infeasible (see Figure 22.3). It was then realised that to say 'yes' to the proposition would in fact rule out one of the strong organizational contenders to house an inpatient unit. When it was clear that no one in the group was willing to do this, it became evident that the answer was 'no': co-location was certainly desirable, but not an absolute requirement. This problem focus was taken no further – it had served its purpose by resolving a troublesome issue, thereby simplifying the remaining problems. In fact the achievement was startling – the perceived link between pediatric and maternity beds had always been seen as a complication that made the problem almost insoluble.

Figure 22.3 Option Graph showing the infeasibilities between the requirement for collocation with neonatal intensive care, and the locations of the proposed pediatric inpatient unit, and the second unit (if any)


Location of Secondary Inpatient #1 (unit)
GOS
RFH
UCLH
Whit
GOS
RFH
UCLH
Whit
None
Location of Secondary Inpatient Unit #2 (unit)
Y
N
UCLH – Y – none is infeasible
(unless reduce to 1 NICU)
if it is infeasible to move NICU from UCH

The process by which this advance was made was typical of the Strategic Choice approach. At any time one of the facilitators was actively engaged with the group, sometimes asking for clarification to avoid miscommunication, sometimes steering the discussion in what seemed likely to be productive directions, and sometimes operating the technical aspects of SCA with ovals and marker pens. The second facilitator would at times be capturing the evolving structure on the STRAD software [8] but more often observing the discussion from a slightly less engaged perspective. (Full details of the STRAD software can be found at the website http://www.btinternet.com/~stradspan/.) This is a valuable backup, as the lead facilitator, in the thick of things, can easily fail to notice aspects of group dynamics or problem content. And of course the roles of the two facilitators were exchanged periodically. (See [9] on the role of the facilitator.)

During any stage of the discussion, aspects of the problem situation surfaced that were clearly relevant but not to the topic immediately under discussion. These were captured (on ovals) for possible later reference. Of particular interest were *uncertainties,* areas of missing information or disagreement whose resolution might remove obstacles to progress. Other aspects were collected together under the heading of *comparison areas,* in effect criteria which could prove relevant to the choice between alternatives.

In this first workshop, attempts were made to achieve further reduction of the disabling complexity of the problem. For example, it became clear that members of the group were using the phrase ‘Ambulatory Care Center’ without a shared understanding of what it would consist of. Would it conduct minor surgery and use anaesthetics? Allow self-referral? Deal with minor injuries? Admit children to general Accident and Emergency? Operate 24 hours per day? These design issues were resolved.

Two initial attempts were made to employ the *comparing* mode of Strategic Choice. In the first, on the assumption that there would only be a single inpatient unit, the relative advantages and disadvantages of two possible locations for it were explored. In the second, the relative merits of having one versus two inpatient units (locations undefined) were examined in a similar way. No clear conclusion was reached; and indeed the exercises pointed up the difficulty of agreeing on just one element of the eventual package while leaving others un-specified. What these exercises did do was to flush out ideas on relevant criteria for future use, and to serve as a rehearsal for later uses of the comparing mode.

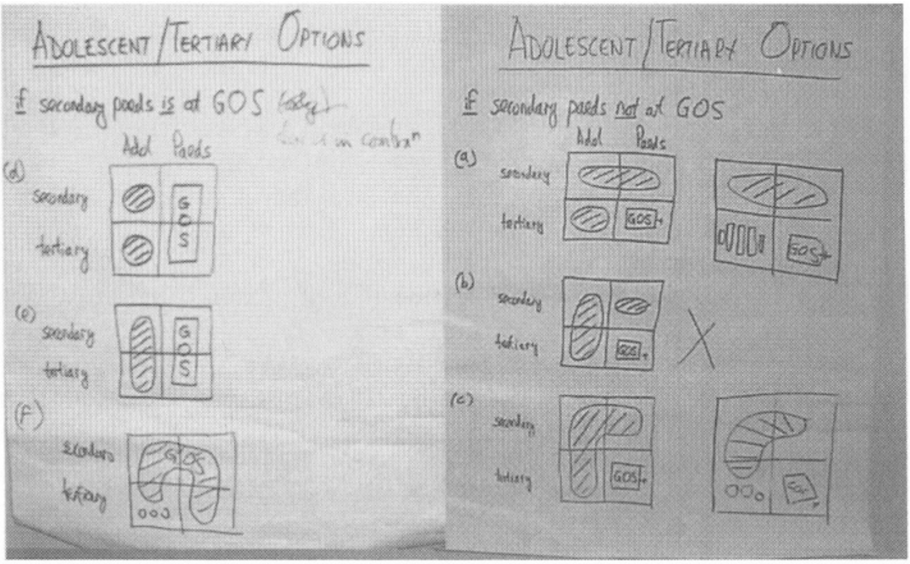
At the end of this first workshop the accumulated uncertainties were reviewed and, for each of those thought to be of significance, a group member agreed to come back with some additional information. Their initials were written on the corresponding ovals, as evidence of their commitment. It was agreed that the second workshop would focus on the relationship between the number (not locations) of inpatient units and Ambulatory Care Centers and the level of community provision; and on how these link to tertiary, adolescent and maternity services.

22.3.2 Workshop 2, February

Before proceeding to new work, the members of the group were asked to review the conclusions reached at the previous workshop and whether they were happy to proceed on the basis of the progress made there. It was important that the group did not feel 'railroaded' but accepted the logic of where the argument had got to. (A similar procedure was followed at the start of Workshop 3 also.)

Two issues that had slowed down progress at the workshop were the relationship of decisions that might be made about the location of secondary inpatient pediatric units to the care arrangements for tertiary pediatric inpatients, and for inpatient secondary and tertiary inpatient adolescents. Some graphical representations of alternative allocations of the resulting four patient categories between Great Ormond Street (the existing tertiary center) and an unspecified secondary pediatric inpatient unit had been pre-prepared by the consultants (see Figure 22.4). These appeared to generate a more focussed discussion, and the group rather swiftly agreed to a modification of one of the schemes illustrated, in which a significant role for adolescent tertiary services would go wherever the adult expertise was located. (Members of the group amended the drawings themselves, always an indication that a representation has proved useful.) The result was that

Figure 22.4 Options for adolescent and tertiary care patterns of provision if pediatric and adolescent services or secondary and tertiary care were to be co-located



these issues lost their ability to entangle the subsequent discussion in unresolved questions.

Other inter-related work that had been commissioned by Workshop 1 was to investigate the 'critical mass' of annual inpatient admissions needed for an inpatient unit, in particular to provide an adequate range of cases for the training of junior doctors; and to clarify existing and predicted activity levels. Discussion of the reported results was not conclusive, but tended to support a single inpatient unit solution.

As agreed previously, the group took as their initial problem focus

- the number of inpatient units
- the number of Ambulatory Care Centers
- the level of increase in community services.

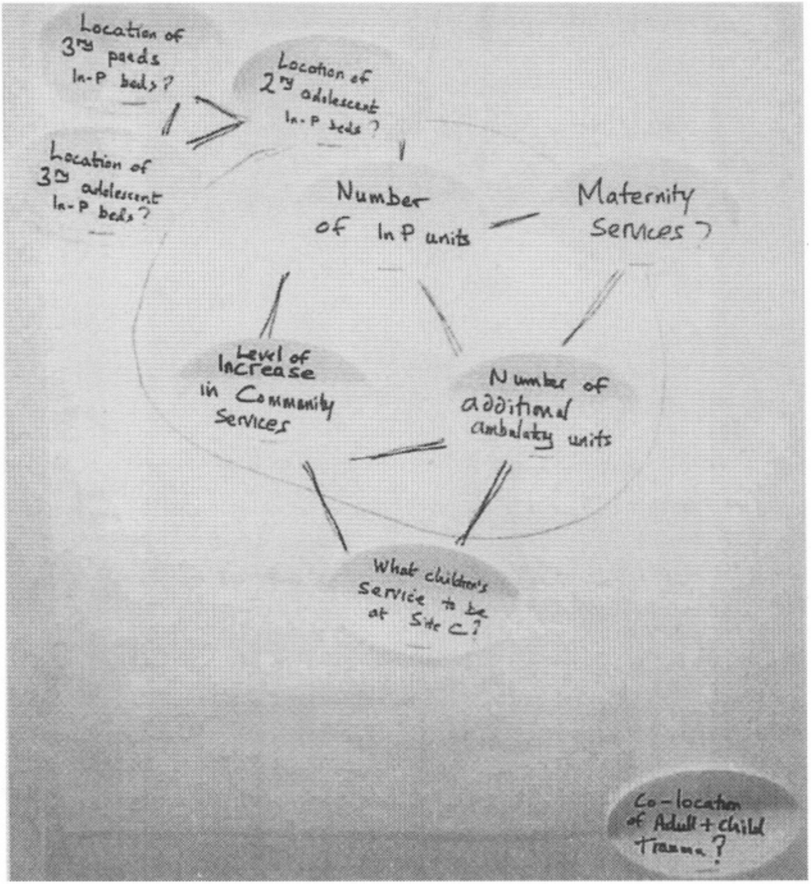
Ovals were used to locate other inter-linked decisions around the boundary of this problem focus, as a guarantor that the impact of any decisions within the focus on these other issues would subsequently be subject to scrutiny (see Figure 22.5). A tabled paper, commissioned by the previous workshop, had specified the days and hours of opening, and consequently the types of patient which could be handled, corresponding to each level of community service investment.

An *option graph* showing the options in each of these three decision areas, and the relationships between them was developed through group discussion, and is shown in Figure 22.6.

Either one or two inpatient units were considered, as well as up to three ambulatory centers, and an increase in community and primary care services that might range from zero to large. The lines drawn in Figure 22.6 are *option bars,* indicating incompatible options. Each of these bars resulted from discussion in the group – e.g., about the level of particular scarce resources needed to maintain that combination of options. This discussion also led to the exclusion of particular options on policy or practicality grounds. Then, by the process called the Analysis of Inter-connected Decision Areas (AIDA), the feasible combinations of options, one from each decision area, were worked out (see Figure 22.7).

There were rather few feasible combinations of options remaining. Discussion of these led the group to a further 'policy' conclusion, that no scheme with only a single 24-hour access point could be contemplated. This meant that any combination involving only a single inpatient unit required a

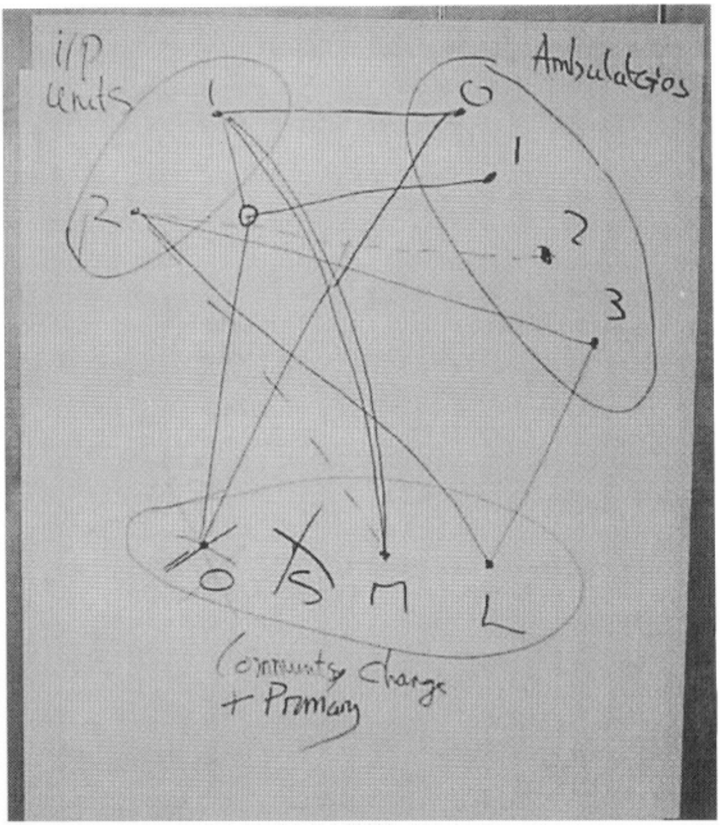
Figure 22.5 Decision focus for Workshop 2*


Location of 3ry paeds In-P beds?
Location of 2ry adolescent In-P beds ?
Location of 3ry adolescent In-P beds ?
Number of In P units
Maternity Services ?
Level of Increase in Community Services
Number of additional ambulatory units
What children's service to be at Site C?
Co-location of Adult + child Trauma ?

* Text of "Post-Its": Location of 3^{ry} pediatric in-patient beds; Location of 3^{ry} adolescent in-patient beds; Location of 2^{ry} adolescent in-patient beds; Number of in-patient units; Maternity services?; Level of increase in community services; Number of additional ambulatory units; What children's services to be at Site C; Co-location of adult and child trauma

'large' increase in community and primary care services (since a 'medium' investment corresponded to seven-day, 8 a.m. to 8 p.m. working, and ambulatory centers had been defined as having approximately 12 hour daily opening times). The remaining schemes, marked B, D, E and F in Figure 22.7, had either one inpatient unit and a large increase in community provision, or two inpatient units and a medium increase in community provision. In the former case there could be either one ambulatory care center or two, while in the latter there could be at most one ambulatory care center.

Figure 22.6 Workshop 2 option graph showing options for number of inpatient units, ambulatory units and increase in provision of community and primary care, and their infeasibilities.

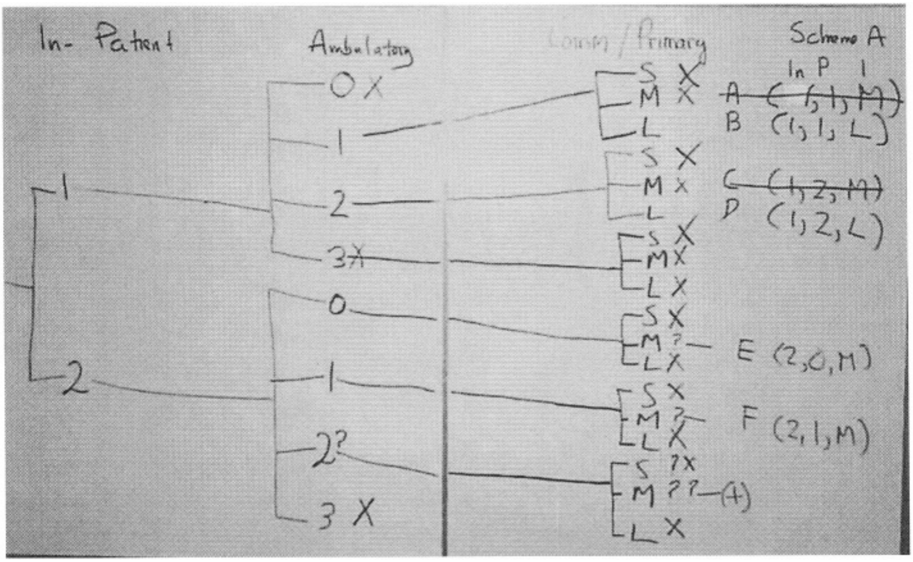


The group was now ready to compare two distinctively different schemes:

- 1 inpatient unit, 2 ambulatory care centers, large community increase, vs.
- 2 inpatient units, 1 ambulatory care center, medium community increase.

These were placed on a standard *comparative advantage* chart, on which the criteria identified at the last workshop were added in agreed order of priority. This chart is shown in Figure 22.8. After discussion of each criterion, group members each marked their estimates of the balance of advantage between the schemes with adhesive stickers. This array was then, in further discussion, consolidated into a range of possibilities and a central point. This assembled information was assessed in a final group discussion, in which the prevailing view was that there was a clear comparative

Figure 22.7 Workshop 2 option tree showing feasible decision schemes defined by numbers of inpatient units and ambulatory units and increase in provision of community and primary care



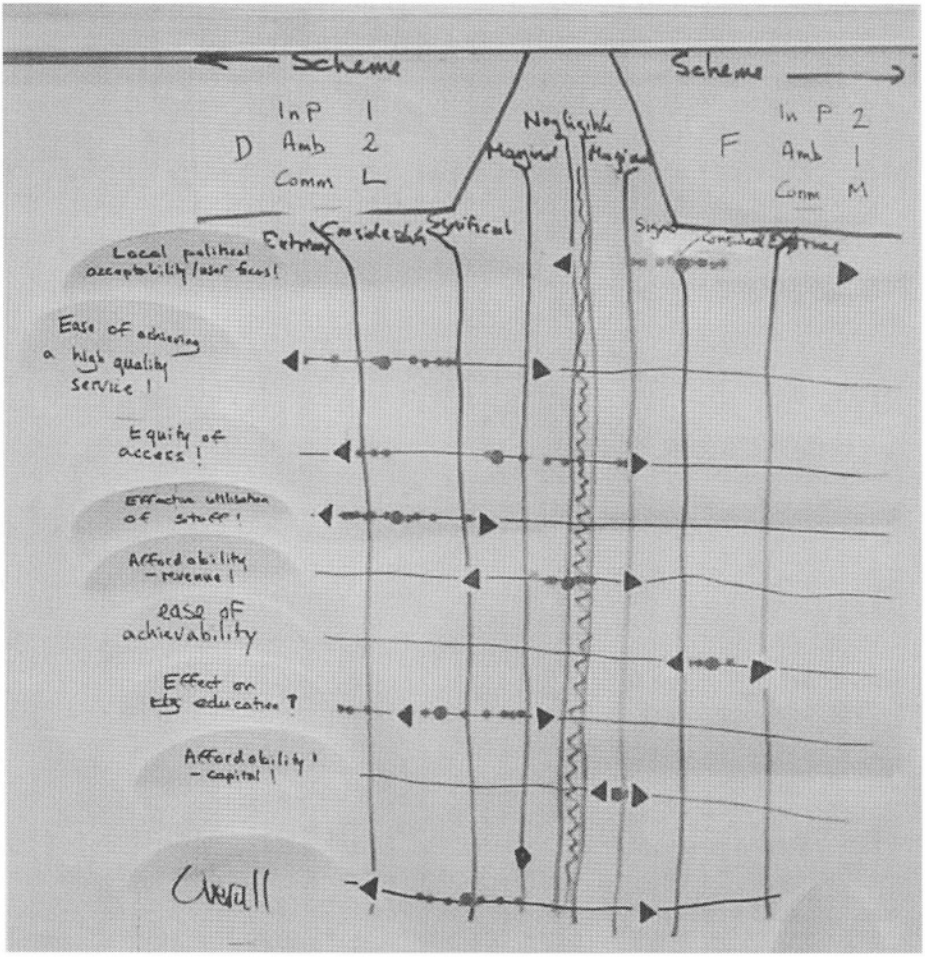
advantage in favor of the single inpatient unit scheme and a large increase in community services.

22.3.3 *Workshop 3, March*

Once again research carried out by the Health Authority in response to uncertainties surfaced at the last workshop was presented. But first the group was asked to confirm whether the decision at the last workshop was for the (1, 2, L) scheme. This generated a lengthy discussion, not all of it directly germane to the decision at hand. Matters debated included the possible sequencing of changes, the needs of different types of patient, connections to other parts of the health service, and the likely public reaction. During this discussion it was agreed to rename Ambulatory Care Centers as 'Specialist Children's Centers' (SpeCCs) to provide a more appropriate and acceptable image. Halfway through the morning, the group was ready to confirm a decision in favor of a single inpatient unit.

After the break, discussion was joined on the question of how many SpeCCs should accompany the unit – i.e., should it be scheme B or scheme D in Figure 22.7? A simplified version of the comparative advantage chart was used (Figure 22.9) in which the criteria that favored scheme B, those that favored scheme D, and those that were neutral between them were

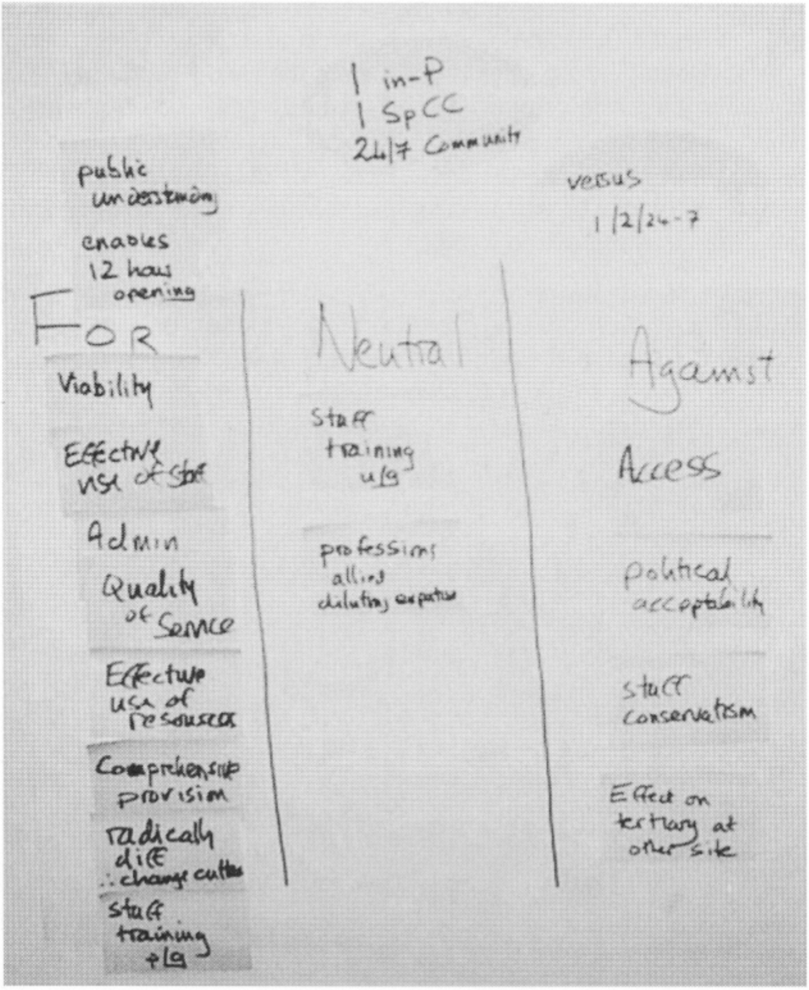
Figure 22.8 Comparative advantage chart*



* Comparison of scheme D (1 inpatient unit, 2 ambulatory units, and large increase in community provision) and scheme F (2 inpatient units, 2 ambulatory unit, and moderate increase in community provision). Is advantage negligible, marginal, significant, considerable or extreme? Text of "Post-Its" for comparison areas: Local political acceptability/user focus; Ease of achieving a high quality service; Equity of access; Effective utilization of staff; Affordability – revenue; Ease of achievability; Effect on education; Affordability – capital; Overall [advantage]

identified. The criteria used for this comparison were generated by the group, and were not those used in the previous workshop, as the issue under consideration was different. The weight of factors in favor of a single SpeCC was fairly rapidly persuasive for the whole group.

Figure 22.9 Comparing number of SpeCCs*



* Text of Text of "Post-Its": [Scheme considered] 1 inpatient unit; 1 SpeCC (Specialist Children's Center); 24/7 community provision.

For: Public understanding; Enables 12 hour opening; Viability; Effective use of staff; Administrative ease; Quality of service; Effective use of resources; Comprehensive provision; Radically different ∴ change culture; Staff training p/g [post-graduate].

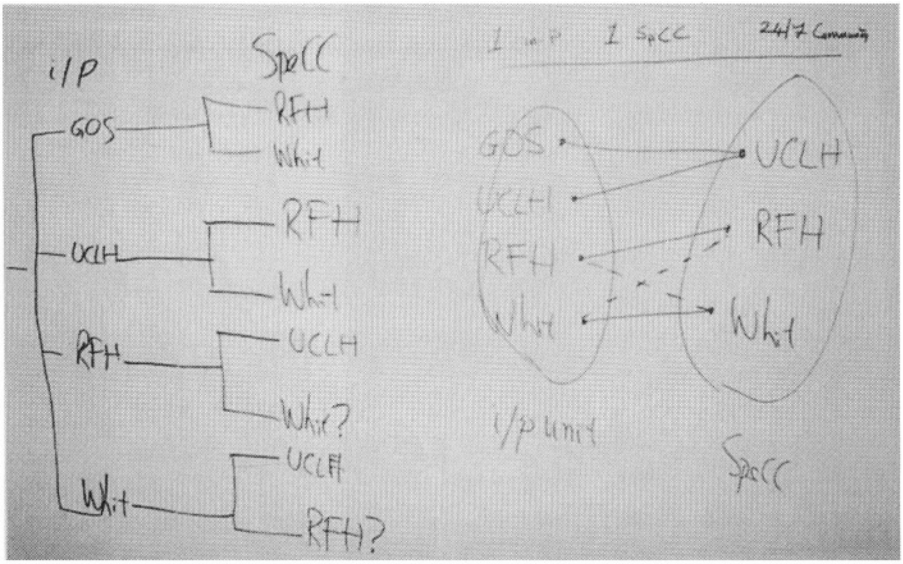
Neutral: Staff training u/g [under-graduate]; Professions allied [to medicine] diluting experience.

Against: Access; Political acceptability; Staff conservatism; Effect on tertiary at other site

Having come down in favor of one inpatient unit and one SpeCC, the remaining question, and the most politically charged, was their location.

Once again an option graph was used to identify the option bars and hence the locational schemes that remained feasible (Figure 22.10). The option bars broadly indicated the impossibility or undesirability of co-location of the units, or of geographical concentration within the Health Authority area. Extensive further discussion followed on availability of space on particular sites, public transport accessibility, political acceptability, ease of implementation and effect on existing services. All these issues and others were captured on ovals. In the process, the schemes under consideration were whittled down from eight to four.

Figure 22.10 Location options for scheme composed of 1 inpatient unit, 1 SpeCC (Specialist Children's Center), and 24/7 community provision



At about this point an unexpected uncertainty surfaced. It became clear in the discussion of some of these criteria that not all group members had secure mental images of the geography of the boroughs and the locations of all of the facilities under discussion. A London street map was hurriedly obtained and roughly transcribed to flip-chart size. It was clear from the reactions that several minds were made up, or at least provisional decisions confirmed, by the provision of this simple graphical aid!

It was agreed that the final assessment should be made by confidential ballot. Each group member was given five adhesive stickers, which they could allocate freely between the alternatives. There was no comparative advantage chart, but the criteria of the previous discussion were displayed for consultation. The result was clear-cut. The same location for the SpeCC received all but two of the 60 'votes'. One location for the inpatient unit received two and a half times as many choices as its nearest competitor. These two locations in combination received 50% more than all other combinations combined. These results were regarded as decisive by all the participants, who accepted it as legitimate, and as the concrete crystallisation of the logic that they had been elaborating and clarifying over the entire workshop process.

It had been a long journey from the state of disabling complexity and uncertainty which the group had experienced at the first of these workshops. The process and its outputs had the assent of the whole group, including representatives of those institutions that would lose services. The workshop concluded with a discussion of procedures for resolving the various issues 'parked' along the way; and on how to take the recommendations forward through the various stages and decision-making required before they could be implemented.

22.4 IMPACT OF THE STUDY

22.4.1 Reporting the outcomes

For legitimacy and implementation to be achieved, the results of the workshops had to be fed back to the Children's Strategy Working Group and also communicated to key individuals and stakeholders. These two strands were progressed through a mixture of formal and informal processes.

The final workshop had been held on a Friday. It was recognized in the final discussion that group members would be under immense pressure to reveal the workshop outcomes as soon as they returned to work. Attempting to keep the recommendations confidential for the time being was simply not an option. This meant that a careful dissemination strategy was crucial if the workshop gains were not be lost through hostile media coverage and instant political opposition. Already between the second and third workshops, and before the most sensitive decisions on unit locations had been reached, some information had leaked and articles had appeared in the local papers headlined "Royal Free fights to save children's casualty services", "Who will care for our children?" and "Hospital plan must not make children suffer".

The Health Authority representatives were instructed to ensure that their Chief Executive and those of the hospital trusts were informed of the workshop outcomes before the end of the weekend. The Health Authority Chief Executive in fact succeeded in briefing all the local Members of

Parliament (MPs), the Minister responsible and the Regional Health Authority by the following Monday.

Presenting the outcomes as a set of recommendations on service organization and not just on the location of inpatient services was identified as important if the proposals were not just to be seen as service closures. Over an extended period there had been a series of hotly contested plans to close hospitals (and accident and emergency units) in London. Some of these could be justified as a re-alignment to take account of population shifts out of Central London; or alternatively to allow concentration into large units that could support increasingly specialised and technologised services. However, popular perception was that such closures were driven by a Government agenda to cut health service costs rather than improve provision. There was thus a raw nerve to be touched.

The way in which the proposals would be seen by the multiple stakeholders not represented in the workshops was therefore a major area of concern. While this would have been true anywhere, the location of Camden and Islington in the center of London and the proximity to the offices of the national media made it even more pressing, as did the presence of the homes of many national journalists in the two boroughs. As one of the local MPs had just ceased being the Secretary of State for Health, another was a current Cabinet Minister and a third was an influential junior minister, the sense of political pressure was even more acute. The workshops were being held during the run-up to the 2001 General Election, expected to be held on May 3. Any publicly aired proposals to close units were likely to become incorporated in election campaigning and would thus potentially receive much publicity but little dispassionate consideration.

The second main strand of the process of feeding the workshop results into the policy process was a report to the Children's Strategy Working Group. This was the group from which the workshop participants were drawn, and whose endorsement of the results of the workshops was required. (Formally the workshops' recommendations were advice to the Strategy Group.) There had already been an interim report produced between the second and third workshops. At this meeting, few members of the workshop group, apart from the Health Authority members, were present. Having put time aside for the workshops, most members had not prioritized attending this meeting as well. The consequence was that the feelings of exclusion felt by people who would have liked to be part of the workshops, but were not, expressed themselves as mistrust of the report of the facilitators and of the process they described. This mistrust was difficult to counter in the absence of participants who could describe their experience of that process.

Consequently, the Health Authority staff and the facilitators prepared more carefully for the final report-back meeting, six weeks after the third workshop. In order to convince the Strategy Group of the robustness both of the recommendations and of the process by which they were reached, effort was put into ensuring that several members of the workshop group attended. Special attention was paid to ensuring that members who were not representatives of the principal hospitals were present as it was felt that, as more disinterested parties, their voices would carry more weight. While there were some reservations, particularly from the representatives of the neighboring health authorities who were concerned that the needs of their residents may not have been considered sufficiently, the workshop outcomes were well received overall and endorsed.

22.4.2 What happened next

At the beginning of April 2001, the government decided to postpone the general election by one month because a major foot-and-mouth disease outbreak made campaigning in many rural areas impractical. One result of this delay was a potential gap in policy announcements by Government ministers; the planned succession of announcements of initiatives, necessary to maintain campaigning momentum was disrupted. The announcement on April 23 by Alan Milburn, the Secretary of State for Health, of a restructuring of the management of the Health Service [10] can in this light be seen as a political initiative to fill a news gap. His proposal to abolish health authorities was totally unexpected and came abruptly in the middle of an already existing process of setting up Primary Care Trusts and transferring budgetary, but not planning responsibilities, from the Health Authorities to the Trusts. The eventuality of such a change had not figured in the uncertainties considered during the workshops, nor could it have done, as even well informed observers had no inkling of this proposal.

The consequences for the reorganization of children's health services were terminal. No Chief Executive would take the risk of becoming embroiled in a potentially controversial service change at this juncture. All the Health Authority officers who would have been responsible for carrying through the changes had to concern themselves with their immediate futures – all of their posts were to be abolished and they had to apply for posts in the new structure or elsewhere. Responsibility for planning health services in Camden and Islington passed to the new North Central London Strategic Health Authority (SHA), one of the 28 new SHAs covering the whole of England to be set up by April 2002.

The North Central London SHA consisted of Camden and Islington together with Enfield and Haringey, and Barnet, the neighboring health authorities that had been more sceptical of the workshop proposals. (These authorities had just been merged in April 2001.) The senior management of the new SHA proved to be drawn from these authorities and few Camden and Islington managers were appointed to senior positions in it. It could have been argued that children's services should be considered across the whole SHA area. However, in fact, the new SHAs were under much closer central scrutiny and direction than the former health authorities. Major changes in service provision were thus more politically exposed. The proposals informed a much wider discussion of children's, young people's neonatal and maternity services across the wider area.

However, there has been action as a direct result of the workshops. Community-based care has been radically reformed on the lines recommended in the workshops: opening hours have been extended and seven-day cover provided. Opening hours are likely to be extended further towards 24 hour, seven-day provision. This was achieved as a direct result of the consensus reached at the workshops, and could be implemented without either a consultation process or sanction from the Department of Health. The provision of ambulatory care has also been strengthened. At the workshops, these changes had been developed and proposed as an integral part of a comprehensive service model which included the desirable and necessary alterations in inpatient provision, rather than as stand-alone initiatives. These other changes at present remain in abeyance – though the pressures which provoked the workshops do not.

The experience of the workshops has also underpinned subsequent moves towards a Children's Services Network in Camden and Islington, and the full advice remains as an available resource when the issues of inpatient care are eventually addressed. They have a continuing status because of the process by which they were reached. As one key participant put it [3],

> "I think one of the features of the group and this piece of work was that it was a well embedded, you know, it's been well embedded in the folklore of Camden and Islington. The tradition of Camden and Islington, for many years, and individuals have been around and around this set of problems and been involved in work over a number of years. And so the people who were involved were all well able, for a fairly sort of strategic exercise, were well able to be articulate and to contribute and to think rationally."

22.5 DISCUSSION

Many lessons can potentially be learned from a rich encounter of this kind. Here we will focus on two broad areas which we think worthy of further attention.

22.5.1 Working between workshops

There is now a considerable literature on the pragmatics of engagements using problem structuring methods (PSMs). (See in particular [11].) Broadly, the literature deals with aspects of the client-consultant interaction in the context of model-based group decision support. There are also discussions dealing with method-specific issues.

The main focus of these accounts is on what happens in the workshop itself. There are of course exceptions to this rule. In her survey of the views of clients of the SODA (Strategic Options Development and Analysis) approach on the role of facilitators, Ackermann [12] explicitly includes a 'pre-workshop' phase in which the consultant establishes the framework of the intervention with the principal contact. The structure of the Strategic Choice approach [7] includes future 'explorations' within the concluding 'progress package' of explicit outputs. This automatically incorporates a perspective on future commitments to be made subsequent to the workshop, once those explorations bear fruit in the reduction of key uncertainties. Also, Mingers and Gill [13] include the possibility of the use of different methodologies not only between different phases of an engagement, but also across different engagements. Wong [14] makes a useful categorisation of the modes of work engaged in by PSM practitioners, namely:

- A workshop – in which the consultant(s) engage simultaneously with the complexity of subject matter, and with the complexity of interaction of the stakeholders about the subject matter.

- An interview – in which the consultant(s) engage with a single member of the group, most commonly to elicit information to structure or populate the model

- The backroom – in which information already elicited from participants either individually or collectively is processed by the consultant(s) alone in preparation for a subsequent interaction with stakeholders.

However, the rule nevertheless persists. These counter-examples broadly take the single one-time workshop as the norm. There is little attention paid to aspects of practice especially relevant to multi-workshop interventions, and to what happens in the gaps between those workshops. The particular opportunities and difficulties of single engagements that incorporate a sequence of workshops are not well addressed.

It may be surprising to those who have not taken part in one, but the amount of analytically-based work that can be achieved in a single one-day workshop is quite limited. Four one and a half hour sessions must find room for mutual introductions and acclimatisation; an introduction to the method to be used; the setting of expectations; a 'scoping' period in which participants are reassured that their particular concerns will be on the table; periodical summarising of the degree of progress made; confirmation from time to time that the progress that appears to have been made does indeed have the positive assent of all participants; and a final period in which the day's events are assessed, and subsequent actions agreed upon or confirmed.

Furthermore, a successful workshop is not 'run' by the facilitator(s). For large periods s/he is silent (though attentive) and the discourse is generated between the participants. The benefit of this in terms of 'ownership' of the process and outcomes is evident. However, there is an equally evident cost in terms of the time-economy of the event – the most effective path between two points will not be a straight line.

The implication of this is that unless there has been a great deal of preparatory work (and quite possibly if there has) it will be rather unusual for a complex set of inter-related issues to have been pursued through to effective closure in a single day's work. It is of course quite possible that sufficient clarity will have been achieved that the subsequent working out of implementation consequences can be left to more conventional, and less labour-intensive, processes. In effect, after the initial stages of problem structuring, what to do will appear 'obvious' (see [15]).

In other cases, however, it may be that the first workshop will, in effect, identify a subset of the issues which the group agrees to prioritize – but without the time to tackle that agreed problem focus adequately. There will be other situations, and the case discussed here is one of them, when the implementation questions are highly political; that is, the interests of stakeholders are likely to be differentially affected by alternative solutions to the identified question. In such cases, the continued involvement of the group of stakeholders in working out the implications of a consensual problem structure is crucial to the legitimacy of any set of proposed commitments.

There are thus a number of situations in which a single engagement will incorporate a number of workshops in sequence. Some of the features that come to the fore when this is the case have to do with the conduct of the workshops, while others concern the potentialities of the spaces between the workshops.

One feature of the first kind is the importance of achieving continuity of membership of the group. Fluidity of attendance can be consistent with continuity of representation of the stakeholders. But it is not compatible with a methodology in which later stages take as given certain assumptions and conclusions agreed at a previous meeting. The result of rotation of membership, or even of designated alternates, is a dilution of ownership of the developing problem structure. Retracing of the earlier stages with the possibility of coming to different conclusions is scarcely a practicable option, given that the majority of the group have traversed this terrain and established their own workable road-map.

Where a sequence of workshops is anticipated, therefore, the selection of committed participants is crucial. They need to be strongly advised of the expectation that they will not allow other engagements to displace their agreement to attend all the component workshops. It follows that the complete set of workshop dates needs to be established in advance. This was the procedure carried out, successfully, in the Camden and Islington study. Attendance was complete and unvarying, except in the case of one Community Health Council, and of one missed meeting by the representative of a non-central stakeholder.

The other principal requirement at all workshops except the first is to pay particular attention to the re-confirmation of the position reached at the end of the preceding workshop. There is more than one reason for doing this. The first is to re-introduce members to the conceptual world which they had been constructing and inhabiting. It will generally have been quite some time since the last meeting – to allow time for inter-workshop activities to be carried out – and memories will need re-activating. The second is that in the intervening time members of the group will have been subject to a range of influences – views of colleagues, unanticipated events – which might have caused them to revise their opinions. Finally, the strength of the conclusions from such an engagement is that of its weakest (i.e., least convinced) link. It follows that no opportunity should be lost to test out the commitment of members to the evolving problem structure. Indeed the public re-affirmation of support for that structure makes it more difficult psychologically for members later to renege from the action consequences of that structure.

It is the gaps between the workshops, however, that present the major additional opportunity for progressing the business of the engagement. The size of this opportunity will, of course, depend on the size of the gap. In one recent case, force of circumstances limited the period between two workshops to little more than the intervening weekend [16]. In general, a longer gap provides more opportunities for inter-workshop activities, but against this must be set any urgency associated with the implementation of conclusions, and also the decay of a sense of involvement and of group identity.

One advantage simply lies in the availability of more time for the consultants to reflect on how to make the activities that they are supervising more productive for the participants. Ackermann and Eden [17], in a description of a case study using the Oval Mapping Technique, repeatedly describe activities that needed to be carried out hurriedly in the interstices of the process. Catching up on material missed or not completely captured on the software, tidying up clusters of concepts, carrying out quick analyses, setting up the elicited material in a form appropriate for presentation to the group, and (especially) reviewing progress with the principle client – all these were conducted in 15-minute coffee breaks, over the lunch period, or in time snatched after the workshop before the consultants had to leave for the airport to travel home. Having to think on one's feet under pressure is undoubtedly a very concentrating experience [18]. Having additional time between workshops does not remove this invention spawned by necessity, but adds the potential for more considered views and more extensive analysis.

In the case described in this chapter, this scope was exploited in a number of ways. Certainly interaction with the 'client' (the Director of Public Health and her team) was used extensively. Other members of the group became aware of this, and there was even some sensitivity about how this selective access might be biasing the process. As one of the participants said in a followup interview,

> "....the process seemed to be reasonably clear and did seem to be based on fair principles. The one worry that occasionally went through me was whether [one of the Health Authority officers] had had pre-meetings with you, and whether in fact we were being led down a pre-laid path. And I don't know. But that was the only worries I ever had in that meeting, was just sometimes she, as an observer seemed to be further down the road than I was. And I wondered whether that was because she'd practised."

On reflection, an explicit advance statement about this aspect might have defused possible anxieties.

Typically there were two meetings with this client group between each workshop. At the first we would discuss the progress at the preceding workshop, and run through the explicit activities agreed to at the workshop to ensure that they actually happened. For the facilitators, these were tasks largely concerned with workshop process. For the Health Authority, these often related to work agreed to reduce identified uncertainties. The opportunity to reduce uncertainties during the course of the workshop sequence, rather than as part of a commitment package to be pursued as a post-workshop task, strengthened the approach.

At the second meeting we would review the new information generated, and discuss detailed plans proposed by the consultants for the structure of the coming workshop. These meetings also gave the clients an opportunity to ensure that we were adequately aware of tensions beneath the surface whose manifestations might not have been easy for us to interpret. These briefings informed both the structure that we proposed, and our handling of issues and individuals on the day.

The available time also enabled the consultants to think intensively and extensively about the way to sequence the procedures that would constitute the next workshop, and also about particular content questions that were proving intractable. An example of the latter was the development of a graphical representation of possible configurations of adolescent and tertiary provision in relation to secondary facilities for children, This proved successful at the second workshop in disentangling what had proved till then to be a disabling thicket too dense to be sorted out in mid-workshop. An example of the former was the decision to develop a mutation of the comparative advantage chart for use at the third workshop in comparing schemes with different numbers of 'SpeCCs'. It was felt by the consultants that a simplified form might be adequate, and would avoid the over-repetitive use of a single tool.

Although we did make use of the month-long intermissions to develop quite elaborate 'running orders' (including contingency plans) for the impending workshop – as we did in initial preparation before the first workshop – these were of variable utility in practice. It was always necessary to deviate from the programme at various points and to improvise as situations developed in unpredicted ways. Sometimes the workshop's path rejoined the anticipated one, and in other cases we proceeded on a different course. Devising the running order was, however, always a valuable use of inter-workshop time. Its existence re-assured the client that the effort and political commitment that they were putting into the workshops was matched by due consideration on our part. And it also ensured that the consultants had journeyed mentally down into the grain of the problem situation. This meant that we were well prepared to respond rapidly and confidently to the unexpected analytic, interpersonal, and inter-organizational challenges that the dynamic of the workshop would throw up.

22.5.2 Problematic implementation

As described above in Section 22.4.2, the workshops, although perceived by participants to be effective events in themselves, substantially failed to bring about the desired changes. Failures of OR interventions, either hard or soft, are seldom reported. (An exception is the review of failures and successes by Tilanus [19].) However, all practitioners know that interventions fail for reasons other than methodological incompetence, and successes are frequently achieved through ungeneralisable and undocumented fixes and hacks [20]. So it useful to reflect on what happened in this case, that made the outcomes so much less than the promise.

We can look at three contributory factors:

- Some key stakeholders were not present
- Some participants could not carry their constituencies
- Unforeseen circumstances

Absent stakeholders In Section 22.2.2 we described the process of deciding upon workshop membership. To realise the advantages of open interaction and engagement between members, workshop numbers need to be limited. In this case, not even all health specialties and roles in the Health Authority area could be represented. (This excluded, crucially as it turned out, representatives from a neighboring health authority.) Non-professional health interests were only represented by the CHCs as permitted intermediaries.

Inevitably, group size limitations meant that key political interests, both local and national, could have no spokespersons in the workshops. However, these interests were not totally unrepresented. Participants brought them into the discourse as comparison areas or uncertainty areas (e.g. local political acceptability and effect on children's and adolescents' mental health services) and at this remove were captured as labels on post-its. Their influence on the workshop processes was through these proxy representations, which were the results of what is described in Actor Network Theory as a series of translations and inscriptions [21, 22]. This theory explores how human and non-human actants are enrolled in a network which may or may not be stable and induce action. In this case the workshops could not stably enrol all the absent key actors in coordinated action, as instanced in the unwanted critical prominence given to the ongoing process by the local newspaper.

One strategy for avoiding this pitfall is to move towards the actual incorporation of more stakeholders into the discursive process. There is now a range of approaches designed for large group interventions, notably Open Space Technology [23], Future Search [24-26] and Team Syntegrity [27]. The principle has been described as "getting the whole system in the room". What is traded off against this inclusivity (and the legitimacy that it imparts) is the possibility of engaged conversation between all participants. The various approaches use different methods, none of them model-based, to synthesise outputs from multiple small group conversations into a large group consensus. The complementary strengths of large group intervention approaches and PSMs suggest a potential for mutual borrowing.

Failure of delegacy Although it did not affect the eventual outcome, it became evident in the immediate aftermath of the workshops that the consultant from one of the hospitals proposed to lose its inpatient department was in some difficulties in maintaining this agreed position inside the hospital. This is a not unfamiliar situation in inter-organizational uses of PSMs. Indeed the phenomenon is widespread – witness the experience through the years of both ambassadors whose negotiated accommodations are repudiated by their governments, and trade union negotiators who fail to get their wage deals endorsed by their members.

We can use a similar analysis based on Actor Network Theory to understand the process of failed delegacy, the question of "who speaks in the name of whom" (Callon [21], p. 214). Participants are involved in a sense-making process [28] which is contingent upon the composition and discourse of the workshop. That which makes sense within the workshop and appears to be a reasonable resolution of conflicting demands may not be seen as sensible when reported back to constituents outside the workshop. If the links in the chain that connect the organization to the workshop through the representative are not sufficiently strong then, in Callon's phrase, "translation becomes treason".

This problematic potential for workshop-based approaches is intensified for a linked series of workshops such as that employed in this case. Revealing work in progress is disruptive to the internal workings of the workshops (and maybe indeed be destabilising if first one option for service relocation is floated and then another); but not doing so weakens the representivity of the participants. They become less able to speak to and for their constituencies: this is the cost of becoming more embedded in the network and worldview of the workshop.

Unforeseen circumstances We have described how seemingly remote occurrences (here, the foot-and-mouth epidemic) disrupted the anticipated sequence of events. This put so much pressure on the Health Authority management, that even their representatives, who had commissioned the workshop with the full backing of their managers and were the most committed to the proposals, could not in the end carry their constituency. In Actor Network Theory terms, allies had not been locked into place and had become implicated in other networks – in this case ensuring they continued to have jobs. The attempt to make the issue of "how do we ensure a critical mass of pediatric patients" become an *Obligatory Passage Point* for all discussion, and action had failed.

Through this rudimentary analysis (which will be elaborated elsewhere) we can see how success within the workshop did not necessarily result in success outside the workshop. Within the workshop, concepts generated in the discourse became fixed as they were written onto post-its and persisted through the workshops. They became 'boundary objects' [29], which inhabit different social worlds [30] and are capable of being interpreted and applied in the different professional forms of life and understandings of the workshop members. However, they were not effective representations to people outside the workshop. Furthermore, participants were not effective as brokers or boundary spanners [31] to communicate effectively the workshop results to other audiences. (The exceptions to this were the Strategy Group, to which the workshop closely related, and initially the Chief Executives of the hospital trusts.) Successful workshops, especially in an inter-organizational field exposed to the public gaze need to be able to transverse boundaries of perception not only between participants but also between participants and wider communities.

As even this introductory account shows, Actor Network Theory provides a framework which illuminates the strengths and potential weaknesses of workshop-based approaches such as SCA. We have adumbrated both the attempted process of translation of absent actors, such as neighboring health authorities, into members of a network rooted in the workshop; and how the process centered on the workshop failed to make its participants part of a stable network which would effect change. A more detailed analysis of this and other workshops would examine how effective the rhetorical devices of SCA – shaping, designing, comparing and choosing – can be in enrolling participants in networks in which their interests are represented and where these new networks arrangements embody irreversible change. Such further analysis would, in particular by paying attention to the workshop and the wider world simultaneously, provide indications of how best to employ these devices. Such an analysis has the potential, therefore, both to respond to the criticisms that have been made of SCA that it has pragmatic effectiveness but no theoretical underpinning; and also to improve practice. Practice improvements may be looked to through the direction of facilitators' attention to the steps necessary to ensuring robust relationships between the activities within the workshop, and actors and actions external to it, thus increasing the likelihood of apparently successful workshops leading to substantive desired change [32].

ACKNOWLEDGEMENTS

The authors would like to acknowledge the contributions made by all members of the workshop, in particular: Dr. Maggie Barker, Sue Dutch and their colleagues at Camden and Islington Health Authority: those who participated in post-workshop interviews; John Friend; and Diane Plamping.

References

[1] Barker, M. (1999). *The Health of Children and Young People in Camden and Islington: Public Health Report 1999.* Camden and Islington Health Authority, London.

[2] Camden and Islington Health Authority and NHS Partners in Camden and Islington (2000). *Improving Health Services for Children and Young People in Camden & Islington.* Camden and Islington Health Authority, London.

[3] Barker, M (2002). Interview, May 14.

[4] Greenberger, M., M.A Crenson and B.L. Crissey (1976). *Models in the Policy Process.* Russell Sage, New York.

[5] Friend, J. and A. Hickling (1997). *Planning under Pressure: The Strategic Choice Approach* (2nd edition). Butterworth-Heinemann, Oxford, UK.

[6] Rosenhead, J. (2001). Robustness analysis: Keeping your options open. In Rosenhead, J. and J. Mingers, Eds., *Rational Analysis for a Problematic World Revisited: Problem Structuring Methods for Complexity, Uncertainty and Conflict.* Wiley, Chichester, UK, 181-207.

[7] Friend, J. (2001). The Strategic Choice Approach. In Rosenhead, J. and J. Mingers, Eds., *Rational Analysis for a Problematic World Revisited: Problem Structuring Methodsfor Complexity, Uncertainty and Conflict.* Wiley, Chichester, UK, 181-207.

[8] Holt, J. (1994). Disarming defences. *OR Insight,* 7, 19-26.

[9] Phillips, L. and M. Phillips (1993). Facilitated work groups: Theory and practice. *Journal of the Operational Research Society,* 44, 533-549.

[10] Department of Health (2001). *Milburn Hands Power to Front-Line Staff: £100m savings for patient services.* http://www.info.doh.gov.uk/doh/IntPress.nsf/page/2001-0200?Open Document. Department of Health Press Release 2001/0200, Accessed March 24, 2003.

[11] Eden, C. and J. Radford, Eds. (1990). *Tackling Strategic Problems: The Role ofGroup Decision Support.* Sage Publishers, London.

[12] Ackermann, F. (1996). Participants' perceptions on the role of facilitators using Group Decision Support Systems. *Group Decision and Negotiation,* 5, 93-112.

[13] Mingers, J. and A. Gill (1997). *Multimethodology: The Theory and Practice of Combining Management Science Methodologies.* Wiley, Chichester, UK.

[14] Wong, H.-Y. (1998). *Making Flexible Planning Decisions: Clarification and Elaboration of the Theory and Methodology of Robustness Analysis.* PhD thesis, London University, London.

[15] Eden, C. (1987). Problem-solving or problem-finishing? In Jackson, M.C. and P. Keys, Eds., *New Directions in Management Science,* Gower, Aldershot, UK.

[16] Horlick-Jones, T., J. Rosenhead, I. Georgiou, J. Ravetz, and R. Lofsted (2001). Decision support for organisational risk management by problem structuring. *Health, Risk and Society,* 3, 141-165.

[17] Ackermann, F. and C. Eden (2001). SODA – Journey making and mapping in practice. In Rosenhead, J. and J. Mingers, Eds., *Rational Analysis for a Problematic World Revisited: Problem Structuring Methods for Complexity, Uncertainty and Conflict,* Wiley, Chichester, UK, 43-60.

[18] Chapman, C., D.F. Cooper, C.A. Debelius, and A.G. Pecora (1985). Problem solving methodology design on the run. *Journal of the Operational Research Society,* 36, 769-778.

[19] Tilanus, C.B. (1985). Failures and successes of quantitative methods in management. *European Journal of Operational Research,* 19, 170-175.

[20] Ciborra, C. (2002). *The Labyrinths of Information: Challenging the Wisdom of Systems.* Oxford University Press, Oxford, UK.

[21] Gallon, M. (1986). Some elements of a sociology of translation: Domestication of the scallops and the fishermen of St-Brieuc Bay. In J. Law, Ed., *Power, Action and Belief: A New Sociology of Knowledge.* Routledge and Kegan Paul, London, 196-233.

[22] Latour, B. (1987). *Science in Action: How to Follow Scientists and Engineers Through Ssociety* (translated by C. Porter). Harvard University Press, Cambridge, MA.

[23] Owen, H. (1992). *Open Space Technology: A User's Guide.* Abbott, Potomac, MD.

[24] Weisbord, M.R. (1987). *Productive Workplaces: Organising and Managing for Dignity, Meaning and Community.* Jossey-Bass, San Francisco, CA.

[25] Weisbord, M.R. (1987). *Discovering Common Ground.* Berrett-Koehler, San Francisco, CA.

[26] Weisbord, M.R. and S. Janoff (1995). *Future Search.* Berrett-Koehler, San Francisco, CA.

[27] Beer, S. (1994). *Beyond Dispute: The Invention of Team Syntegrity.* Wiley, Chichester, UK.

[28] Weick, K.E. (1995). *Sensemaking in Organizations,* Sage Publications, Thousand Oaks, CA.

[29] Star, S.L. and R.J. Griesemer (1989). Institutional ecology, 'translations', and boundary objects: amateurs and professionals in Berkeley's Museum of Vertebrae Zoology, 1907-39. *Social Studies of Science,* 19, 384-420.

[30] Bowker, G.C. and S.L. Star (1999). *Sorting Things Out: Classification and Its Consequences.* MIT Press, Cambridge, MA.

[31] Wenger, E. (2000). Communities of practice and social learning systems. *Organization,* 7 , 225-246.

[32] Jackson, M. (1991). Review *of Rational Analysis for a Problematic World* Edited by J. Rosenhead. *Systems Practice,* 4 , 258-290.

23 MODELING MEDICAL TREATMENT USING MARKOV DECISION PROCESSES

Andrew J. Schaefer[1,2,3], Matthew D. Bailey[1],
Steven M. Shechter[1] and Mark S. Roberts[2,3]

[1] Department of Industrial Engineering
University of Pittsburgh
Pittsburgh, PA 15261

[2] Department of Medicine
University of Pittsburgh
Pittsburgh, PA 15261

[3] Center for Research on Health Care
University of Pittsburgh
Pittsburgh, PA 15261

SUMMARY

Medical treatment decisions are often sequential and uncertain. Markov decision processes (MDPs) are an appropriate technique for modeling and solving such stochastic and dynamic decisions. This chapter gives an overview of MDP models and solution techniques. We describe MDP modeling in the context of medical treatment and discuss when MDPs are an appropriate technique. We review selected successful applications of MDPs to treatment decisions in the literature. We conclude with a discussion of the challenges and opportunities for applying MDPs to medical treatment decisions.

KEY WORDS

Markov decision processes, Stochastic dynamic programs, Optimal medical treatment, Stochastic optimal control, Medical decision making

23.1 INTRODUCTION

Medical treatment decisions must be made sequentially and in an uncertain environment. A physician determining a course of treatment must consider the patient's current health, as well as the best treatment decisions in the future. One important source of uncertainty is that different patients will respond to treatments differently. Other sources of uncertainty include the availability of scarce resources, such as cadaveric organs for transplantation, and human behavior, such as the response time for individuals to react to stroke symptoms. In current medical practice, the vast majority of these treatment decisions are made using ad hoc or heuristic strategies. However, there is a growing feeling among medical practitioners that some treatment decisions are too complicated to solve accurately using intuition alone [1,2]. The evidence for this includes psychological experiments that indicate that short-term memory has a limited capacity to handle multiple memory constructs, and a substantial body of evidence suggesting a large variation in clinical practice [1, 3-5].

Physicians will always need to make subjective judgments about treatment strategies. However, mathematical decision models that provide insight into the nature of optimal decisions can aid treatment decisions. *Markov decision processes (MDPs)* (also known as stochastic dynamic programs) are an appropriate and under-utilized technique for certain types of treatment decisions. MDPs find optimal solutions to sequential and stochastic decision problems. The major advantage of MDPs is their flexibility. Although virtually every medical decision can be modeled as an MDP, the technique is most useful in classes of problems involving complex, stochastic and dynamic decisions, for which MDPs can find optimal solutions.

An MDP is similar to a Markov process (or Markov model, as it is known in the medical decision making literature), except that the decision maker must make decisions at various time epochs. The goal of an MDP is to provide an optimal *policy,* which is a decision strategy to optimize a particular criterion such as maximizing a total discounted reward. In this way, MDPs differ from other stochastic modeling techniques such as discrete-event simulation or Markov processes. Such techniques may be used to evaluate the consequences of a fully specified stochastic model, but they do not allow for the stochastic optimization of that model; they evaluate just one particular policy at a time. To evaluate exhaustively every feasible policy in this manner may be computationally prohibitive. MDPs not only provide the consequences of a policy, they guarantee that no better policy exists.

MDPs also have drawbacks. As the size of the problem increases, MDPs become harder to solve exactly. However, many techniques for finding approximate solutions to MDPs exist. This has been a fertile research area recently, but not in the context of medical treatment decisions [6, 7]. Perhaps the biggest hindrance to the broader application of MDPs is data. Obtaining quality medical data is very difficult and expensive. It is common for a large medical study to cost several million dollars. MDPs are even more data-intensive than other stochastic modeling techniques. This is because the transition probabilities governing the stochastic process, as well as the rewards, are permitted to vary according to the decision made at each decision epoch. While this flexibility is a large advantage in treatment decisions, it means that for every possible description of patient health and every possible treatment, an MDP requires enough observations to estimate accurately transition probabilities to the next epoch. In practice, this typically means that quality data covering thousands of patients is necessary for a successful and realistic MDP model. Although the use of such large patient series is not common, the increasing use of electronic medical records systems is enhancing researchers' ability to utilize large amounts of clinical data from thousands of patients [8].

In Section 23.2 we provide formal models of MDPs and discuss implementation issues such as algorithms and efficiency issues. In Section 23.3 we consider modeling issues particular to applying MDPs to health care problems. In Section 23.4 we provide a selective literature review of previous successful applications of MDPs to medical treatment problems. For each article, we describe the medical application, modeling issues and the solution technique. Finally, in Section 23.5 we provide some conclusions and discuss the future of applying MDPs to medical treatment problems.

23.2 FUNDAMENTALS OF MDP METHODOLOGY

Markov decision processes, or stochastic dynamic programs, are a general framework for modeling dynamic systems under uncertainty. Under mild separability assumptions, discrete-time MDPs can be applied to a variety of systems where decisions are made sequentially to optimize a stated performance criterion. An MDP binds previous, current, and future system decisions through the proper definition of system states, defined as variables that contain the relevant information for making future decisions. The system model evolves in the following manner: The condition or state of the system is observed (or partially observed), an action is taken, a reward is received (or cost incurred), and the system transitions to a new state according to a known probability distribution. The state variables must be defined so that given the current state of the system the future transitions and rewards are independent of the past. This is the standard assumption of a Markov process. MDPs are typically used to model dynamic systems; therefore the decisions are assumed to occur sequentially. However, static decisions can also be modeled using MDPs when the problem's decisions or reward structure are separable: then a one-time decision can be optimized by decomposing it into a sequence of sub-decisions.

23.2.1 Finite-horizon MDPs

We now introduce the fundamentals of MDP methodology. For more complete coverage we refer the reader to Puterman, Bertsekas, or Bellman [9-11]. Following the notation of Puterman, the basic model of a finite-horizon, discrete-time MDP is defined by $(S, A, p_t(\cdot|\cdot,\cdot), r_t(\cdot,\cdot), N)$, where S is the set of defined states and for every state $s \in S$, A is the set of all feasible actions or decisions and A_s are those actions available at state s. The system progresses to state s' from state s when action $a \in A_s$ is chosen at decision epoch t, $(t = 1,\ldots,N)$, with known probability transition $p_t(s' \mid s,a)$. When action $a \in A_s$ is chosen from state s at decision epoch t, a reward $r_t(s,a)$ is received. We define a policy $\pi = \{d_1, d_2, \ldots, d_{N-1}\}$ as a sequence of decision rules, where a decision rule is a mapping from states to actions, so that $d_t(s) \in A_s$. The application of a policy π induces a probability distribution over the states at various stages, where the state of the system after t transitions is X_t and the action chosen, Y_t, is a function of this state. The objective is to compute the policy that maximizes a given criterion in expectation.

Three commonly used criteria (when beginning in state s) are: the total expected reward,

$$v_N^{\pi}(s) = E_s^{\pi}\left\{\sum_{t=1}^{N-1} r_t(X_t,Y_t) + r_N(X_N)\right\};$$

the total discounted expected reward,

$$v_N^{\pi}(s) = E_s^{\pi}\left\{\sum_{t=1}^{N-1} \lambda^{t-1} r_t(X_t,Y_t) + \lambda^{N-1} r_N(X_N)\right\},$$

for $0 \leq \lambda < 1$; and the average reward per stage,

$$v_N^{\pi}(s) = E_s^{\pi}\left\{\frac{1}{N}\left(\sum_{t=1}^{N-1} r_t(X_t,Y_t) + r_N(X_N)\right)\right\}.$$

For a finite N, the optimal policy for both the average reward per stage and the total reward criterion are equivalent. For the infinite-horizon case, which will be discussed shortly, there is a distinction.

We will present the fundamentals of the total discounted expected reward criterion. Under the standard assumptions of a basic MDP model, where S and A are finite and the rewards are bounded, i.e. $|r_t(s,a)| \leq M < \infty$ for every state-action pair (s, a), and $t \leq N$, then $v_N^{\pi}(s)$ exists and is bounded. We seek a policy π^* such that $v_N^{\pi}(s) \leq v_N^{\pi^*}(s)$ for every $s \in S$. As a result of the principal of optimality [11], the separability of the MDP decisions and rewards can be exploited to decompose this N-period problem into a sequence of $N - 1$ single-stage problems, by recursively solving backward from stage $N - 1$ to 1:

$v_N(s) = r_N(s)$ for every $s \in S$, and

$$v_t(s) = \max_{a \in A_s} \left\{ r_t(s,a) + \sum_{s' \in S} \lambda p(s' \mid s,a) v_{t+1}(s') \right\}$$

$$\text{for } t = N-1, \ldots, 1 \text{ and } s \in S.$$

Here $v_t(s)$ is the total discounted-expected reward of the $N - t$ stage problem beginning in state s at stage t or as a single-stage problem with terminal rewards $v_{t+1}(s')$, which are known at the time $v_t(s)$ is computed. This is the true computational benefit of MDPs, the ability to reduce a problem into manageable subproblems and still attain the optimal solution. The optimal policy is defined to be the sequence of decision rules, mapping states to the actions that maximize the above recursion, i.e.

$$d_t(s) = \arg\max_{a \in A_s} \left\{ r_t(s,a) + \sum_{s' \in S} \lambda p(s' \mid s,a) v_{t+1}(s') \right\}$$

$$\text{for } t = N-1, \ldots, 1 \text{ and } s \in S.$$

In the above solution and model we assumed that the decision horizon was a finite N. Often there is no defined horizon or the number of stages is so large that it may be approximated by an infinite horizon. In these instances we utilize the techniques discussed in the next section.

23.2.2 Infinite-horizon MDPs

Infinite-horizon models require an infinite amount of data. Therefore, it is typically assumed that data are time-homogeneous or changing so slowly that homogeneity is a reasonable assumption. As a result, the state of an infinite-horizon MDP must be carefully defined to ensure that the system transitions are stationary. If the data are naturally time-dependent, the time-homogeneity assumption can be satisfied by properly augmenting the state definition with the time at which a system transition occurs. The presence of stationary system transitions allows for the use of several elegant solution techniques and easily characterized optimal policies. We replace the above finite-horizon criteria with an infinite-horizon variant by taking the limit of each measure as N goes to infinity. Unlike the total expected reward and discounted expected reward criteria, the analysis and solution methodologies for the average reward criterion depend on the structure of the underlying Markov processes [12]. Again we focus on the problem of maximizing a stream of discounted rewards, which is assured to converge as a result of the bounded rewards assumption.

One of the key insights into infinite-horizon MDPs is that as a result of the assumptions of an infinite horizon, time-homogeneity, and Markov property, under a stationary policy π, i.e. $d_t(s) = d(s)$ for all $s \in S$ and $t = 1,2,\ldots$, the expected reward vector is also stationary

$$v_t^{\pi}(s) = v^{\pi}(s) \text{ for } t = 1,2,\ldots \text{ and } s \in S\,,$$

and $v^{\pi}(s)$ is the unique solution to the set of equations:

$$v^{\pi}(s) = r\,(s,d(s)) + \sum_{s' \in S} \lambda p(s' \mid s,d(s))\, v^{\pi}(s') \text{ for } s \in S\,. \qquad (1)$$

It is well known that a stationary policy is optimal for these MDPs. In addition, the optimal vector v^* is the solution of the following equations, known as Bellman's equations:

$$v^*(s) = \max_{a \in A_s} \left\{ r\,(s,a) + \sum_{s' \in S} \lambda p(s' \mid s,a)\, v^*(s') \right\} \text{ for } s \in S\,.$$

Given any initial bounded vector v_0, it can be shown that the following sequence converges to a solution of Bellman's equations:

$$v_k(s) = \max_{a \in A_s} \left\{ r\,(s,a) + \sum_{s' \in S} \lambda p(s' \mid s,a)\, v_{k-1}^*(s') \right\} \text{ for } s \in S \text{ and } k = 1, 2, \ldots. \qquad (2)$$

However, this solution procedure, known as value iteration, may require an infinite number of iterations [13]. As a result, another technique, policy iteration, is typically used to search over the finite space of policies [11, 14].

In policy iteration, we begin with a policy π^0, evaluate that policy by solving the set of linear equations in (1) to find v^{π^0}, use this value to choose the actions that maximize the equations in (2) to perform a policy improvement step, and determine the next policy π^1. This process is continued until identical policies are found in subsequent iterations. Each iteration results in a policy with an improved optimal reward vector and therefore, for an MDP with finite state and action spaces, policy iteration will terminate with the optimal policy in a finite number of steps. There are several variants of the above techniques; however, the most successful solution methodologies will typically exploit the natural structure of a particular problem instance.

23.2.3 Partially observed MDPs

The above finite- and infinite-horizon MDPs fall into a broader class of MDPs that assume perfect state information – in other words, an exact description of the system. However, often such precision is either too strong an assumption or is not plausible within the model. For example, the state of an MDP could be results from a series of medical tests. These results may supply a better idea of the true state of the patient, but are subject to the error of the tests. Extensions of MDPs, called partially observed Markov decision processes (POMDPs), have been developed to deal with imperfect information [15, 16]. In these models it is assumed that uncertainty exists in the transitions of the system itself and in our knowledge of which state the system truly occupies. Therefore, the objective is to find an optimal policy based on the observations of the system and the previous decision rules applied. It is possible to replace the partially observed state with a sufficient statistic that can be interpreted as a likelihood estimation of the true state of the system given the observations seen. In this manner, the model can be transformed to one with perfect information using the sufficient statistic as the state definition [17]. However, this conversion results in computationally intractable models for systems with even moderately sized underlying true state spaces. As a result, heuristics or approximation techniques must be employed to effectively generate solutions to realistic problem instances.

23.2.4 Semi-Markov decision processes

The above discussion focused on models where the time between decision epochs is fixed and has no effect on the rewards of the system. However, in health care and other applications, decisions may occur over continuous time intervals, such as when varying treatments can be administered. The time between these transitions may depend on the action selected or may occur randomly. In these instances, an extension of MDPs called semi-Markov decision processes (SMDPs) can be employed. These models allow system transitions to occur in continuous time and allow for the inclusion of a probability distribution over the amount of time spent in a state. Through problem transformations and redefinitions, techniques and solution algorithms analogous to those of discrete-time MDPs have been developed for this class of problems [18, 19].

23.3 MODELING ISSUES

23.3.1 Benefit of MDP modeling over traditional decision modeling in health care

For simple medical treatment decisions, a decision tree can be utilized to discover the best course of action. A terminal node of a decision tree usually represents the expected utility (such as life expectancy or quality-adjusted life years) of a patient whose health progression follows that branch of the tree. The path to that terminal node may be complex, and the calculation of that value, requires knowing how the patient may transition between various health states from the initial decision point until death. Modeling these transitions in a standard tree requires a large number of nodes representing multiple time periods in the model, resulting in a tree explosion [20]: the situation is even more complex if the decision can be made at various times, which requires the use of *embedded* decision nodes, making the analysis and interpretation of standard trees almost impossible. As the complexity of the problem increases, the standard decision tree becomes impractical.

Markov models are popular in medical decision making because they can handle some of the difficulty described above. They allow for a simpler representation of the future states and possible transitions that may occur until the patient dies. Solutions to Markov models are obtained via matrix algebra, cohort simulations, or Monte Carlo simulations. Markov models have their limitations, however, because they are not well suited to handle the situation in which decisions may be made at multiple time points. This deficiency of traditional Markov models is precisely the advantage of using Markov *decision* processes for treatment decisions.

Rather than evaluating a decision tree based on a one-time decision (as is often the case in traditional decision trees and Markov models), MDPs allow the “do-nothing” option in each time period and consider the “do-something” option at any later decision epoch [21]. For example, organ transplantation can be modeled as an MDP in which the action each time a donor organ becomes available is to either accept the organ or reject it and wait for a better one. The MDP methodology is especially beneficial because it offers the flexibility of choosing possibly different actions across multiple time periods according to the patient’s state. For example, a doctor treating an

HIV-infected patient using highly active anti-retroviral therapy (HAART) may consider different doses and different combinations of drugs at different times during the course of treatment. The action chosen depends on the patient's state, which could include side effects, level of CD4 cells and viral RNA, signs of drug resistance, and degree of adherence to the regimen. Just about any situation where one wants to optimize a process over multiple time periods can be modeled using an MDP. As discussed above, though, exact solutions for large-scale problems may be computationally infeasible and one may need to resort to approximate heuristics.

23.3.2 Issues in modeling disease treatment decisions

Many MDP applications in health care must address the same important modeling issues. For example, MDPs that attempt to optimize a treatment plan or surgery time for a disease require a model of how a patient's health evolves both before and after an intervention. In the case of the optimal time to transplant a liver from a living donor, it is important to develop both a good natural history model of how a patient's health changes in the absence of a transplant and a post-transplant survival model that determines when a patient dies. The natural history model is used to determine transition probabilities between health states from one period to the next if the patient chooses to wait another day for the transplant. In MDP terminology, the survival model determines a terminal reward – the expected remaining life of the patient after receiving a new liver – when the transplant action is chosen.

Another modeling issue in health care MDPs is determining the rewards associated with actions. Optimal disease treatments are usually concerned with maximizing both total life years and quality of life. The quality-adjusted life year (QALY) is a popular measure in the medical literature that blends these two goals [22, 23]. This approach considers a patient's utility for various health states and multiplies the length of life under these health states by the utility weight. One can assess these utilities in various ways including the standard gamble, the time-tradeoff, and the visual rating scale [24]. When quality adjustment is used, the decision to wait another day for treatment or surgery can have very different payoffs for different patients. As Ahn and Hornberger note, for example, some kidney patients may not mind dialysis as much as others and hence would be willing to wait longer for a better donor match [25].

An important area of research in medical treatment decisions concerns the correct way to discount future health consequences. A ubiquitous model to handle this is the discounted-utility (DU) model in which the same discount rate (appropriately compounded) is applied to all future outcomes [26]. In this way, outcomes that occur earlier are preferred to equally valued outcomes that occur later. Over the last couple of decades, however, there has been much research questioning the normative aspects of the DU model [27]. Some of this research demonstrates preference reversals as the time until an event draws nearer, which is inconsistent with DU theory. For example, one study showed that one month before birth, many women wanted to avoid using anesthesia, but during labor they often changed their mind and preferred the anesthesia [28]. Such reversals can be handled by an alternative discounting model – hyperbolic discounting [29]. Other observed phenomena that are inconsistent with traditional DU models include sign effects (where gains are discounted more than losses), magnitude effects (where small outcomes are discounted more than larger ones), and preferences for improving sequences over worsening sequences [27].

A common and recommended practice in cost-effectiveness analyses is to use the same discount rate for both monetary and health outcomes [30]. However, people usually do not discount these two types of outcomes in the same way [31]. Rather, people often demonstrate higher discount rates for health than for money, and, moreover, do not demonstrate a correlation between discount rates in these areas [31]. This suggests that we must pay careful attention to the valuation and discounting of outcomes in an MDP.

23.3.3 Appropriateness of MDPs

Under mild assumptions about the reward functions, any discrete-time sequential decision under uncertainty can be modeled as an MDP. However, data limitations and computational effort may impose limits on one's ability to solve large-scale MDP models in health care. MDP models differ from other models used for treatment decisions. A discrete-event simulation estimates the behavior of a system under uncertainty but is generally unable to make optimal decisions within the simulation. An exception is optimization via simulation, in which parameters governing the simulation are optimized by estimating gradients [32]. In contrast, an MDP allows decisions to be embedded within a Markov process. Rather than an estimate of system behavior, an MDP implicitly considers all possible decision rules or policies and produces the one that behaves the best under a given optimality criteria.

23.3.4 State definition

Selecting the appropriate level of descriptive detail contained in the states of an MDP model is extremely important. From a modeling perspective, the more detailed the information contained in the states the better, since this detail provides a greater distinction among patients. However, increasing the state space makes the model more difficult to solve. Furthermore, data limitations may make a large state space undesirable. For instance, there may be state-action pairs (s,a) for which few or no clinical observations occurred. This is typically the case in health care models. States can either be functions of physiological measures (e.g. laboratory values, heart rate, CD4 counts) or can be defined based on subjective judgments such as survival probability.

When insufficient data exist to derive a transition probability distribution or estimated rewards for a set of state-action pairs, two main modeling approaches can be used. One method is to aggregate states judiciously and/or actions to accumulate enough observations for sufficient estimates. For this approach it is important that the aggregated states and/or actions can be justified clinically, since the model cannot distinguish among different patients in the same state. The other approach is to use empirical models of clinical phenomena to estimate the effect of one state-action pair by considering similar state-action pairs for which sufficient data exist. For instance, a statistical model such as a regression model might be able to estimate the effects of a particular state-action pair by considering the results of all states with the same action. This approach may be more successful in estimating rewards than transition probabilities.

23.4 APPLICATIONS OF MDPs TO MEDICAL TREATMENT DECISIONS

We summarize previous successful applications of MDPs to medical treatment decisions. Despite the appropriateness of MDPs for medical treatment decisions, the fact that relatively few such applications exist illustrates the difficulties in developing successful applications.

Epidemic Control Lefèvre developed a continuous-time MDP formulation to address the problem of controlling an epidemic in a closed population of N people [33]. The state of the system was described by the number of people infected, and the rest of the population was considered susceptible. Transition probabilities depended on the rate of infection from some external causes, the internal rate of disease transfer from those infected to the uninfected, and the rate at which the infected recovered from the disease. At any point in time, the decision-maker could choose two parameter levels: 1) the amount of the population to quarantine, and 2) the amount of medical treatment to apply to the infected population. Utilizing these definitions, the model minimized the total expected discounted cost over an infinite horizon where the costs incorporated the social cost of people being infected, the cost of quarantining, and the cost of administering medical treatment to those infected. Rather than use real data to solve an instance of the problem, Lefèvre developed the structure of the optimal policy according to the form of the various input parameters. In order to do this, he used a technique that allows one to convert a continuous-time MDP into an equivalent discrete-time MDP [19, 34].

Drug Infusion Hu et al. considered the problem of choosing an appropriate drug infusion plan for the administration of anesthesia [35]. The main decision in this problem was the level at which to set the drug infusion rate to reach a target concentration. Too much anesthesia can cause problems with blood pressure, heart rate, or recovery from the anesthetic state, but too little anesthesia can make the patient more aware of the painful operation. They modeled the problem as a POMDP, which in its pure form was computationally unsolvable. Fast heuristics were necessary for this problem since the maintenance of drug concentrations at target levels is very time sensitive.

One of the main difficulties in this problem arose from the inability to directly observe patient parameters such as anesthesia concentration in the blood and the clearance rate of the drug. This lead to two main issues in the model: 1) the best way to estimate the prior and posterior distributions for these parameters (i.e., whether to use a continuous or discrete distribution), and 2) how much to emphasize active versus passive gathering of information (i.e., how much cost should be incurred now to obtain useful information that can be used more effectively later). The authors developed their own discretization technique for estimating the parameter distribution. This technique has most of the advantages of using continuous and discrete distributions without incurring high computational costs. They applied six approximation methods to determine suboptimal though useful treatment strategies. Three of these treatment strategies emphasized active gathering of information, and the other three strategies emphasized passive gathering. Based on their results, they planned on implementing one of the passive gathering policies into the STANPUMP program at Stanford Medical School, which administers intravenous anesthetics.

Kidney Transplantation Ahn and Hornberger described a model of kidney transplantation that allowed patients to accept or reject an offered kidney based on the quality of the organ [25]. For a potential kidney, they estimated the one-year graft survival of that kidney in a certain patient. For that patient, they also determined the one-year graft survival acceptance threshold that maximized his or her quality-adjusted life expectancy (QALE). The QALE was based on patient-specific ratings for being in different health states. Rather than solve the problem explicitly as an MDP, the authors restricted their search to threshold policies, thereby reducing the problem to finding the optimal threshold level. If the expected one-year graft survival for the kidney-patient pair exceeded the threshold, the patient accepted the transplant; otherwise the patient rejected it. Their model was further simplified by having just five states: 1) alive on dialysis waiting for a transplant, 2) not eligible for transplantation, 3) received a functioning renal transplant, 4) transplant failed, and 5) death. They assumed that patients transitioned between the different states according to a Markov chain with probabilities based on published graft and patient survival rates in the United States.

Spherocytosis Treatment Magni et al. used an MDP approach to decide on therapy for mild hereditary spherocytosis, a disease that causes the chronic destruction of red blood cells [21]. For patients with a mild form of this disease, the main medical treatments considered were prophylactic splenectomy and/or cholecystectomy or no surgery at all. The state of the patient was described through the severity of gallstones and the presence of or years since removal of the spleen. The authors considered gallstone natural history, risk of surgical mortality, and natural causes of death in deriving transition probabilities. They estimated these probabilities and quality-of-life utilities based on published mortality tables and previous studies. They assumed that decisions were made every year with the overall objective of maximizing the patient's quality-adjusted life years. The optimal solution to the MDP model resulted in the following strategy: If a six-year old patient does not have gallstones, then as long as she does not develop gallstones, wait until she is fifteen and then perform splenectomy surgery. If gallstones do appear before the age of fifteen, then both cholecystectomy and splenectomy are suggested.

Treatment of Ischemic Heart Disease Hauskrecht and Fraser applied a POMDP formulation to the problem of treating patients with ischemic heart disease (IHD) [36]. IHD results from the heart not receiving adequate oxygen and is usually caused when the coronary arteries narrow. For patients with this disease, physicians must choose among various diagnostic procedures (such as an angiogram or one of many varieties of stress test), which may be followed by a therapeutic intervention such as medication, surgery (such as angioplasty or bypass surgery), or nothing at all. The state of the patient was described by a variety of variables including the level of coronary heart disease, ischemia level, history of coronary artery bypass grafting, history of percutanerous transluminal coronary angioplasty, and stress test results. The uncertainty of the patient health state arises from the inability to know exactly the level of coronary artery occlusion or the homodynamic impact of that occlusion on myocardial ischemia. Some variables, such as level of chest pain, are directly observable.

Hauskrecht and Fraser framed their POMDP as an infinite-horizon discounted model that seeks a treatment strategy that minimizes total lifetime costs (where the costs incorporate duration of life, quality of life, and monetary costs). To solve their model they used heuristic procedures along with methods that take advantage of special problem structure. They validated their model by devising treatments for ten case patients and then having a cardiologist evaluate their model's treatment strategy. Almost all of the model's recommendations were deemed clinically reasonable, though the experiment also revealed areas for model improvement. Overall, their POMDP formulation was very effective and efficient in generating good treatment strategies for IHD.

Breast Cancer Screening and Treatment Ivy used a POMDP to develop a cost benefit analysis of mammogram frequency and treatment options for breast cancer [37]. The goal was to minimize the total expected cost over a patient's lifetime, where costs were based on the patient's condition, exams, and treatment options. The model consisted of three states: no disease, non-invasive breast cancer, and invasive breast cancer. It was assumed that all patients started in the no-disease state, transitioned to the non-invasive state after a random number of years (according to a geometric distribution based on age) and then transitioned to the invasive stage after another random number of years (the model was flexible enough to relax the assumptions that all non-invasive cancers became invasive or that one must enter the non-invasive state before reaching the invasive state). The part of the model that was partially observable was the patient's condition. Two types of exams – clinical breast exams (CBE) and mammograms – could be performed to get information about the patient's state. At the beginning of each time period, the decision-maker must choose whether to perform a CBE alone or a CBE with a mammogram. If a mammogram was performed and the results were abnormal then the decision-maker could choose either a lumpectomy or a mastectomy. If the mammogram was normal the decision-maker could choose to cease treatment. Using estimates from the literature on costs, test specificity, test sensitivity, and disease progression rates, Ivy solved the dynamic program and characterized optimal decision regions based on the perceived probabilities of the different states of breast cancer.

Liver Transplantation Alagoz et al. presented an MDP model for deciding the optimal time to perform a living-donor liver transplantation [38]. In these types of transplants, the friend or relative of a patient agrees to donate a portion of her liver, and the livers of both the donor and the patient regenerate to a normal size. The goal of the model was to determine when to perform the surgery in order to maximize the expected life years of the patient. The model considered the daily decision of whether or not to transplant. If a transplant was performed, the reward was the expected remaining life years post-transplant, and this was based on survival-analysis estimates [39]. If no transplant was performed, then the patient died in the next day with some probability or transitioned to another health state and increased her life by one day. The transition to other health states was governed by a natural history model of pre-transplant survival [39]. Alagoz et al. used the policy iteration algorithm to solve the MDP and generated an optimal stationary policy to transplant or wait at least another day as a function of the liver quality and the patient health at the start of the day [38].

23.5 CONCLUSIONS

MDPs are a powerful and appropriate technique for medical treatment decisions. MDPs provide optimal policies to stochastic and dynamic decisions. Examples of such decisions naturally arise in finding optimal disease treatment plans. Despite a wealth of potential applications, there have been very few successful applications of MDPs in the medical arena. This is due to several factors, particularly heavy data requirements and computational limitations. However, several recent trends appear to help ameliorate these limitations. First, the medical community is rapidly developing a more quantitative understanding of disease progression and the effects of treatment options. Additionally, the operations research/management science community is improving the solution methodology for MDPs, particularly approximate solutions of MDPs. Also, computing capacity continues to become cheaper. Finally, more hospitals are using electronic medical record systems to gather large amounts of patient data. This confluence of factors will open the door for the increased application of MDPs to medical treatment problems.

Acknowledgments

This work was supported in part by National Science Foundation grant DMI-0223084.

References

[1] Morris, A.H. (2000). Developing and implementing computerized protocols for standardization of clinical decisions. *Annals of Internal Medicine,* 132, 373-83.

[2] Tversky, A. and D. Kahneman (1982). Availability: a heuristic for judging frequency and probability. In *Judgment Under Uncertainty: Heuristics and Biases,* D. Kahneman, P. Slovic and A. Tversky, (Eds.), Cambridge University Press, New York.

[3] Pilote, L., R.M. Califf, S. Sapp, D.P. Miller, D.B. Mark, W.D. Weaver, J.M. Gore, P.W. Armstrong, E.M. Ohman and E.J. Topol for the GUSTO-1 Investigators (1995). Regional variation across the United States in the management of acute myocardial infarction. *New England Journal of Medicine,* 333, 565-572.

[4] Nattinger, A.B., M.S. Gottlieb, J. Veum, D. Yahnke and J.S. Goodwin (1992). Geographic variation in the use of breast-conserving treatment for breast cancer. *New England Journal of Medicine,* 326,1102-7.

[5] Wennberg, J. and A. Gittelsohn (1973). Small area variations in health care delivery. *Science,* 182, 1102-1108.

[6] Van Roy, B. (2002). Neuro-dynamic programming: Overview and recent trends. In *Handbook of Markov Decision Processes: Methods and Applications,* E. Feinberg and A. Schwartz, (Eds.), Kluwer Academic Press, Boston, MA.

[7] de Farias, D.P. and B. Van Roy (2003). The linear programming approach to approximate dynamic programming. *Operations Research* 51, 850-856.

[8] Tierney, W.M., J.M. Overhage and C.J. McDonald (1995). Toward electronic medical records that improve care. *Annals of Internal Medicine,* 122, 725-726.

[9] Puterman, M.L. (1994). *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons, New York.

[10] Bertsekas, D.P. (2001). *Dynamic Programming and Optimal Control.* Athena Scientific Press, Belmont, MA.

[11] Bellman, R.E. (1957). *Dynamic Programming.* Princeton University Press, Princeton, NJ.

[12] Arapostathis, A., V. Borkar, E. Fernandez-Gaucherand, M.K. Ghosh and S.I. Marcus (1993). Discrete-time controlled Markov processes with average cost criterion: a survey. *SIAM Journal on Control and Optimization,* 31, 282-344.

[13] Shapley, L.S. (1953). Stochastic games. *Proceedings of the National Academy of Sciences of the United States of America,* 39, 1095-1100.

[14] Howard, R.A. (1960). *Dynamic Programming and Markov Processes.* Technology Press of Massachusetts Institute of Technology, Cambridge, MA.

[15] Lovejoy, W.S. (1991). A survey of algorithmic methods for partially observed Markov decision problems. *Annals of Operations Research,* 28, 47-66.

[16] White, C.C. and W.T. Scherer (1989). Solution procedures for partially observed Markov decision processes. *Operations Research,* 37, 791-797.

[17] Streibel, C.T. (1965). Sufficient statistics in the optimal control of stochastic systems. *Journal of Mathematical Analysis and Applications,* 12, 576-592.

[18] Jewell, W.S. (1963). Markov-renewal programming I: Formulation, finite return models; Markov-renewal programming II, infinite return models, example. *Operations Research,* 11, 938-971.

[19] Serfozo, R. (1979). An equivalence between continuous and discrete time Markov decision processes. *Operations Research,* 27, 616-620.

[20] Roberts, M.S. and F.A. Sonnenberg (2000). Decision modeling techniques. In *Decision Making in Health Care,* F. A. Sonnenberg and G. Chapman, (Eds.), Cambridge University Press, Cambridge, UK.

[21] Magni, P., S. Quaglini, M. Marchetti and G. Barosi (2000). Deciding when to intervene: a Markov decision process approach. *International Journal of Medical Informatics,* 60, 237-253.

[22] Torrance, G.W. (1976). Social preferences for health states: an empirical evaluate of three measurement techniques. *Socio-Economic Planning Sciences,* 10, 129-136.

[23] Torrance, G.W., D.H. Feeny, W.J. Furlong, R.D. Barr, Y. Zhang and Q. Wang (1996). Multiattribute utility function for a comprehensive

health status classification system. Health Utilities Index Mark 2. *Medical Care,* 34, 702-722.

[24] Drummond, M.F., B. O'Brien, G.W. Stoddart and G.W. Torrance (1997). *Methods for the Economic Evaluation of Health Care Programmes.* Oxford University Press, Oxford.

[25] Ahn, J.H. and J.C. Hornberger (1996). Involving patients in the cadaveric kidney transplant allocation process: A decision-theoretic perspective. *Management Science,* 42, 629-641.

[26] Samuelson, P. (1937). A note on measurement of utility. *Review of Economic Studies,* 4, 155-161.

[27] Frederick, S., G. Loewenstein and T. O'Donoghue (2002). Time discounting and time preference: A critical review. *Journal of Economic Literature,* XL, 351-401.

[28] Christensen-Szalanski, J.J. (1984). Discount functions and the measurement of patients' values. Women's decisions during childbirth. *Medical Decision Making,* 4, 47-58.

[29] Kirby, K.N. and N.N. Markovic (1995). Modeling myopic decisions: Evidence for hyperbolic delay-discounting within subjects and amounts. *Organizational Behavior and Human Decision Processes,* 64, 22-30.

[30] Gold, M.R., J. Siegel, L. Russell and M. Weinstein, Eds. (1996). *Cost-Effectiveness in Health and Medicine.* Oxford University Press, New York.

[31] Chapman, G.B. (2003). Time discounting of health outcomes. In *Time and Decision: Economic and Psychological Perspectives on Intertemporal Choice,* G. Loewenstein, D. Read and R. F. Baumeister, (Eds.), Russell Sage Foundation, New York.

[32] Pflug, G. and U. Dieter (1992). *Simulation and Optimization: Proceedings of the International Workshop on Computationally Intensive Methods in Simulation and Optimization, held at the International Institute for Applied Systems Analysis (IIASA), Laxenburg, Austria, August 23-25,1990.* Springer-Verlag, Berlin.

[33] Lefevre, C. (1981). Optimal control of a birth and death epidemic process. *Operations Research,* 29, 971-982.

[34] Lippman, S. (1973). Applying a new technique in the optimization of exponential systems. *Operations Research,* 23, 687-710.

[35] Hu, C., W.S. Lovejoy and S.L. Shafer (1993). Comparison of some suboptimal control policies in medical drug therapy. *Operations Research,* 44, 696-709.

[36] Hauskrecht, M. and H. Fraser (2000). Planning treatment of ischemic heart disease with partially observable Markov decision processes. *Artificial Intelligence in Medicine,* 18, 221-244.

[37] Ivy, J.S. (2002). A maintenance model for breast cancer detection and treatment. Submitted for publication.

[38] Alagoz, O., A.J. Schaefer, L.M. Maillart and M.S. Roberts (2002). Determining the optimal timing of living-donor liver transplantation using a Markov decision process (MDP) model. *Medical Decision Making,* 22, 558 (abstract).

[39] Roberts, M.S. and D.C. Angus (2002). The optimal timing of liver transplantation: Final report R01 HS09694. University of Pittsburgh, Pittsburgh, PA.

24 DYNAMIC INFLUENCE DIAGRAMS: APPLICATIONS TO MEDICAL DECISION MODELING

Gordon B. Hazen

Department of Industrial Engineering and Management Sciences
Northwestern University
Evanston IL 60208

SUMMARY

Influence diagrams are now a well established tool for modeling in decision analysis. Recently, *dynamic* influence diagrams have been applied to help structure stochastic processes. This chapter discusses dynamic influence diagrams for structuring continuous-time Markov chains, with particular focus on medical decision modeling. We describe our freely available Excel-based software package *StoTree,* in which dynamic influence diagram models may be readily formulated and solved. We present medical applications as examples, including a previously published cost-effectiveness analysis for total hip replacement.

KEY WORDS

Influence diagram, Stochastic tree, Dynamic Bayes net, Medical decision analysis, Medical cost-effectiveness, Total hip arthroplasty

24.1 INTRODUCTION

Influence diagrams are well-known graphical tools for formulating and solving decision problems under uncertainty [3, 17, 20, 26]. They are useful both as problem formulation tools and as a compact means of presenting the probabilistic structure of a completed model. An influence diagram can in principle represent any decision problem involving probabilistic uncertainty. However, the representation of a *stochastic process* model is at best cumbersome. We propose here a graphical extension of the influence diagram that we call the *dynamic* influence diagram, which facilitates the representation of stochastic processes, while retaining the formulation and presentation advantages of an influence diagram.

Whereas chance nodes in an influence diagram represent random variables, nodes in a dynamic influence diagram may represent random variables that change state over time. Just as in an influence diagram, arrows in a dynamic influence diagram indicate influence in the sense of probabilistic dependence. But now the meaning is that transition rates or transition probabilities in the influenced node may depend on the state of the influencing node or nodes. Alternately, a transition within an influencing node may trigger a transition in an influenced node. In either case, cycles of influence are possible, as for example, when the state of one node influences a transition rate in another node, whose state in turn influences a transition rate in the original node. Cycles of this kind are not permitted within a conventional influence diagram.

Any stochastic process can in principle be represented by a dynamic influence diagram. However, the notion of dynamic influence diagram is motivated by applications in medical treatment decision analysis, where conventional decision analytic approaches are popular, but uncertainty in long-term outcome is most naturally modeled as a stochastic process. Such models can be formulated as dynamic influence diagrams in which loosely coupled processes (for example, background mortality, disease progression and prosthesis loosening) are modeled separately and linked by triggers and transition rate dependencies. By formulating the model as a dynamic influence diagram, the analyst may decouple the model formulation process into manageable components that can later be linked appropriately. When complete, a dynamic influence diagram then provides a graphical overview of model structure.

Stochastic tree models [11, 12] and Markov chain models are particularly useful in medical models that must capture the long-term consequences of interventions. A stochastic tree is essentially a transition diagram for a continuous-time Markov model augmented by chance and decision nodes under the conventions for decision trees. Our modeling software *StoTree,* which runs in the Microsoft Excel spreadsheet environment, allows a user to graphically depict the internal structure of each node in a dynamic influence diagram on a worksheet in an Excel workbook. Each worksheet/node constitutes an independent stochastic tree model. The user may link these models by specifying triggers and transition rate dependencies, thereby implicitly specifying the influence arrows in a dynamic influence diagram. Moreover, *StoTree* allows the user to calculate mean quality-adjusted life years (QALYs) by rolling back the resulting stochastic tree in a manner analogous to decision tree rollback [11, 13]. All stochastic tree diagrams in this chapter are screen captures from *StoTree.* We will describe this software in more detail below.

We discuss dynamic influence diagrams in the next section. Following that, we give a graphical presentation of a dynamic influence diagram model for the cost-effectiveness of joint replacement surgery [2]. After giving a short description of our *StoTree* software, we conclude by summarizing the cost-effectiveness results from our joint replacement model.

24.2 DYNAMIC INFLUENCE DIAGRAMS

Influence diagrams are widely accepted tools for formulating and solving decision problems under uncertainty. Figure 24.1 presents an example influence diagram taken from our joint replacement model. In an influence diagram, oval (or circular) nodes are called *chance nodes* and represent uncertain variables; rectangular nodes are called *decision nodes* and represent decisions; and an arrow between two nodes indicates that the parent node *influences* the child node in a probabilistic sense. If the influence is deterministic, that is, if the child variable is a function of the state of the parent variable, then the child node is given a doubly outlined border. Doubly outlined nodes with no parents are therefore constants. (This is the traditional graphical convention, but others are utilized as well – see [3].)

We introduce a new feature into influence diagrams by allowing designated oval nodes to represent variables that may change state over time. Such variables are stochastic processes, and we call the associated oval node a *stochastic node.* We distinguish stochastic nodes from chance nodes by adding a wavy arrow below the node description. Figure 24.2 portrays an influence diagram containing a stochastic node *Prosthesis Status.* We call this a *dynamic influence diagram.*

The node *Prosthesis Status* in Figure 24.2 represents the stochastic process that is depicted as a stochastic tree in Figure 24.3. In Figure 24.3, an initially functioning prosthesis is subject to infection failure at an average rate of *rInfection* per year. Surgery immediately follows infection failure. If the surgery is successful (with probability *PISucc*), the prosthesis returns to its functioning state. If the surgery fails (with probability *PIFail*), it is repeated. There is a small chance *pIMort* of surgical mortality.

Figure 24.1 An influence diagram involving chance, decision and deterministic nodes

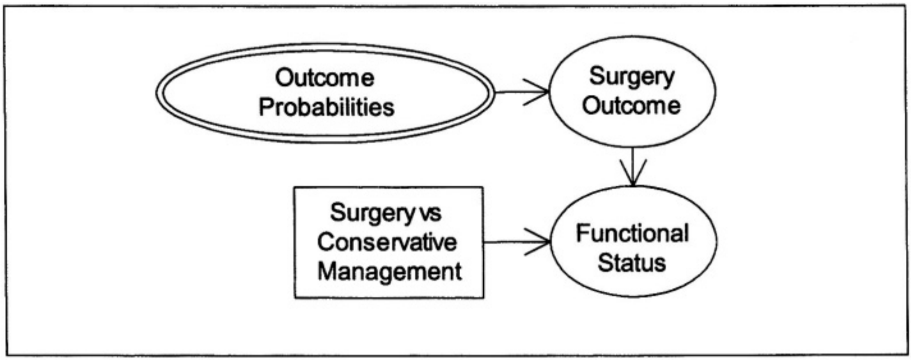


Figure 24.2 A dynamic influence diagram containing a node representing a stochastic process. Directed cycles are not permitted in conventional influence diagrams, which do not contain stochastic nodes. However, cycles involving stochastic nodes are allowable, and represent dynamic interaction between the variables in the cycle.

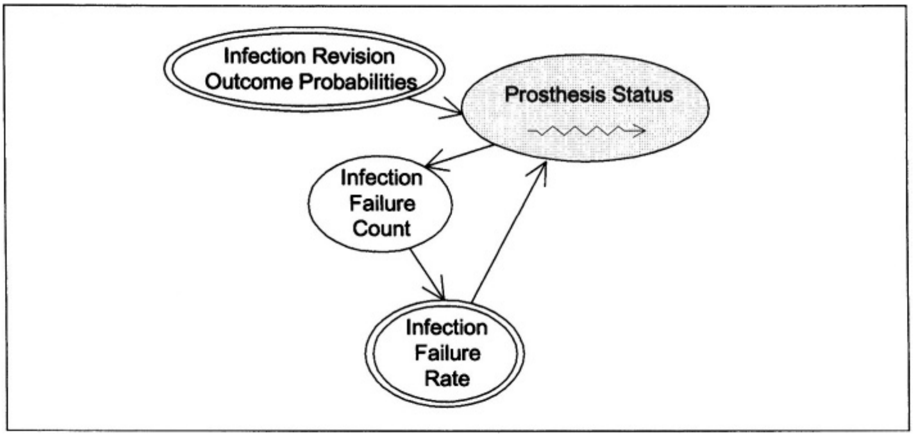
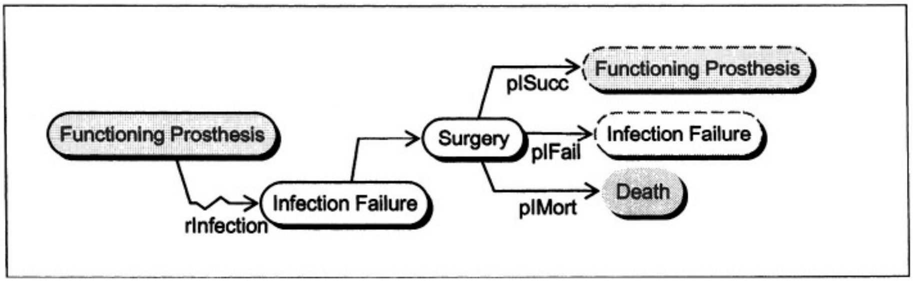


Figure 24.3 A stochastic tree diagram of the stochastic process *Prosthesis Status.*



The node *Infection Failure Count* in the influence diagram of Figure 24.2 represents the stochastic process depicted in Figure 24.4. *Infection Failure Count* serves to count the cumulative number of infections (up to three) in the process *Prosthesis Status.* It has no transition arrows: a patient's state changes only when triggered by an infection failure in the latter process. Therefore, we have included no wavy arrow in its node in Figure 24.2. The process *Infection Failure Count* is required because the infection failure rate *rInfection* of Figure 24.3 is higher when there have been more infection failures. This dependence is indicated in the influence diagram of Figure 24.2, where the node *Infection Failure Rate* is a deterministic function of *Infection Failure Count. Infection Failure Rate* in turn influences *Prosthesis Status,* thereby inducing a directed cycle of influences. Although directed cycles are forbidden in traditional influence diagrams, they are allowable in dynamic influence diagrams because nodes in a cycle may change state over time.

This example illustrates the use of a dynamic influence diagram for medical decision analysis, but in fact such diagrams can serve as graphical models for many stochastic processes. For instance, Figure 24.5 depicts a dynamic influence diagram for an M/M/1/4 queue, and Figure 24.6 displays the components. In general, a dynamic influence diagram can portray any generalized semi-Markov process (GSMP) (see, for example, [8] or [29] for a discussion of GSMPs).

24.3 AN OSTEOARTHRITIC JOINT REPLACEMENT MODEL

Total hip arthroplasty (THA) and total knee arthroplasty (TKA) have proven to be clinically reliable and durable procedures for the surgical treatment of severe osteoarthritis of the hip and knee [7, 19, 24, 25, 27]. An estimated 120,000 THAs are performed per year in North America [10], the majority of

Figure 24.4 The stochastic process *Infection Failure Count.* This process changes state only when triggered by an infection failure in *Prosthesis Status after THA.* Therefore, no transition arrows are present.

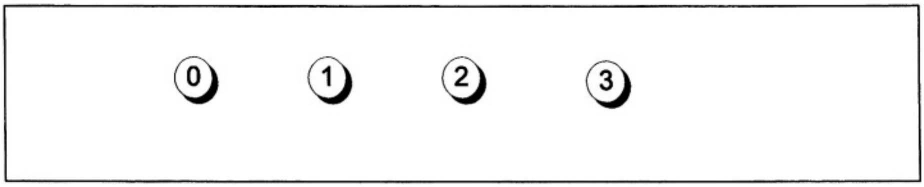
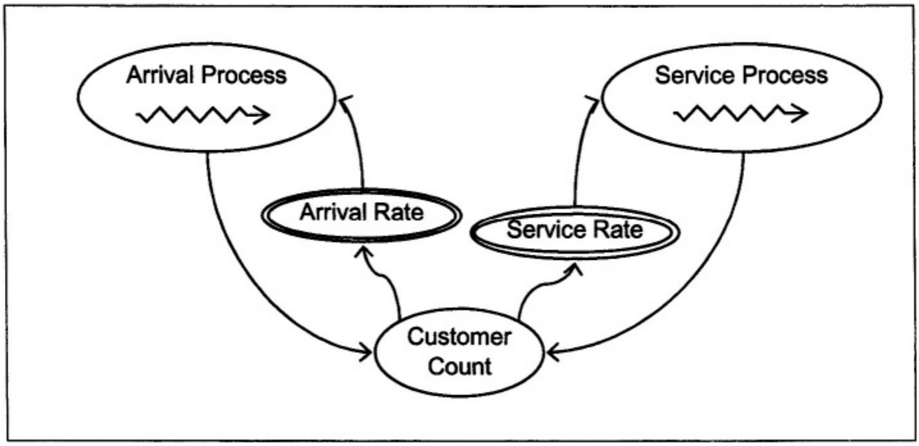


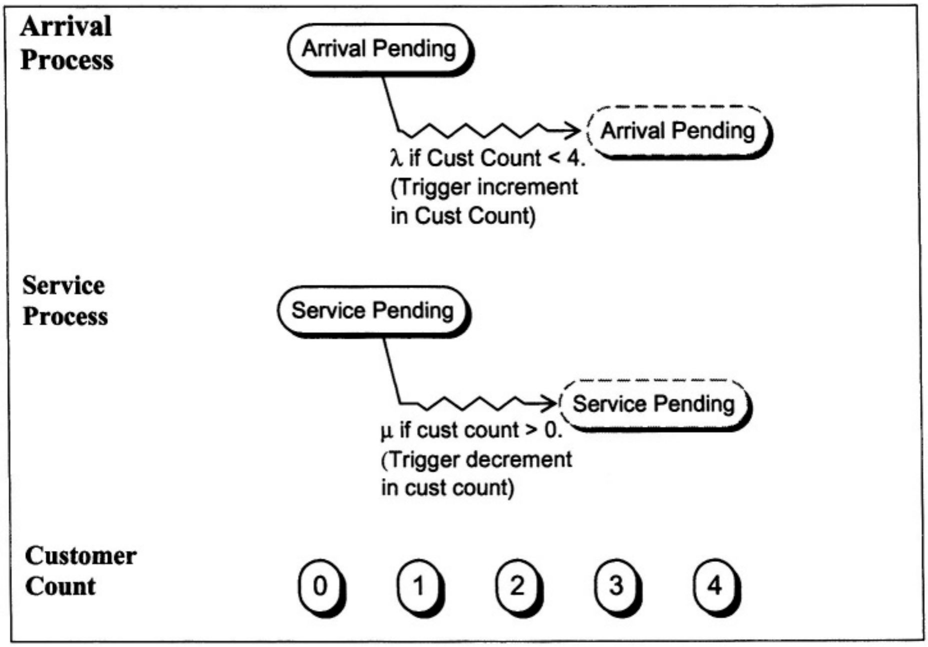
Figure 24.5 A dynamic influence diagram for a single-server queue



which are for patients with hip osteoarthritis. In the United States alone, 125,000 to 140,000 TKAs are performed annually [14, 23], with osteoarthritis and rheumatoid arthritis accounting for over 90% of these operations. Because these operations are performed predominantly on the elderly, their frequency is expected to increase as the population ages [15].

Although THA and TKA seem justified in terms of clinical success, they are particularly vulnerable to scrutiny in economic terms, for several reasons: Their indications are generally elective; their target population is largely geriatric; and they are high-technology procedures, much more expensive in the short term than simple medical management. From a societal or policy perspective, the cost-effectiveness of these procedures is therefore of particular interest. Moreover, because these procedures do not extend life, their cost-effectiveness must be measured in terms of dollars per *quality-adjusted life year* (QALY).

Figure 24.6 Components of the dynamic influence diagram of Figure 24.5

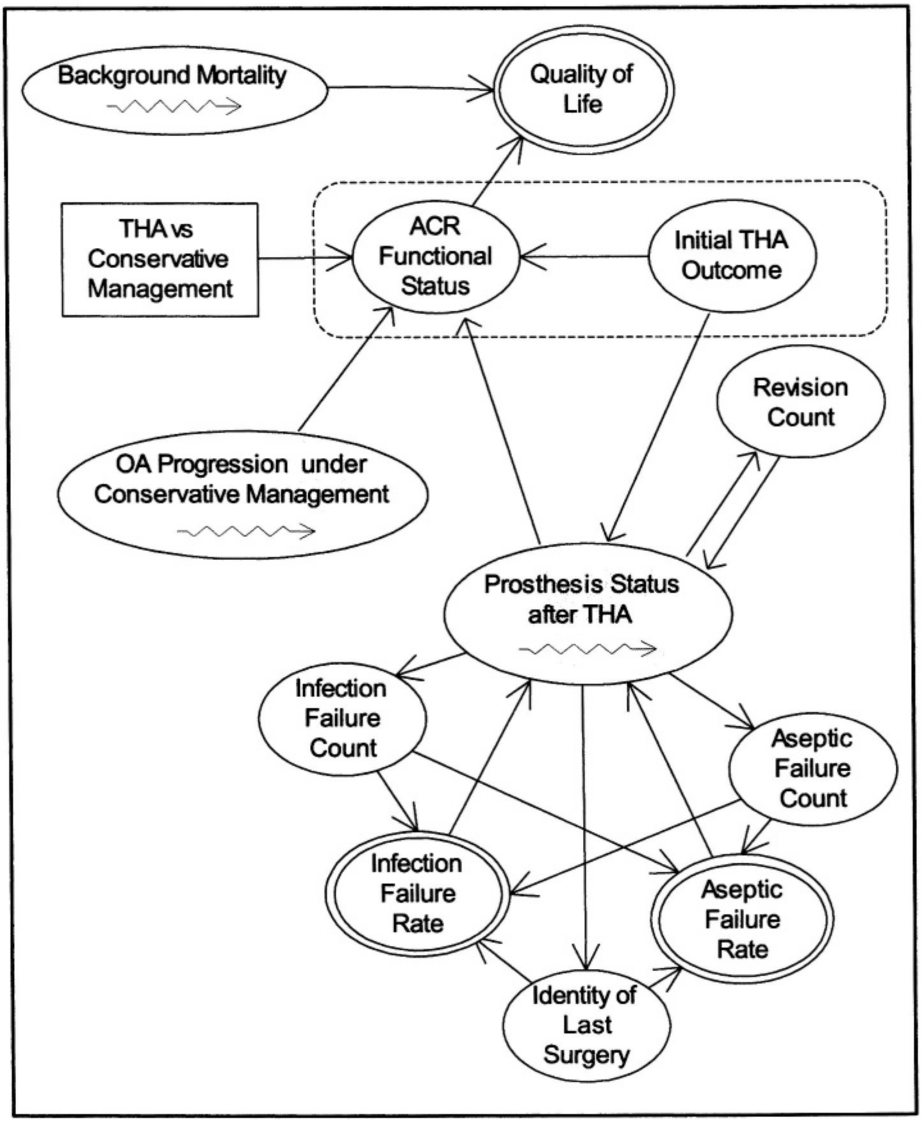


Although a few studies address the cost effectiveness of these procedures in the short term, we found no cost-effectiveness analyses for THA or TKA that considered long-term issues such as the need for revision surgeries, or worsening osteoarthritis and its associated custodial care costs. We therefore constructed decision-analytic models of the short- and long-term consequences of THA and TKA [2, 9]. The results indicate that when improvements in quality of life are included, THA and TKA can be among the most cost-effective of medical procedures, comparable or superior to well-accepted procedures such as cardiac bypass or renal dialysis. In fact, for some patients, these procedures can be *cost saving,* improving quality of life *and* reducing long-term costs compared to conservative medical management. We present here our THA model as an illustration of a modeling effort using dynamic influence diagrams.

24.3.1 An influence diagram model for joint replacement

Figure 24.7 presents our complete dynamic influence diagram model for the choice between THA and conservative management for hip osteoarthritis. The diagram contains three major stochastic nodes. For convenience, the

Figure 24.7 A dynamic influence diagram for our model of the choice between THA and Conservative Management. The two nodes surrounded by a the dashed line were grouped into a single component in our model.



nodes enclosed by the dashed line were combined into a single component in our model.

We discuss the detailed structure of each factor below. However, first we use Figure 24.7 to give an intuitive overview of the model. The purpose of the model is to calculate optimal mean quality-adjusted lifetime, and to that end, the node *Quality of Life* is present in the upper portion of Figure 24.7. *Quality of Life* is a deterministic function of *ACR Functional Status.* The latter is a four-level functional status scale adopted by the American College of Rheumatology (see below). *ACR Functional Status* depends on the decision node *THA versus Conservative Management.* If the choice is THA, then *ACR Functional Status* depends on *Initial THA Outcome,* which indicates the outcome of the initial hip replacement surgery, and *Prosthesis Status after THA,* a stochastic process describing prosthesis failure over time and subsequent revision surgeries. Prosthesis failures can be due to infection (infection failure) or mechanical failure (*aseptic* failure). If the choice is conservative management, then *ACR Functional Status* depends on *OA Progression under Conservative Management,* a stochastic process describing the functional deterioration of the hip due to osteoarthritis.

Prosthesis Status after THA interacts with a number of stochastic and chance nodes. First, if *Initial THA Outcome* is surgical failure, then an aseptic revision surgery is triggered in *Prosthesis Status after THA.* The stochastic nodes *Infection Failure Count* and *Aseptic Failure Count* record the cumulative number of prosthesis failures due respectively to infection and to mechanical failure. The stochastic node *Identity of Last Surgery* has possible values *Initial THA, Aseptic Revision* and *Infection Revision,* corresponding to the possible types of the most recent surgery. The variables *Infection Failure Rate* and *Aseptic Failure Rate* determine the average prosthesis failure rates in *Prosthesis Status after THA.* These rates are deterministic functions of *Infection Failure Count, Aseptic Failure Count,* and *Identity of Last Surgery.*

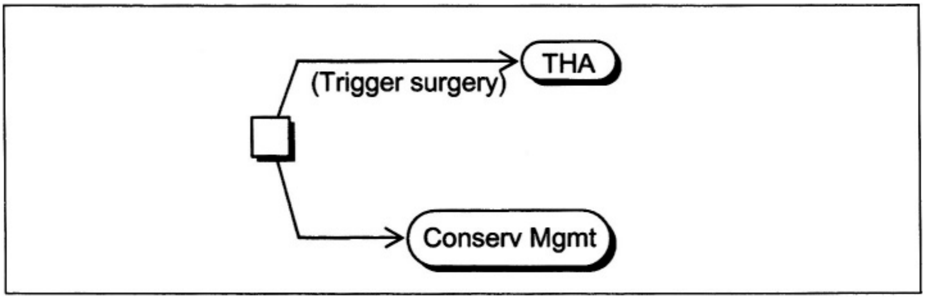
Finally, the stochastic node *Background Mortality* represents mortality due to causes unrelated to THA or conservative management of osteoarthritis. Its only effect is to reduce *Quality of Life* to zero when mortality occurs.

24.3.2 Modeling the initial THA decision

The structure for the decision node *THA vs. Conservative Management* is depicted in Figure 24.8. The structure of the combined nodes within the dashed rectangle in Figure 24.7 is shown in Figure 24.9. Here we depict the outcome of the initial THA surgery, should THA be chosen, as well as the subsequent functional status under either THA or conservative management.

We chose functional class as our primary measure of effectiveness for THA, adopting the American College of Rheumatology (ACR) functional status classification [28] for use in hip osteoarthritis, described in Table 24.1. We assumed the patient is initially in functional class III, a common status of individuals deciding whether to undergo THA. From Figure 24.9 we see that initial surgical outcome is ACR functional class I, II, or III. Functional class IV cannot be reached initially, but entry into this state can be triggered by progression of osteoarthritis or subsequent prosthesis failure. The quality-of-life values qI, qII, qIII, qIV we assigned to these functional classes are specified in Figure 24.9 as well.

Figure 24.8 The structure of the decision node *THA vs. Conservative Management*



24.3.3 Modeling prosthesis status after THA

The structure of the node *Prosthesis Status after THA* from Figure 24.7 is depicted in Figure 24.10. Following initial THA, the patient occupies the state *Daily Living,* in which the prosthesis is subject to both aseptic and infection failure. Revision surgery is undertaken after prosthesis failure. In practice, available bone stock limits the number of revision surgeries that can be undertaken. We assumed at most three revision surgeries were possible.

The node *Revision Count* (structure not shown, but identical to Figure 24.4) counts the cumulative number of revision surgeries. Should an aseptic or infection failure occur when *Revision Count* is equal to 3, then the *No Revision* branch is taken.

As the influence diagram in Figure 24.7 and the tables in Figure 24.10 indicate, prosthesis failure rates and the outcome probabilities for aseptic revision depend on the type of the most recent surgery, as well as the cumulative number of aseptic revisions and the cumulative number of infection revisions. These cumulative counts are kept by the stochastic components *Aseptic Failure Count* and *Infection Failure Count* (identical to

Figure 24.9 The structure of the combined nodes depicting the outcome of initial THA surgery and subsequent ACR functional status. The patient initially occupies the state "ACR Class III". If THA is chosen in Figure 24.8, then transition is triggered from "ACR Class III" to "Surgery". The state "ACR Class IV" can be reached only when triggered by events in other components of the model.

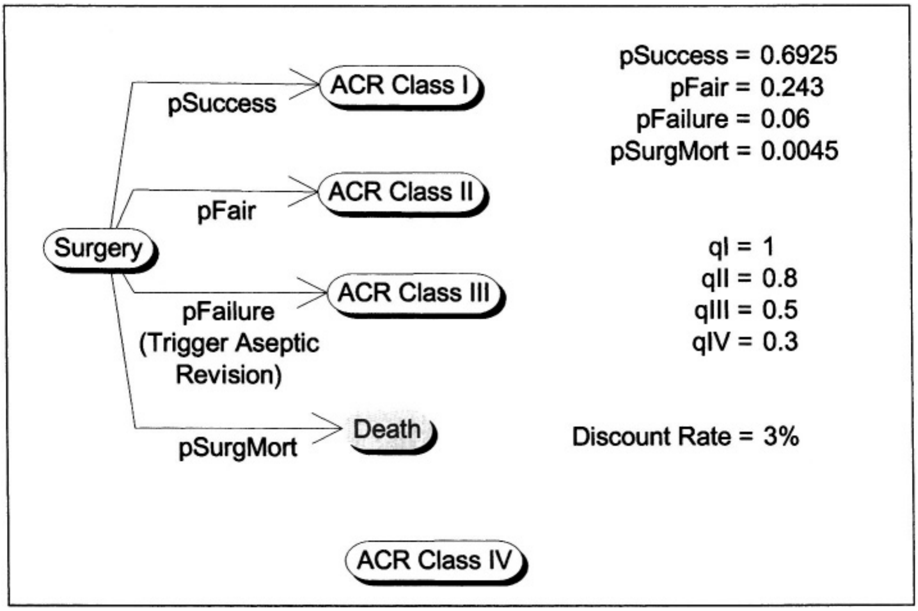
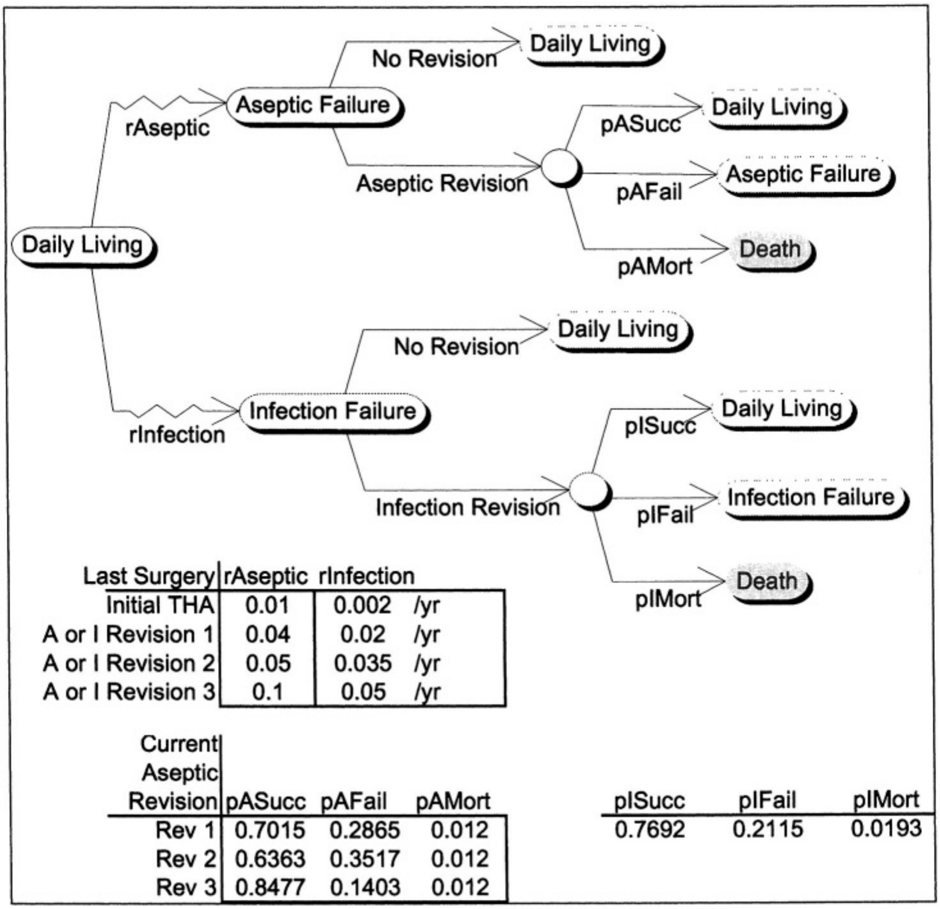


Table 24.1 American College of Rheumatology (ACR) functional classifications for hip osteoarthritis.

Class	Description
I	Complete ability to carry on all usual duties without handicap
II	Adequate for normal activities despite handicap of discomfort or limited motion in the hip
III	Limited only to little or none of duties of usual occupation or self-care
IV	Incapacitated, largely or wholly bedridden or confined to wheelchair, little or no self-care

Figure 24.4), whose levels are incremented by one, respectively, when an aseptic or infection failure occurs. The type of the last surgery is recorded in the stochastic component *Identity of Last Surgery* shown in Figure 24.11.

Figure 24.10 The structure of the node *Prosthesis Status after THA* from Figure 24.7. Rates of aseptic failure and infection failure are given in the accompanying tables, as are the surgical outcome probabilities. Both rates and probabilities depend on the levels of the auxiliary nodes *Aseptic Failure Count, Infection Failure Count, Revision Count,* and *Last Surgery.*



Last Surgery	rAseptic	rInfection
Initial THA	0.01	0.002 /yr
A or I Revision 1	0.04	0.02 /yr
A or I Revision 2	0.05	0.035 /yr
A or I Revision 3	0.1	0.05 /yr

Current Aseptic Revision	pASucc	pAFail	pAMort
Rev 1	0.7015	0.2865	0.012
Rev 2	0.6363	0.3517	0.012
Rev 3	0.8477	0.1403	0.012

pISucc	pIFail	pIMort
0.7692	0.2115	0.0193

24.3.4 Stochastic factor for conservative management

The stochastic node *OA Progression under Conservative Management* in Figure 24.7 represents a two-state stochastic process in which the initial state occupied is functional class III but later transition may occur to functional class IV. Figure 24.12 depicts this process as a stochastic tree in which the rate of transition from state III to state IV occurs at an average rate of *rNatural* per year.

Figure 24.11 The stochastic factor *Identity of Last Surgery.* Transitions in this factor can occur only when triggered by revision surgery in the factor *Prosthesis Status after THA.*

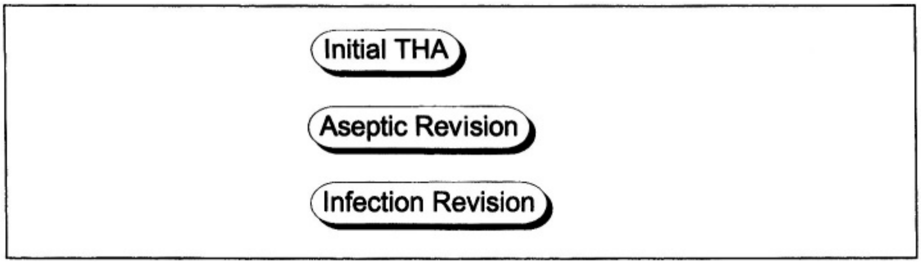
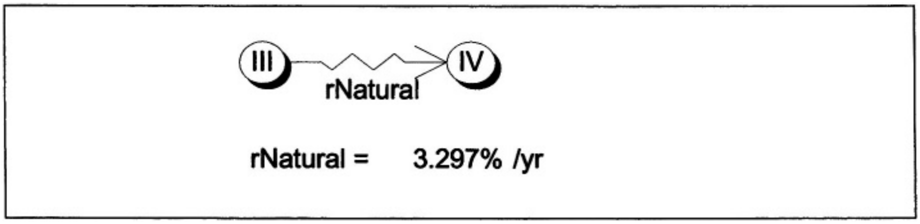


Figure 24.12 The node *OA Progression under Conservative Management* in Figure 24.7 represents the stochastic process depicted here, in which transition occurs from functional status III to IV at an average rate of *rNatural* = 3.297% per year.

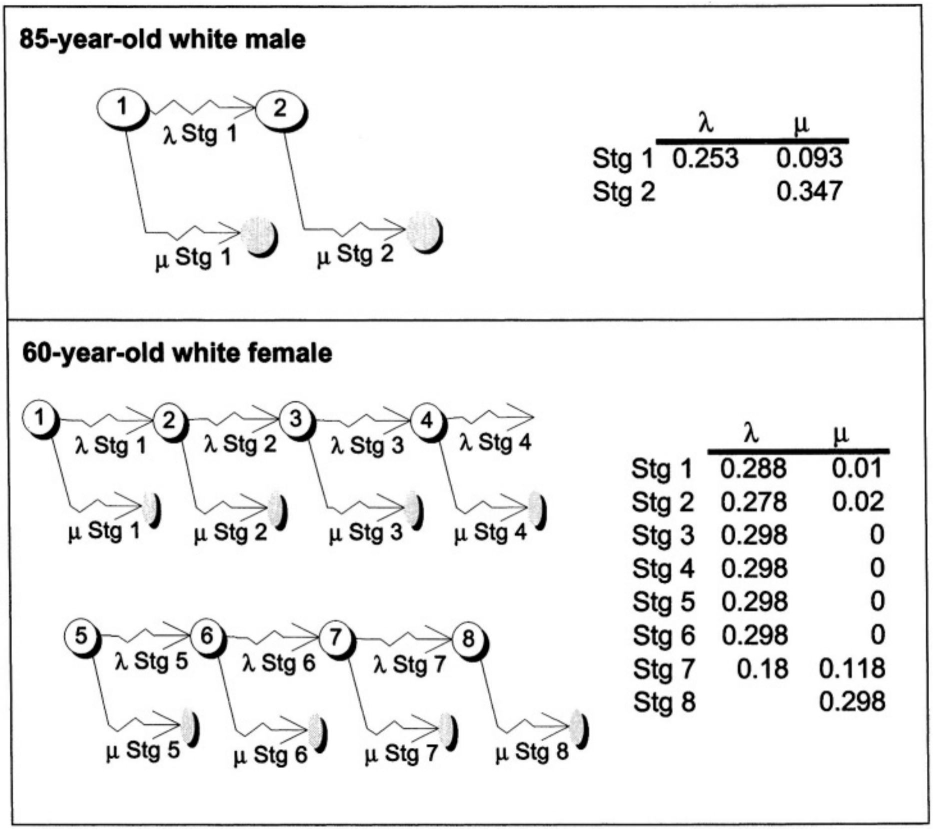


24.3.5 Mortality structure

The stochastic node *Background Mortality* in Figure 24.7 represents mortality due to causes unrelated to hip osteoarthritis or hip replacement. The underlying stochastic process is depicted by one of the *Cox* stochastic tree models shown in Figure 24.13. The Cox model [5] is a useful way to implement a phase approximation [18] for human survival times. For a particular age and gender, a Cox model with the appropriate number of stages and parameter values has survival probabilities that quite closely approximate the true ones [22]. For example, Figure 24.14 compares the true survival probabilities for a 60-year-old white female with the survival probabilities for the Coxian model of Figure 24.13.

Factoring background mortality from the overall model in this way results in a useful *modularity* property: Should one wish to run the model for a different age or gender, one need not alter any other part of the model. Instead, one can merely substitute the appropriate age- and gender-specific Cox mortality component.

Figure 24.13 Cox stochastic trees that closely approximate observed mortality for the given ages and genders. Numbered nodes may be thought of as life stages, between which transition occurs at rates λ Stg 1, λ Stg 2, ... Empty nodes represent mortality, to which transition occurs with stage-dependent rates μ Stg 1, μ Stg 2, ...



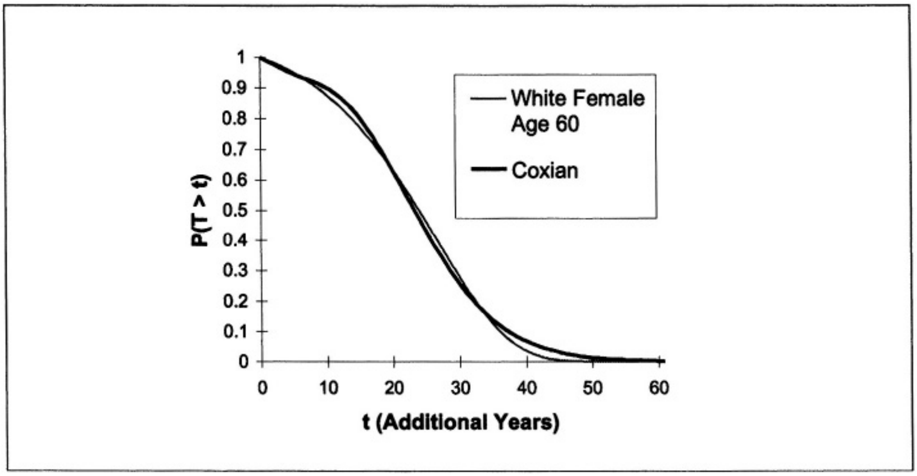
	λ	μ
Stg 1	0.253	0.093
Stg 2		0.347

	λ	μ
Stg 1	0.288	0.01
Stg 2	0.278	0.02
Stg 3	0.298	0
Stg 4	0.298	0
Stg 5	0.298	0
Stg 6	0.298	0
Stg 7	0.18	0.118
Stg 8		0.298

24.4 SOFTWARE FOR FACTORED STOCHASTIC TREE MODELING

StoTree is a graphical interface tool for the formulation and solution of factored stochastic tree models, implemented as a Microsoft Excel add-in. A *factored* stochastic tree is a stochastic tree whose set of possible states is the Cartesian product of the sets of possible states of its components, or factors. Each component, or factor, represents a node in a dynamic influence diagram model of the process. The StoTree add-in and supporting documentation are

Figure 24.14 Survival probabilities for a white female age 60 (U.S. Center for Health Statistics, 1991) are quite closely approximated by the survival probabilities for the Cox model of Figure 24.13.



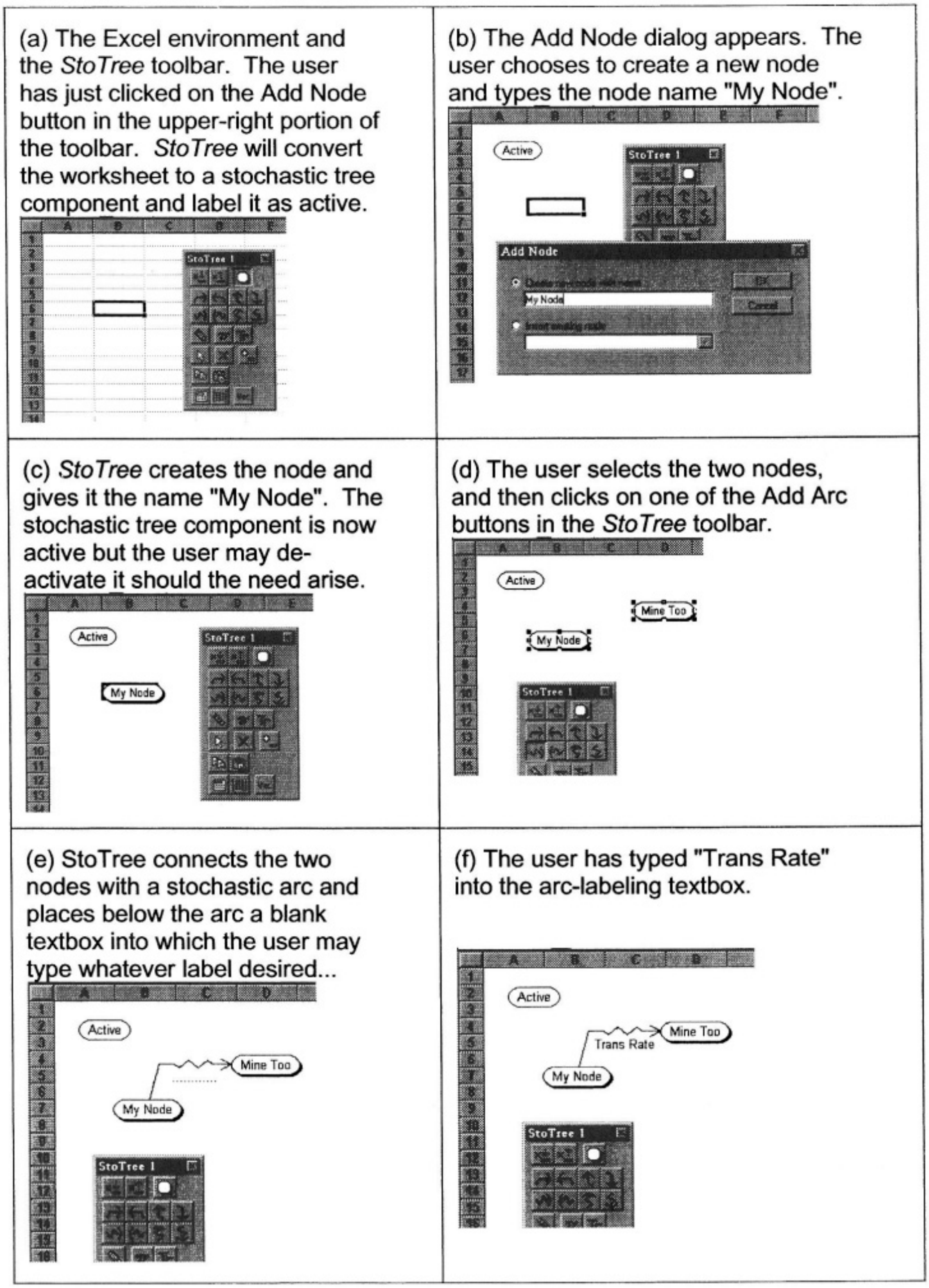
available for downloading at the web site www.iems.nwu.edu/~hazen/. Operating in a spreadsheet environment, the *StoTree* user can access all the usual features of that environment in addition to those of *StoTree.* The modeling effort begins with graphical model composition and parameter specification. The user concludes by using the rollback features in *StoTree* to calculate average quantities of interest such as mean quality-adjusted life years or average costs.

24.4.1 Graphical model composition and parameter specification

Figure 24.15 illustrates typical first steps of a user-initiated *StoTree* session. Via point-and-click operations, the user can create nodes, name and position them as desired, and connect nodes with either stochastic or chance arcs, to which the user may attach any desired label. Arcs may be drawn in any of the four compass directions. Each worksheet in the Excel workbook to which nodes have been added is regarded by *StoTree* as a component in a multi-factor stochastic tree, or equivalently, as a node in a dynamic influence diagram. Capabilities not shown in Figure 24.15 include subtree copy and paste, node and arc deletion, and tree redraw.

Once graphical structure has been specified, the user can associate numerical parameters with nodes and arcs. Parameters for arcs consist of probabilities for chance arcs, rates for stochastic arcs, and tolls for either type of arc.

Figure 24.15 The initial graphical model construction stages in *StoTree*



(a) The Excel environment and the *StoTree* toolbar. The user has just clicked on the Add Node button in the upper-right portion of the toolbar. *StoTree* will convert the worksheet to a stochastic tree component and label it as active.

(b) The Add Node dialog appears. The user chooses to create a new node and types the node name "My Node".

(c) *StoTree* creates the node and gives it the name "My Node". The stochastic tree component is now active but the user may de-activate it should the need arise.

(d) The user selects the two nodes, and then clicks on one of the Add Arc buttons in the *StoTree* toolbar.

(e) StoTree connects the two nodes with a stochastic arc and places below the arc a blank textbox into which the user may type whatever label desired...

(f) The user has typed "Trans Rate" into the arc-labeling textbox.

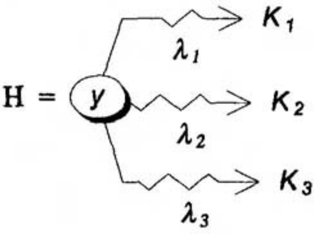
Node parameters consist of a quality or cost rate, and a discount rate. These are used in the rollback operation described below. Arc parameters can depend on the state of other factors. The user may also add one or more triggers on any arc, which force transitions in other specified components when that arc is traversed. All of these features are accessed via dialog boxes called up from the *StoTree* toolbar.

24.4.2 Rollback algorithm

StoTree implements a form of the *Markovian* utility function [13]. Let $y^t h$ denote a duration t visit to state y followed by any other sequence h of states and durations. If x is the previously visited state, then the Markovian utility assigned to $y^t h$ is equal to

$$u(y^t h \mid x) = w(y \mid x) + \int_0^t v(y) e^{-a(y)s} ds + e^{-a(y)t} u(h).$$

Here $w(y/x)$ is a toll associated with the transition from x to y; $v(y)$ is a quality rate specific to state y, and $a(y)$ is a discount rate. At a stochastic fork



in which subtrees K_i are reached from state y at competing rates λ_i, the rollback equation for calculating expected utility can be shown to take the simple form

$$E[u(H) \mid x] = w(y \mid x) + \frac{v(y) + \sum_i \lambda_i E[u(K_i \mid y)]}{a(y) + \sum_i \lambda_i}.$$

At a chance fork with associated probabilities p_i, the usual probability-weighted average is used:

$$E[u(H) \mid x] = w(y \mid x) + \sum_i p_i E[u(K_i \mid y)].$$

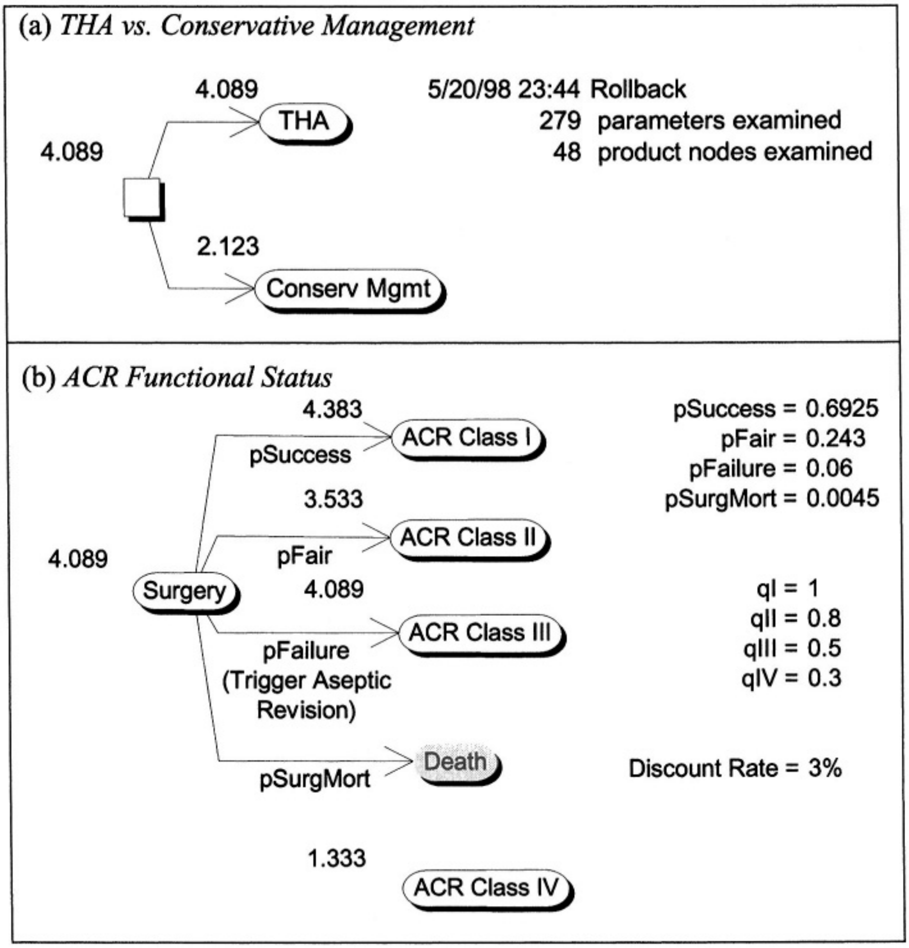
StoTree repeatedly evaluates these equations for the user-specified multi-factor stochastic tree. In this context, the states y are vectors $y = (y_1, \ldots, y_m)$, where y_i is the state of the i^{th} factor. If the i^{th} factor contains k_i non-death states, then in principle there are $k_1 \cdot k_2 \cdot \ldots \cdot k_m$ product states $y = (y_1, \ldots, y_m)$ for which expected utility must be calculated. For example, the full THA model specified above has $3\cdot5\cdot2\cdot5\cdot4\cdot4\cdot3\cdot4\cdot k_8 = 28800\cdot k_8$ possible product states, where k_8 is the number of stages in the Coxian model (equal above to 2 or 8). In Excel's slow computing environment, such a large number of product states would give unacceptably lengthy rollback times. However, *StoTree* bypasses this problem by calculating expected utility only for those product states that are *reachable* from the initial combination of states in each factor. For example, due to the many triggers present in the THA model, the number of reachable product states is only $4 + 22\cdot k_8$. *StoTree* identifies reachable product states by performing a breadth-first traversal of the product tree beginning at the combination of user-specified initial states. lifetimes for our THA model, as displayed in two of the model's worksheets. A useful feature available in *StoTree* is the linking of rollback entries to cell values. For example, at the chance fork in Figure 24.16b, the user can enter all probability parameters as cell references. *StoTree* then incorporates these cell references into the rollback formulas. The result is that the rollback entries will change when values in referenced cells change. For example, should the cell entry 0.6925 for *pSuccess* be changed, then the rollback values will also change when the spreadsheet is recalculated. This can be very useful for sensitivity analysis.

StoTree displays rollback values next to the corresponding nodes in each worksheet. For example, Figure 24.16 displays mean quality-adjusted lifetimes for our THA model, as displayed in two of the model's worksheets. A useful feature available in *StoTree* is the linking of rollback entries to cell values. For example, at the chance fork in Figure CHN.16b, the user can enter all probability parameters as cell references. *StoTree* then incorporates these cell references into the rollback formulas. The result is that the rollback entries will change when values in referenced cells change. For example, should the cell entry 0.6925 for *pSuccess* be changed, then the rollback values will also change when the spreadsheet is recalculated. This can be very useful for sensitivity analysis.

24.5 COST-EFFECTIVENESS OF JOINT REPLACEMENT

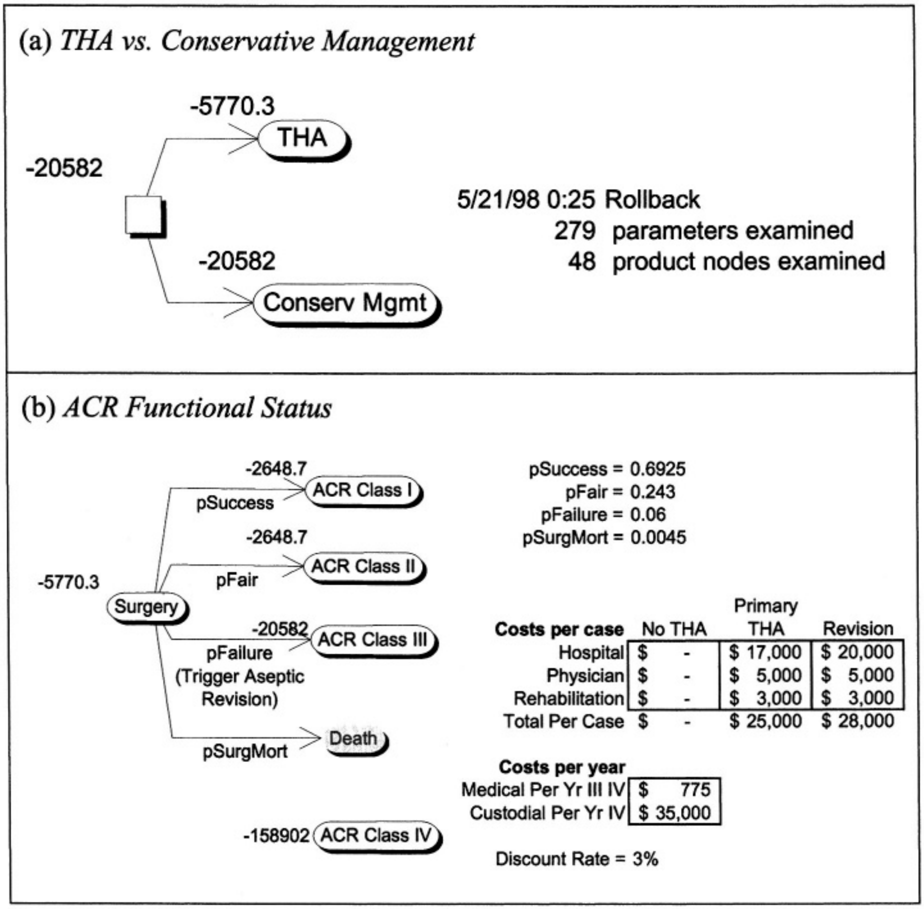
Thus far we have only discussed the computation of mean quality-adjusted lifetimes for the hip replacement decision. In order to conduct a cost-effectiveness analysis, one must also calculate mean lifetime costs. This is

Figure 24.16 Rollback results for the THA model for the case of a white male aged 85, as displayed in the model components *THA vs. Conservative Management* and *ACR Functional Status.* Rollback quantities are displayed adjacent to the appropriate nodes, and indicate the mean number of quality-adjusted life years remaining beginning at that state, assuming all other components occupy their initial states.



easily accomplished in a stochastic tree model: Simply replace quality rates with ongoing cost rates, and include one-time costs as tolls on the appropriate arcs. Figure 24.17 displays both ongoing and one-time costs for our THA model, as well as the rollback results using these costs.

Figure 24.17 Calculation of expected costs for a white male aged 85 via rollback in *StoTree.* The cost data used is shown next to the *ACR Functional Status* tree. Total lifetime discounted cost for conservative management is $20582, and for THA is $5770.30 + $25000 = $30770.30. Conservative management is less costly at this age.



5/21/98 0:25 Rollback
279 parameters examined
48 product nodes examined

pSuccess = 0.6925
pFair = 0.243
pFailure = 0.06
pSurgMort = 0.0045

Costs per case	No THA	Primary THA	Revision
Hospital	$ -	$ 17,000	$ 20,000
Physician	$ -	$ 5,000	$ 5,000
Rehabilitation	$ -	$ 3,000	$ 3,000
Total Per Case	$ -	$ 25,000	$ 28,000

Costs per year	
Medical Per Yr III IV	$ 775
Custodial Per Yr IV	$ 35,000

Discount Rate = 3%

Overall cost-effectiveness results for our THA model are presented in Table 24.2. For a white male aged 85, the cost-effectiveness ratio is $5183 per QALY gained, a value superior to well accepted procedures such as cardiac bypass and renal dialysis [2]. For a white female aged 60, the procedure both improves quality of life *and* reduces costs. Although THA does not extend life expectancy, the intuitive rationale for its superiority is clear from the table: Compared to conservative management, THA reduces average time

Table 24.2 The results for THA versus conservative management based on our stochastic tree model. For a white female aged 60, THA saves costs *and* increases quality-adjusted life expectancy. For a white male aged 85, the marginal cost-effectiveness ratio of $5183 is superior to accepted procedures such as renal dialysis or cardiac bypass.

	White Male Age 85		White Female Age 60	
	THA	Conserv	THA	Conserv
Mean Dctd Years in				
ACR Class I	2.944	0	9.717	0
ACR Class II	1.384	0	5.078	0
ACR Class III	0.063	3.952	0.608	11.39
ACR Class IV	0.022	0.49	0.48	4.634
Mean Dctd Life Expect	4.413	4.442	15.883	16.024
Mean Dctd QALY	4.089	2.123	14.23	7.087
Mean Dctd Costs				
Primary Surgery, Rehab	$25,000	-	$25,000	-
Revision Surgery, Rehab	$4,929	-	$12,284	-
Ongoing Medical	$66	$3,442	$843	$12,421
Custodial	$776	$17,140	$16,789	$162,175
Total	$30,771	$20,582	$54,916	$174,596
Marginal cost	$10,189		$(119,680)	
Marginal effectiveness (QALY)	1.966		7.143	
Marginal CE ratio	$5,183		-	

spent in functional classes III and IV, which are expensive, low-quality health states.

24.6 CONCLUSION

We have introduced here a new graphical tool, the dynamic influence diagram, for constructing and portraying complex stochastic models in medical decision analysis. These models can be used to guide individual decision making, or can be applied at the societal level to determine the cost-effectiveness of these procedures. We have illustrated their application to modeling joint replacement decisions. However, a dynamic influence diagram can in principle be used to model many stochastic processes. We have represented stochastic nodes in our diagram as stochastic trees, but alternate stochastic process models such as discrete-time Markov chains could be used as components in a dynamic influence diagram.

The probabilistic components of influence diagrams have been well-studied as *Bayes nets* or *knowledge maps* in the artificial intelligence literature (e.g., [4, 21]). *Dynamic* Bayes nets have also received attention (e.g., [1, 6, 16]), although continuous-time formulations such as those considered here have not been treated using our graphical formalism. For static Bayes nets, a graphical property known as *d-separation* guarantees conditional independence in the probabilistic sense. Moreover, a Bayes net is a *minimal I-map* for the probability distribution it represents: that is, every graphically inferred independence statement is correct, and no net with fewer arcs has this property [21]. Two open research questions are (1) what graphical characterization of conditional independence is possible in a dynamic Bayes net/ influence diagram, and in particular whether or in what way the d-separation notion carries over; and (2) how the characterization of static Bayes nets as minimal I-maps extends to dynamic Bayes nets.

The use of dynamic influence diagrams and corresponding factored stochastic trees permits a modular approach to model construction, facilitates the presentation of the models, and opens models to inspection by other parties. The graphical modeling tool *StoTree* enables users to formulate and solve factored stochastic tree models in a user-friendly spreadsheet environment.

References

[1] Boyen, X. and D. Koller (1998). Tractable inference for complex stochastic processes. In *Uncertainty in Artificial Intelligence: Proceedings of the Fourteenth Conference (UAI-1998),* Morgan Kaufmann Publishers, San Francisco, CA, 33-42.

[2] Chang, R.W., Pellissier, J.M., and G.B. Hazen (1996). A cost-effectiveness analysis of total hip arthroplasty for osteoarthritis of the hip. *Journal of the American Medical Association,* 275, 858-865.

[3] Clemen, R.L. (1996). *Making Hard Decisions: An Introduction to Decision Analysis.* Duxbury Press, Belmont, CA.

[4] Cowell, R.G., Dawid, P.A., Lauritzen, S.L., and D.J. Spiegelhalter (1999). *Probabilistic Networks and Expert Systems.* Springer Publishers.

[5] Cox, D.R. (1959). A use of complex probabilities in the theory of stochastic processes. *Proceedings of the Cambridge Philosophical Society*, 51, 313-319.

[6] Dean, T. and K. Kanazawa (1989). A model for reasoning about persistence and causation. *Artificial Intelligence,* 93, 1–27.

[7] Dennis, D.A., Clayton, M.L., O'Donnell, S., Mack, R.P., and E.A. Stringer (1992). Posterior cruciate condylar total knee arthroplasty. Average 11-year follow-up evaluation. *Clinical Orthopedics,* 281, 168-176.

[8] Glasserman, P. (1990). *Gradient Estimation via Perturbation Analysis.* Kluwer Academic Publishers, Boston, MA.

[9] Gottlob, C.A., Pellissier, J.M., Wixson, R.L., Stern, S.H., Stulberg, S.D., and R.W. Chang (1996). The long-term cost-effectiveness of total knee arthroplasty for osteoarthritis. Working paper, Department of Preventive Medicine, Northwestern University Medical School, Chicago, IL.

[10] Harris, W.H. and C.B. Sledge (1990). Total hip and total knee replacement. *New England Journal of Medicine,* 323,725–731.

[11] Hazen, G.B. (1992). Stochastic trees: A new technique for temporal medical decision modeling. *Medical Decision Making,* 12, 163-178.

[12] Hazen, G.B. (1993). Factored stochastic trees: a tool for solving complex temporal medical decision models. *Medical Decision Making,* 13, 227-236.

[13] Hazen, G.B. and J.M. Pellissier (1996). Recursive utility for stochastic trees. *Operations Research,* 44, 788-809.

[14] Hoffman, A.A., Rubash, H.E., Crofford, T.W., Fletcher, F.J., and L.S. Crosset (1993). Knee and leg: reconstruction. In *Orthopaedic Knowledge Update 4,* Frymoye, J.W., Ed.. Rosemont, American Academy of Orthopaedic Surgeons, 603-624.

[15] Liang, M.H., Cullen, K.E., Larson, M.G., Thompson, M.S., Schwartz, J.A., Fossel, A.H., Roberts, W.N., and C.B. Sledge (1986). Cost-effectiveness of total joint arthroplasty in osteoarthritis. *Arthritis and Rheumatology,* 29, 937-943.

[16] Murphy, K. (In press). Dynamic Bayesian Networks (Draft). *Probabilistic Graphical Models,* Joran, M., Ed.

[17] Nease R.F. and D.K. Owens (1997). Use of influence diagrams to structure medical decisions. *Medical Decision Making,* 17, 263-275.

[18] Neuts, M.F. (1994). *Matrix-Geometric Solutions in Stochastic Models: An Algorithmic Approach.* Dover Publications, New York.

[19] NIH Consensus Development Panel on Total Hip Replacement (1995). Total hip replacement. *Journal of the American Medical Association,* 273,1950-1956.

[20] Owens, D.K., Shachter, R.D., and R.F. Nease (1997). Representations and analysis of medical decision problems with influence diagrams. *Medical Decision Making,* 17, 241-262.

[21] Pearl, J. (1991). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference.* Morgan Kaufman Publishers, San Francisco, CA.

[22] Pellissier, J.M. and G.B. Hazen (1994). Coxian mortality factors for stochastic tree modeling (abstract). *Medical Decision Making,* 14, 439.

[23] Quam, J.P., Michet, C.J., Wilson, M.G., Rand, J.A., Ilstrup, D.M., Melton, L.J., and S.L. Wallrichs (1991). Total knee arthroplasty: a population based study. *Mayo Clinic Proceedings,* 66, 589-595.

[24] Ranawat, C.S., Flynn, W.F., Saddler, S., Hansraj, K.K., and M.J. Maynard (1993). Long-term results of the total condylar knee arthroplasty. A 15-year survivorship study. *Clinical Orthopedics,* 286, 94-102.

[25] Scuderi, G.R., Insall, J.N., Windsor, R.E., and M.B. Moran (1989). Survivorship of cemented knee replacements. *Journal of Bone and Joint Surgery,* 710B, 798-803.

[26] Shachter, R.D. (1986). Evaluating influence diagrams. *Operations Research,* 34, 871-882.

[27] Stern, S.H. and J.N. Insall (1992). Posterior stabilized prosthesis. Results after follow-up of nine to twelve years. *Bone and Joint Surgery,* 74-A, 980-986.

[28] Steinbroker, O., Traeger, C.H., and R.C. Batterman (1949). Therapeutic criteria in rheumatoid arthritis. *Journal of the American Medical Association,* 140, 659-662.

[29] Whitt, W. (1980). Continuity of generalized semi-Markov processes. *Mathematics of Operations Research,* 5, 494-501.

25 ASSESSMENT OF THE BENEFITS OF ANESTHESIA PATIENT RISK REDUCTION MEASURES

M. Elisabeth Paté-Cornell

Department of Management Science and Engineering
Stanford University
Stanford, CA 94305

SUMMARY

This chapter presents an analytical framework based on Bayesian analysis that was used to evaluate human and management risk factors for patients undergoing anesthesia, and the effects of a variety of proposed measures for mitigating those risks. More specifically, the analysis considers the frequency and the effects of various risk factors, the extent to which safety measures based on management improvements can decrease the chances that they occur, and the effects of these safety measures on patient risk. The analysis demonstrates that the accident sequences that had received the most attention because they had made the headlines were not the largest contributors to the overall patient risk. The analysis finds that most of the problems are not caused by rare events, but by more mundane factors such as fatigue and poor supervision of residents. Closer supervision of residents, periodic re-certification and simulator training appeared to be among the most potentially effective measures for reducing patient risk. A similar model can be applied to other medical problems involving risk, such as assessing the performance of surgeons or early assessment of medical devices (before comprehensive testing on large populations).

KEY WORDS

Risk analysis, Bayesian analysis, Anesthesia

25.1 RISK ANALYSIS IN THE ABSENCE OF A COMPLETE STATISTICAL DATABASE

The standards for risk analysis, in the medical field, rest on the use of "evidence-based" studies and statistical data sets involving substantial time series. Such data are gathered, for example, when explicitly required for the approval of a new procedure or a new drug. For existing procedures, however, the data may not exist, and it may not be in the legal interest of the parties involved to gather them. In the case of anesthesia accidents, for example, data exist for the cases where death or severe brain damage occurred in the course of a benign operation. But during a normal operation in which the patient's death can be attributed to his or her condition, the actual probability of an anesthesia accident is not known because the exact causes may not be clear and it may not be in the best interest of the medical institutions to gather such information.

A number of potential measures that may reduce the risk of an anesthesia accident can be considered; but without statistical data, their benefits cannot be easily evaluated, and such statistical data are unlikely to be gathered because they may be damaging to the current system. Therefore, another approach is needed to identify effective risk reduction measures and evaluate their benefits, which as is shown below, could be substantial.

Although a small proportion of anesthesia accidents is caused by technical malfunctions of anesthesia machines, such incidents are rare. Most accidents are caused by human errors. Thus, effective risk reduction measures are generally not technical but instead concern mainly the management of the personnel. Trained anesthesiologists as well as nurse anesthesiologists can experience a number of problems, ranging from substance abuse to poor supervision, which may affect their performance and increase the risk to the patients. The question is thus to know: (1) what is the frequency and what are the effects of these problems, (2) to what extent can a number of safety measures based on management improvements decrease the chances that they occur, and (3) what are the effects of these safety measures on patient risk.

In this chapter we present an analytical framework based on Bayesian analysis, which is different from classical statistical models. This approach involves three steps. First, we analyze the probability of an anesthesia accident based on the different sequences of events that can constitute accident scenarios and on their probabilities. Second, we examine the kinds of problems that anesthesiologists can experience that can affect their performance. For a random operation, we assess the probability that an anesthesiologist experiences any one of these problems. We then link the presence of such problems to the value of parameters in a patient risk model (e.g., the frequency of accident initiators) in order to estimate the effect of these problems on the safety of the patient. Finally, in the third phase of the analysis, we consider a number of potential management measures that a hospital's administration might implement to reduce the chances that an anesthesiologist in any operation is affected by any of these problems. These include, for example, improved supervision of residents and periodic re-certification of all practitioners. We then estimate the risk reduction associated with these measures by assessing their effects on the probabilities of different problems in the overall population of anesthesiologists. In the second phase we computed the probability that an anesthesiologist experiencing a particular type of problem might cause an accident. We can therefore compute the risk reduction benefits associated with reducing the chances that an anesthesiologist experiences various problems.

In order to make the analysis manageable, we decided to restrict it to "healthy" patients, that is, those for whom the initial cause of surgery cannot explain death or brain damage during the operation. These could be, for example, otherwise healthy patients undergoing knee surgery. We also chose to consider only medical doctors trained as anesthesiologists, as opposed to nurse anesthesiologists. Finally, we considered only large Western-style hospitals as opposed to outpatient clinics, which may experience different kinds of problems.

This study [1-3] was motivated by a number of accidents that had occurred shortly before, and had appeared in newspaper headlines and outraged the public. These accidents ranged from the case of an alcoholic who had been allowed to continue practicing when his problem was widely known among his colleagues, to that of a surgeon and an anesthesiologist who let a patient die while they engaged in a physical fight over whether or not to terminate the operation in the face of an emergency. Another key motivation for this work was the realization by hospital administrators that substance abuse could be more widespread than previously suspected, both among young residents using some of the drugs available in the operating room for recreational purposes, and among older practitioners abusing alcohol to relieve the stress of the profession. As we shall see, however, substance abuse, intolerable as it is in such a profession, is only a minor contributor to patient risk. Most of the risk to patients is rooted in much more mundane problems, such as insufficient supervision of residents or lack of periodic training among practitioners who may have been out of school too long to remember how to react in the face of rare events.

Our analysis is based on the techniques of engineering risk analysis, combining systems analysis and Bayesian probability in what is often referred to as probabilistic risk assessment (PRA) (e.g., [4]). This technique has been used to analyze poorly known systems before much experience has been gathered, such as nuclear reactors [5] and space systems [6]. The analysis described here relies on a similar forward-looking probabilistic model of the different scenarios of anesthesia accidents. For probabilistic assessments, these classes of scenarios must be structured into a mutually exclusive, collectively exhaustive set. Therefore, we identify first the different incidents that can start an accident sequence (for example, an overdose of the anesthetic drug). We then consider the different events (detection, diagnosis, correction) that can follow the initiating event, and we perform a stochastic dynamic analysis of these accident sequences. For simplicity, we use a discrete Markov model to represent the evolution of the accident sequence [7]. Therefore, we must assess the probability of transition between each pair of states in any given time period. These states are characterized both by the state of the patient (e.g., deprived of oxygen because of a tube disconnect) and by the state of the "anesthesia system" (anesthesiologist, nurses, surgeon, etc.), representing individuals who may or may not have detected or diagnosed the problem. At the end of the analysis, the question is: what is the probability that the patient has recovered or has died (or experienced severe brain damage) given the time elapsed between the initiating event and the end of the accident sequence. In the case of malignant hyperthermia, this evolution can be very quick. Although the condition is rare, the patient can die quickly unless proper measures are taken: if deprived of oxygen, the patient may be fine after a minute or so but experience severe problems or die after two minutes or more.

We did not have much statistical data for the specific parameters of the model, for instance the probability that in any operation of the considered class, the anesthesiologist is severely fatigued. We did have two kinds of statistical data: the rate of accidents in the cases that we considered, which was estimated in the literature to be about one in ten thousand (10^{-4}) [e.g., 8], and the rate of initiating events per operation. The latter came from a data set gathered in Australia as part of the study called AIMS [9] in which researchers at an Adelaide hospital had gathered anonymous records after each operation of any incident that may have occurred. Therefore, we had a relatively reliable estimate of the rate of occurrence of the initiating events.

We needed to assess the rate of problems that the anesthesiologists might experience in a set of operations, and the effect of these problems on the chances that the anesthesiologists cause an incident (initiating event of an accident sequence) and the chances that the anesthesiologists detect, diagnose and correct the problem they caused. For these, we used expert opinions that were carefully gathered from a diverse group of experts from two countries (the U.S. and Australia). These involved anesthesiologists who were remarkably candid about the kind of problems that are common in the profession, and willing to assess their rates of occurrence. We also interviewed surgeons who had had the opportunity to observe anesthesiologists in their practice, operating room nurses, patient organizations and lawyers. Among them, the most helpful, perhaps, were operating room nurses, especially those with long experience in several large hospitals.

This data structure allowed us to start our computation with reliable statistics of accident initiators and to verify at the end, based on accident statistics, that the results involving the use of expert opinions were consistent with the observed rate of accidents. These statistics at both ends of the problem thus provided us with a "reality check" that gave us some degree of confidence in the computations.

We did not analyze the costs of the different measures considered. Some of these costs are theoretically negligible, such as the cost of proper supervision of residents, which should be enforced anyway. Some costs are relatively low, such as the cost of an annual medical checkup for practitioners. Other costs, such as for regular re-certification, are perhaps higher.

This analysis provides a blueprint for a different approach to the evaluation and correction of problems early, before vast amounts of data have been gathered. The analysis therefore has the potential to save a substantial number of lives. It is based on data that are relatively easy to collect *ex ante,* and in particular on experience that can be related in simple terms; but it does not require wild guesses of the global result, which would be too complex to imagine out of the blue. For example, rather than asking experts to guess directly the benefits of particular safety measures such as the reduction of time on duty, we asked for the relative frequencies of specific types of errors among young residents versus experienced practitioners.

25.2 THE RISK ANALYSIS MODEL EXTENDED TO HUMAN AND MANAGEMENT FACTORS FOR THE CASE OF ANESTHESIA PATIENT RISKS

As described above, the analysis involves three phases: 1) patient risk analysis based on identification of accident scenarios and on their probabilities; 2) development of a model of the problems that may be experienced by anesthesiologists, and of their effects on the parameters of the risk analysis and thus on our estimate of the corresponding patient risk; and 3) identification of potential safety measures and policies, and assessment of their effect on the probability distribution of the different types of problems that can be experienced by an anesthesiologist in any given operation. The third phase allows computation of the risk reduction benefits of these different policies.

The model is based on a general approach developed by the Stanford Engineering Risk Research Group called "System-Action-Management" or SAM. The principle is to analyze first the failure risks of a physical (e.g., an engineered) system, the effects of different decisions and actions of operators and technicians on the parameters and variables of the risk model, and the effects of management decisions and policies on the behavior of these agents [10, 11]. Although causality flows from management to the system's failure risk, the analysis starts with the system and extends to management factors. This allows us to target the computation to the relevant risk factors. This type of analysis has been applied to other problems including, for example, a study of the risks of an accident on offshore oil platforms [12].

25.2.1 The patient risk model

The structure of the patient risk model is based on the identification of the different initiating events and on the dynamic analysis of the event sequence that follows each event. The probability that a patient experiences an anesthesia accident *(AA),* whether death or brain damage, is a function of the probability of the initiating events and of the probability of an accident conditional on each initiating event. Letting *p(.)* be the probability of an event per operation and $p(X|Y)$ be the probability of event *X* conditional on event *Y,* the probability of an anesthesia accident per operation is thus:

$$p(AA) = \sum_i p(IE_i) \times p(AA \mid IE_i) \qquad (1)$$

The first term of equation (1) is the probability of an initiating event $p(IE_i)$. Table 25.1 summarizes the initiating events that were identified, their probabilities per operation as estimated through numerical databases as well as expert opinions, and their relative fractions among incidents.

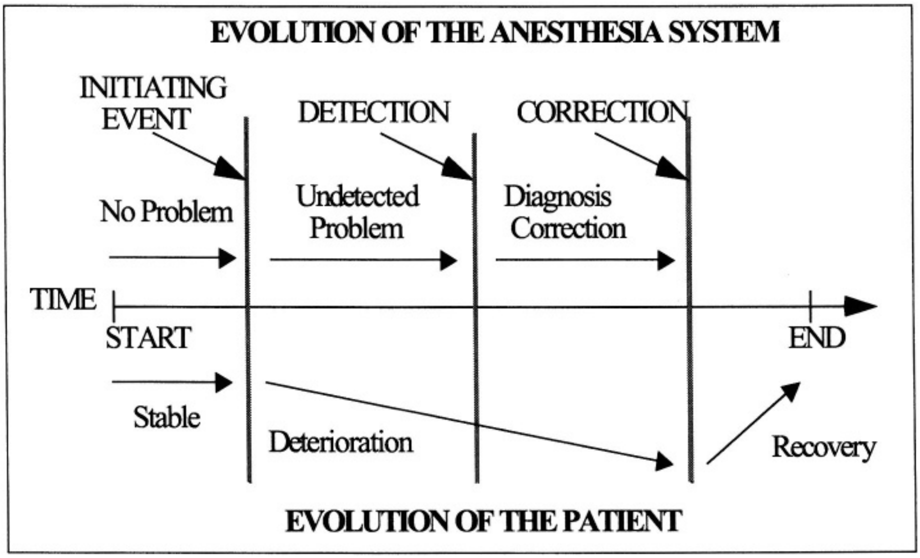
The second term of equation (1) is the probability of an anesthesia accident conditional on each initiating event, $p(AA/IE_i)$. We compute that term using a Markov model that involved transitions among "super states". Each super state combines one state of the patient and one state of the "anesthesia system". For example, one such super state could be the following: the patient is in a state of hypoxemia following a tube disconnect, *and* the "anesthesia system" has not yet detected the disconnection. Each state of the overall system is thus described both by a state of the patient and by a phase of the anesthesia system. Figure 25.1 shows the parallel evolution over time of both the anesthesia system (above the time axis) and the patient state (below that axis).

Table 25.1 Anesthesia accident initiators and their estimated probabilities (data)*

Initiating Event	Probability per operation	Fraction
Breathing Circuit Disconnect	7.2×10^{-4}	34%
Esophageal Intubation	2.6×10^{-4}	12%
Nonventilation	8.1×10^{-4}	38%
Malignant Hyperthermia	1.3×10^{-5}	1%
Anesthetic Overdose	1.8×10^{-4}	8%
Anaphylactic Reaction	1.2×10^{-4}	6%
Severe Hemorrhage	2.5×10^{-5}	1%

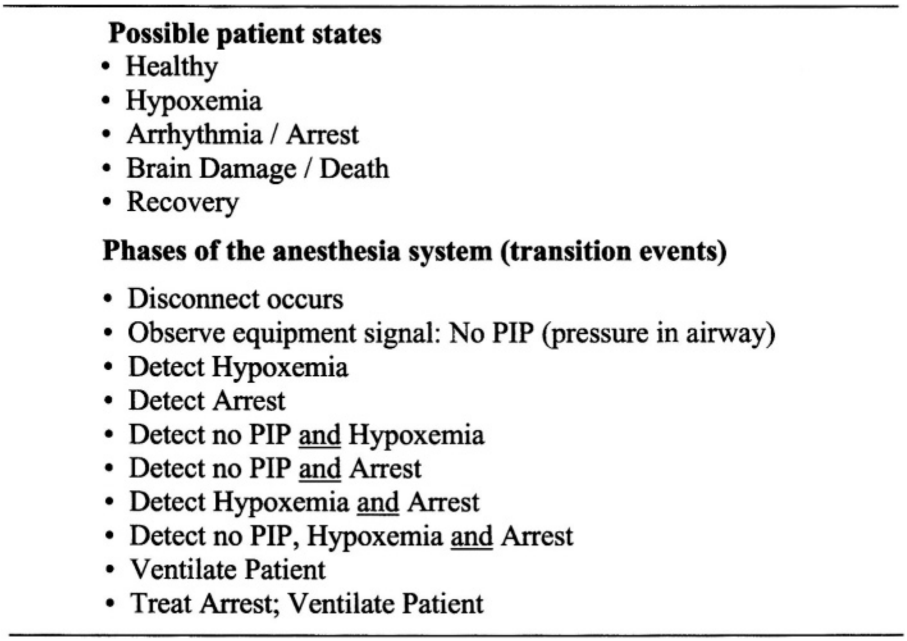
*Sources: [1-2]

Figure 25.1 Evolution of the patient state and of the anesthesia system following the occurrence of an accident initiator such as a tube disconnect*



*Sources: [1-2]

Table 25.2 Possible states of the patient and phases of the anesthesia system over time for the case of a tube disconnect*



Possible patient states

- Healthy
- Hypoxemia
- Arrhythmia / Arrest
- Brain Damage / Death
- Recovery

Phases of the anesthesia system (transition events)

- Disconnect occurs
- Observe equipment signal: No PIP (pressure in airway)
- Detect Hypoxemia
- Detect Arrest
- Detect no PIP and Hypoxemia
- Detect no PIP and Arrest
- Detect Hypoxemia and Arrest
- Detect no PIP, Hypoxemia and Arrest
- Ventilate Patient
- Treat Arrest; Ventilate Patient

*Sources: [1-2]

Table 25.2 summarizes the patient states and the phases of the anesthesia system that we considered, for the example case of a disconnection of the tube that brings oxygen to the patient's lungs.

The results of the analysis, for all possible initiating events, are summarized in Table 25.3, which shows the contribution of each of the possible accident initiators to the overall probability of an anesthesia accident.

Note that breathing problems (breathing circuit disconnect, esophageal intubation and nonventilation) and anesthesia drug-related problems (inhaled anesthetic overdose and anaphylactic reaction) contribute about equally to the overall risk, while rare events such as malignant hyperthermia contribute little to the risk even though they have a relatively high probability of causing a severe accident if they occur.

25.2.2 Analysis of the effect of the state of the anesthesiologist on the patient risk

In this second phase, we first assess the probability that in a given operation, the anesthesiologist experiences a particular type of problem, then the probability of an anesthesia accident conditional on the state of the

Table 25.3 Probability of an anesthesia accident given an initiating event and contribution of each initiator to the overall risk*

Initiating Event	$p(AA/IE_j)$	$p(IE_j)$	Contribution to Overall Risk
Breathing Circuit Disconnect	0.018	7.2×10^{-4}	18%
Esophageal Intubation	0.024	2.6×10^{-4}	9%
Nonventilation	0.021	8.1×10^{-4}	24%
Malignant Hyperthermia	0.16	1.3×10^{-5}	3%
Inhaled Anesthetic Overdose	0.037	1.8×10^{-4}	9%
Anaphylactic Reaction	0.22	1.2×10^{-4}	37%
Severe Hemorrhage	0.014	2.5×10^{-5}	<<1%

*Sources: [1-2]

anesthesiologist. The latter is shown in equation (2), in which SA_j represents the state of the anesthesiologist (presence or not of different problems indexed in j) and where the probabilities of initiating events and anesthesia accidents given these initiating events are also conditioned by this state.

$$p(AA/SA_j) = \sum_i p(IE_i/SA_j) \times p(AA/IE_i/SA_j) \qquad (2)$$

The anesthesiologist states SA_j affect the probability of an anesthesia accident in two ways. First, they increase the chances that a practitioner error causes a problem that starts an accident sequence (for example, that s/he administers an overdose of anesthetic drug to a patient). Second, they can decrease the probability that in due time; the practitioner observes, diagnoses and properly treats the problem that has occurred. For each type of problem, we asked our experts (ten anesthetists and a number of surgeons and operating room nurses) to assess the multipliers of the probabilities of each initiating event, and of the average time that it takes for the practitioner to detect, diagnose and properly treat the problem. Based on these data, we ran the model described in the previous section for each type of practitioner problem and computed the resulting probability of an anesthesia accident.

Table 25.4 displays the nature and the probabilities of the different anesthesiologist problems considered here, and their effects on the risk of an anesthesia accident per operation. Table 25.4 shows, for example, that severe distraction and fatigue can seriously affect the performance of the anesthesiologist and increase by a factor of ten the probability of an accident compared to that in the case of a problem-free practitioner. Note, of course, that the mean probability of such an accident remains unchanged at about $7x10^{-5}$ since the previous results were based on the general population of anesthetists, including by definition those who experienced such problems.

Table 25.4 Performance shaping factors among anesthesiologists, probabilities of different problems per operation and resulting probability of an anesthesia accident*

Anesthesiologist State (SA_j)	**p(SA_j)**	**p(AA| SA_j)**
Problem-free	0.53	6.7×10^{-6}
Fatigue	0.10	8.6×10^{-5}
Cognitive Problems	0.04	8.6×10^{-5}
Personality Problems	0.04	8.6×10^{-5}
Severe Distraction	0.03	2.0×10^{-4}
Drug Abuse	0.03	1.0×10^{-4}
Alcohol Abuse	0.04	1.0×10^{-4}
Aging/Neurological Problems	0.03	1.0×10^{-4}
Lack of Training	0.12	1.3×10^{-4}
Lack of Supervision	0.04	4.9×10^{-4}

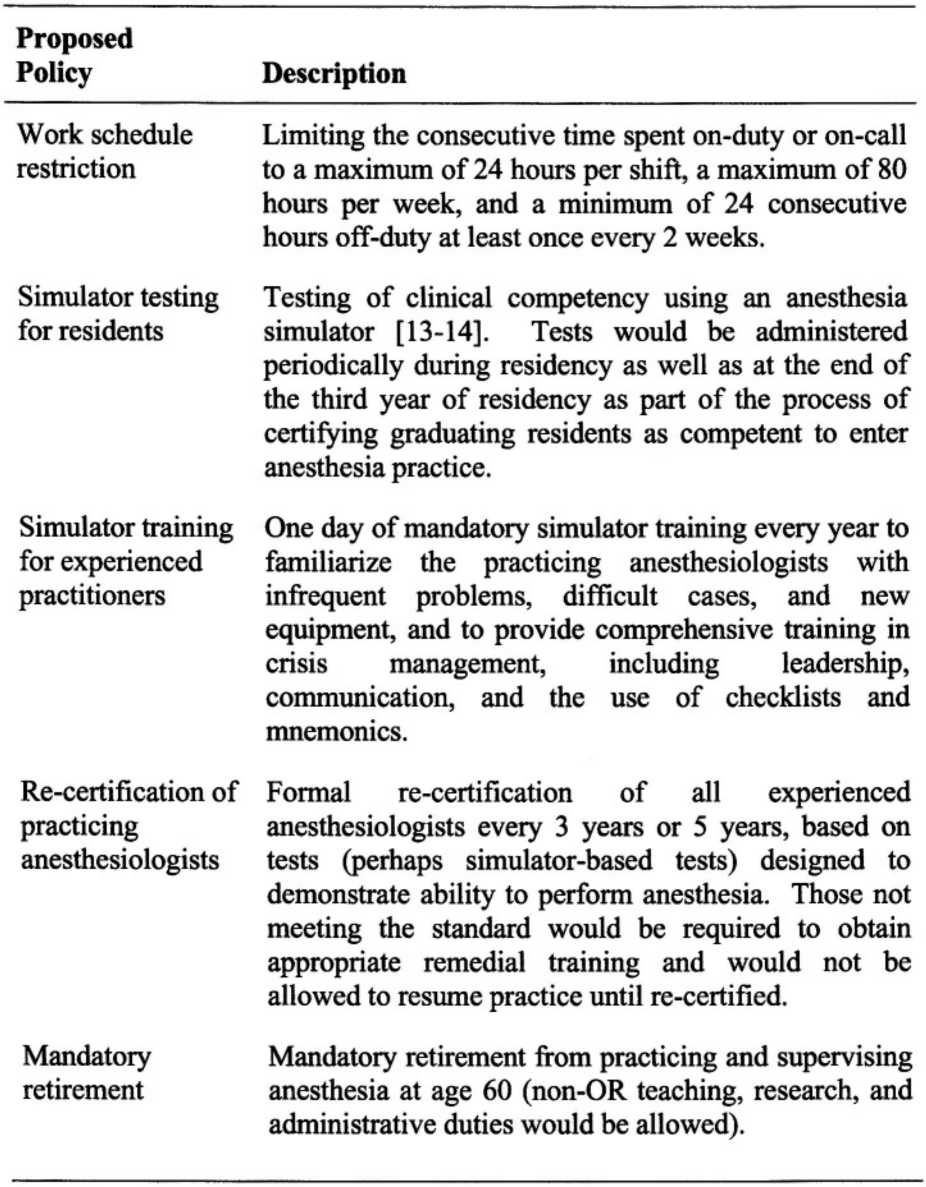
*Sources: [1-2]

25.2.3 Risk-reduction benefits of some management measures

The state of the anesthesiologist may be affected by workplace policies. For example, an anesthetist's ability to perform his or her job may be affected when s/he is assigned to work long hours, when there is little monitoring of anesthetists' general state of health, or when there is insufficient supervision of residents and trainees. Insufficient supervision is not supposed to occur when established standards are followed. In practice, however, the supervisor who is expected to be available in the operating room within two minutes should an incident occur may be 15 minutes away in his or her office.

We considered a number of policy changes that could be made to improve anesthesiologist states. Table 25.5 shows a description of the policies that we examined.

Table 25.5 Descriptions of policy changes evaluated*



Proposed Policy	Description
Work schedule restriction	Limiting the consecutive time spent on-duty or on-call to a maximum of 24 hours per shift, a maximum of 80 hours per week, and a minimum of 24 consecutive hours off-duty at least once every 2 weeks.
Simulator testing for residents	Testing of clinical competency using an anesthesia simulator [13-14]. Tests would be administered periodically during residency as well as at the end of the third year of residency as part of the process of certifying graduating residents as competent to enter anesthesia practice.
Simulator training for experienced practitioners	One day of mandatory simulator training every year to familiarize the practicing anesthesiologists with infrequent problems, difficult cases, and new equipment, and to provide comprehensive training in crisis management, including leadership, communication, and the use of checklists and mnemonics.
Re-certification of practicing anesthesiologists	Formal re-certification of all experienced anesthesiologists every 3 years or 5 years, based on tests (perhaps simulator-based tests) designed to demonstrate ability to perform anesthesia. Those not meeting the standard would be required to obtain appropriate remedial training and would not be allowed to resume practice until re-certified.
Mandatory retirement	Mandatory retirement from practicing and supervising anesthesia at age 60 (non-OR teaching, research, and administrative duties would be allowed).

Table 25.5 (cont.) Descriptions of policy changes evaluated*

Proposed Policy	Description
Drug testing	Monthly random testing for drug abuse. Anyone identified and confirmed to have a problem (through appropriate follow-up testing) would not be allowed to return to practice until he/she had undergone treatment, demonstrated that the problem had been satisfactorily resolved, and met appropriate re-entry and follow-up criteria.
Alcohol testing	Monthly random testing for alcohol abuse, similar to drug testing described above.
Annual medical examination	Mandatory annual medical examination; may identify problems such as substance abuse, age-related performance deterioration, chronic fatigue, etc.
Supervision of residents	Strict supervision rules for residents through all 3 years of residency: a supervisor is required to be available in the OR in less than two minutes at all times during an operation. The supervisor is expected to intervene personally if there is any question of patient safety, and residents are instructed to contact the supervisor sooner rather than later.

*Sources: [1-2]

We assessed the potential effect of each measure on the probability per operation that the practitioner experiences a particular type of problem (and the probability that the problem is eliminated). We thus reassessed, for each measure, the probability distribution of the different types of anesthetist problems ($p(SA_j)$). These new probabilities, in turn, decrease patient risk by increasing the chances that the anesthesiologist is problem-free. Letting M_k denote the different management measures that were considered, the probability of an anesthesia accident per operation, given that management measure M_k has been implemented, can be computed as follows:

$$p(AA \mid M_k) = \sum_j p(SA_j \mid M_k) \sum_i p(IE_i \mid SA_j) \times p(AA \mid IE_i, SA_j) \qquad (3)$$

Table 25.6 displays the risk reduction benefits that were computed for each policy. In some cases we assumed that the policy transferred the anesthesiologists who were originally afflicted by a specified problem into the "problem-free" category. For example, we assumed that periodic re-certification and simulator training would give experienced practitioners who did not operate often enough, a chance to encounter rare problems on a machine before doing so in the operating room.

The results shown in Table 25.6 were somewhat surprising. Our study had been motivated in large part by the fear of substance abuse among both the youngest and the oldest anesthesiologists, but policies whose goal is to detect them are not among the most effective. This is true for two reasons: the base rate of substance abuse is rather low, and the effectiveness of the potential measures is questionable because anesthesiologists may be able, with training, to escape detection.

Table 25.6 Effects of proposed policy changes on the anesthesia patient risk*

Policy	Effects of Policy	Replacement	Risk with Policy ($\times 10^{-5}$)	Risk Reduction (%)
Base Case (current policies)		--	7.12	--
Work schedule restriction	Fatigue cut 50%	Problem-free	6.72	6
Simulator testing for residents	Cognitive problems cut 90% Personality problems cut 50%	New dist'n	7.02	2
Simulator training for practitioners	Lack of training cut 75%	Problem-free	5.98	16

Table 25.6 (cont.) Effects of proposed policy changes on the anesthesia patient risk*

Policy	Effects of Policy	Replacement	Risk with Policy (x 10^{-5})	Risk Reduction (%)
Re-certification every 3 years	Decreases lack of training, aging, cognitive, personality problems For 10 re-certs: 84% reduction	Problem-free	5.06	29
Re-certification every 5 years	Decreases lack of training, aging, cognitive, personality problems For 6 re-certs: 67% reduction	Problem-free	5.48	23
Mandatory retirement	Affects 10% of operations: Aging, lack of training, alcohol abuse more heavily weighted	New dist'n	6.89	3
Drug testing	Drug abuse cut 95%	New dist'n	7.03	1
Alcohol testing	Alcohol abuse cut 90%	New dist'n	6.97	2
Annual medical examination	Aging/neurological problems cut 75% Drug, alcohol abuse cut 25% Fatigue cut 10%	New dist'n**	6.92	3
Supervision of residents	Lack of supervision cut 50%	Problem-free	6.16	14

*Sources: [1-2]

**Except Fatigued replaced by Problem-free. New dist'n: new distribution of the probability of problems among practitioners with increase in the probability of "problem-free".

We found that the real problems were closer to home than expected. Most of the problems were not caused by rare events, but by more mundane factors such as fatigue and poor supervision of residents. Closer supervision of residents and periodic re-certification and simulator training appeared to be among the most potentially effective measures.

We did not include any cost considerations in our analyses. Some of the costs (e.g., cost of re-certification) depend on how practitioners' time is valued. Practitioners will tend to value their time at the rate at which they are paid in the operating room, whereas some economists would value their time at the cost that they are willing to pay for their leisure time.

This analysis shows that the base rate of anesthetists' problems, their effects on patient safety, and the anticipated decrease of problem probabilities as a result of safety measures, all contribute to the risk-reduction benefits. The analysis allowed identification of the most severe problems and of the most effective measures, regardless of the nature of the last accident and the publicity that may have surrounded it.

25.3 CONCLUSIONS

When the costs and the benefits of possible safety measures are easy to assess and priorities are clear, risk quantification may not be necessary. This is not always true, however, as intuition and perceptions can be deceptive. Perceptions of risk can be distorted by media reports, or by an emphasis on the sensational rather than the obscure. When abundant statistics are available, they may be sufficient for computing the corresponding risks. Otherwise, probabilistic methods based on systems analysis, Bayesian probability and dynamic stochastic models can be used to assess the risk based on a mix of statistics, physical models and expert opinion. We have illustrated that approach in this chapter.

This type of analysis provides a logical framework in which to gather the available knowledge when decisions need to be made before large data sets can be gathered and perfect information obtained. The anesthesia patient example provided here showed that the accident sequences that had received the most attention because they had made the headlines were not the largest contributors to the overall patient risk. Although a full study should also include the costs (which are likely to vary widely according to circumstances), we have shown in this chapter how the probabilistic risk analysis framework can be used to set priorities among different policies, based here on their risk reduction benefits.

What was done here for trained anesthesiologists could be transferred directly to a study of the performance of nurse anesthesiologists. It would be interesting to know, given the difference of costs, whether there is a significant difference or not in the risk to the patients. The model could also be applied to other medical cases. What was done for anesthesia could also be done for surgery in general, or for particular types of operations. The performance of surgeons, like that of anesthesiologists, is affected by factors that include fatigue, time pressures, sometimes substance abuse, or simply situations for which they do not have the necessary training. In a different vein, the same Stanford group is currently working on a method for early assessment of medical devices, in the design phase before large statistical studies can be performed to satisfy FDA requirements [15]. The method is based on risk analysis including a system's analysis not only of the device but also of the performance of the people who will use it in the future. It allows computing the risks of failure before large amounts are spent or the device fails in operation after FDA approval.

Acknowledgments

The original study was funded by the Anesthesia Patient Safety Foundation, Cambridge, Massachusetts, and performed at Stanford University, by a team including Linda Lakats (then, Ph.D. student), Dean Murphy (then, Ph.D. student), and Professor Elisabeth Paté-Cornell from the Department of Industrial Engineering and Engineering Management, and Professor David Gaba from the Department of Anesthesiology in the School of Medicine.

References

[1] Paté-Cornell, M.E., D.M. Murphy, L.M. Lakats and D.M. Gaba (1996). Patient risk in anesthesia: Probabilistic risk analysis, management effects and improvements. *Annals of Operations Research,* 67, 211-233.

[2] Paté-Cornell, M.E., L.M. Lakats, D.M. Murphy, and D.M. Gaba (1997). Anesthesia patient risk: A quantitative approach to organizational factors and risk management options. *Risk Analysis,* 17, 511-523.

[3] Paté-Cornell, M.E. (1999). Medical application of engineering risk analysis and anesthesia patient risk illustration. *American Journal of Therapeutics,* 6, 245-255.

[4] Kaplan, S. and B.J. Garrick (1981). On the quantitative definition of risk. *Risk Analysis,* 1, 11-27.

[5] US Nuclear Regulatory Commission (1975). *The Reactor Safety Study.* WASH-1400, Washington, DC.

[6] Paté-Cornell, M.E. and P.S. Fischbeck (1994). Risk management for the tiles of the space shuttle. *Interfaces,* 24, 64-86.

[7] Hillier, F.S. and G.J. Lieberman (1990). *Introduction to Operations Research.* McGraw-Hill, New York.

[8] Davies, J. M., and L. Strunin (1984). Anesthesia in 1984: How safe is it? *Canadian Medical Association Journal,* 131, 437-441.

[9] Runciman, W.B., A. Sellen, R.K. Webb, J.A. Williamson, M. Currie, C. Morgan, and W.J. Russell (1993). Errors, incidents and accidents in anaesthetic practice. *Anaesthesia and Intensive Care,* 21, 506-519.

[10] Murphy, D.M. and M.E. Paté-Cornell (1996). The SAM framework: A systems analysis approach to modeling the effects of management on human behavior in risk analysis. *Risk Analysis,* 16, 501-515.

[11] Paté-Cornell, M.E. and D.M. Murphy (1996). Human and management factors in probabilistic risk analysis: the SAM approach and observations from recent applications. *Reliability Engineering and System Safety,* 53, 115-126.

[12] Paté-Cornel1, M.E. (1990). Organizational aspects of engineering system safety: the case of offshore platforms. *Science,* 250, 1210-1217.

[13] Gaba, D.M. (1992). Improving anesthesiologists' performance by simulating reality. *Anesthesiology,* 76, 491-494.

[14] Gaba, D.M. (1994). Human work environment and simulators. In Miller, R.D., Ed., *Anesthesia.* Churchill Livingstone, New York, Chapter 85.

[15] Pietzsch, J.B., T.M. Krummel and M.E. Paté-Cornell (2002). Early technology assessment of new medical devices and procedures: A systems analysis approach using probabilistic modeling. *Proceedings of ISTAHC2002,* 18th Annual Meeting of the International Society of Technology Assessment in Health Care, Berlin, Germany, June 2002.

26 AN ASTHMA POLICY MODEL

A. David Paltiel[1], Karen M. Kuntz[2], Scott T. Weiss[3] and Anne L. Fuhlbrigge[3]

[1] Department of Epidemiology and Public Health
Yale School of Medicine
New Haven, CT 06520

[2] Department of Health Policy and Management
Harvard School of Public Health
Boston, MA 02108

[3] Channing Laboratory, Department of Medicine
Harvard Medical School
Brigham and Women's Hospital
Boston, MA 02115

SUMMARY

Despite many proven advances in patient care over the last 10 years, asthma continues to impose a large and growing burden on society. Persistent clinician non-adherence to recommended practice is well documented, but little is known about the clinical impact and economic costs of alternative approaches to asthma patient care. In this chapter, we introduce the Asthma Policy Model, a state-transition simulation that we have developed to forecast asthma-related symptoms, acute exacerbations, quality-adjusted life expectancy, health care costs, and cost-effectiveness. We begin with a detailed survey of the epidemiological, clinical, and policy context that motivates our work. With a modeling audience in mind, we then describe the considerations that produced the current analytic structure and input datasets. We illustrate the policy relevance of the model by describing our recent work on the cost-effectiveness of inhaled corticosteroid therapy in a population of adult patients with mild-to-moderate disease. We close the chapter with a discussion of plans for future refinements and applications.

KEY WORDS

Asthma, Inhaled corticosteroids, Cost-effectiveness analysis, Decision analysis, Markov models, State transition models

26.1 INTRODUCTION

Over the last 10 years, the scientific basis for effective therapeutic intervention in asthma has been established. The challenge now is to translate this evidence base into long-term clinical outcomes and economic terms so that it can be understood by decision makers whose choices determine the financing and delivery of patient care. The policy modeling approach to this question is justified by the immediacy and uncertainty of the context in which asthma-related decisions are made. In this section, we briefly describe that context. We begin by describing the growing public health burden imposed by asthma (26.1.1) and inventorying recent developments in patient care (26.1.2). We then turn to a discussion of practice variation in asthma management and the failure of high-profile guidelines to promote a consistent standard of care (26.1.3). We note the lack of program evaluation research in asthma (26.1.4) and argue that a policy modeling approach is both appropriate and necessary.

26.1.1 The growing social burden of asthma

Asthma is one of the most prevalent diseases of children and adults in the United States, affecting approximately 13.7 million Americans [1]. Temporal trends in hospitalization rates indicate that asthma morbidity is increasing. Among children, asthma is the most frequent cause of hospitalization and school absences and is estimated to account for 28 million restricted activity days annually [2]. Pediatric hospitalization rates have increased by approximately 300 percent over the past 20 years [3]. Hospitalization rates among adults have also increased, rising from 10 in 10,000 in 1980 to 13 in 10,000 in 1985 [4]. Although death from asthma is relatively uncommon, the mortality rate in the U.S. has gradually increased over the past 15 years among both children and adults [5].

Asthma also imposes a large economic burden. In 1992, Weiss and colleagues [6] estimated $3.6 billion in total direct costs in the U.S., a figure that included hospitalization and emergency department (ED) services ($1.9 billion), prescription medications ($1 billion), and physician services ($493 million). The indirect costs of lost income and productivity were found to contribute an additional $2.6 billion to the societal bill. More recent figures from the National Heart, Lung, and Blood Institute show direct costs for the year 2000 of $8.1 billion, with lost productivity costing an additional $4.6 billion [7].

The impact of asthma is felt disproportionately in historically underserved communities in the U.S. Asthma mortality among U.S. blacks is approximately twice that of U.S. whites. The rate of hospitalization for asthma is also substantially greater among U.S. blacks and Hispanics overall than among whites. These differences are seen in all age groups, but are especially striking among children. National data for 1987 reveal that black children less than 5 years old were almost three times as likely as white children to be hospitalized for asthma [4, 8, 9]. Similar disparities are observed for the very poor and those living in the inner city [8, 9]. The etiologies for the higher disease prevalence and greater rates of health care utilization, morbidity, and mortality among disadvantaged populations are not fully established. Postulates include increased exposure to smoke and allergens [10-12], reduced access to health care [6, 13], and inappropriate disease management [14-18].

These figures depict asthma as a large and growing public health problem. They also highlight a common misperception about the disease: Although asthma is generally thought of as a low-grade, chronic ailment, the numbers reveal a condition in which 43% of the costs are related to hospitalization and ED services. This suggests that there is room for improvement in patient management. It also speaks to the potential usefulness of formal evaluation in promoting better resource allocation and more efficient, more equitable patient care.

26.1.2 Rapid emergence of new treatment alternatives

Insights into the central roles of airway inflammation and hyperresponsiveness in asthma have led to a marked change in recommended preventive measures and therapies in the last 15 years [19]. Today, the goals of asthma management include: prevention of chronic symptoms, maintenance of near-normal pulmonary function and activity levels, and prevention of recurrent exacerbations while minimizing adverse drug effects. While pharmacologic therapy is the central component of asthma management, several non-pharmacologic approaches also contribute to improved outcomes.

Pharmacologic interventions "Quick relief medications (most notably, short-acting beta-agonists) remain the most commonly prescribed asthma therapy. As stand-alone monotherapy, however, they are generally believed to be appropriate only for those with the mildest form of disease. Persons with persistent disease require the addition of "controller" medications, with inhaled corticosteroids (ICS) forming the mainstay of anti-inflammatory therapy [20]. Newer, longer-acting beta-agonists (duration of action in excess of 12 hours) are in the early stages of distribution [21]. Other agents (leukotriene modifiers, IgE, Anti IL-4, Anti IL-5) are under investigation [22].

Non-pharmacologic interventions Patient education and self-management interventions aim to promote better understanding of the nature of disease, environmental modification, compliance with prescribed medications, early recognition and response to acute events, appropriate use of devices (such as metered dose inhalers), enhancement of patient/family psychological resources, and promotion of coping strategies. Asthma management programs seek to promote greater provider and patient attention to the process of care and thereby to improve patterns of inpatient service and rates of ICS prescribing [23]. Other interventions aim at reducing exposure to environmental risk factors and specific allergens [24-27]. Finally, treatment of coexisting, upper respiratory symptoms (e.g., with intranasal steroids and oral antihistamines) has been shown to produce short-term improvements in lung function and symptoms [28-31].

The expansion of treatment choices in asthma carries with it a host of uncertainties about what is effective, what is not effective, and at what cost. In time, randomized controlled trials and clinical experience may resolve some of these questions. However, the decision to wait for better information is a decision that itself carries costs and consequences. For choices that cannot wait, a model-based approach offers a formal framework for organizing information.

26.1.3 Limited success of national practice guidelines

The National Asthma Education and Prevention Program (NAEPP) Guidelines for the Diagnosis and Management of Asthma [1] were developed to help bridge the gap between research and clinical practice. The guidelines describe four components of asthma management: 1) use of objective measures of lung function to assess severity and monitor the course of disease; 2) pharmacologic therapy focused on long-term management of airway inflammation; 3) non-pharmacologic measures to diminish or eliminate factors that precipitate asthma symptoms; and 4) patient education. The guidelines recommend a "step" approach to asthma therapy based on disease severity, emphasizing the need to manage persistent disease more aggressively with consideration to earlier use of anti-inflammatory therapy.

Clinical guidelines in other diseases have been shown to decrease the variance between physician practice patterns and accepted standards [32-34]. However, judging by the persistent variation in asthma practice patterns in the U.S., nonadherence to the NAEPP guidelines is widespread [35]. This raises the concern that increased morbidity and cost related to inadequate asthma care may soon cause payors to demand the delivery of guideline-concordant care. It also speaks directly to the policy relevance of a model-based approach in evaluating the clinical and economic costs that society incurs by failing to provide universal, comprehensive, state-of-the-art care.

26.1.4 Limited program evaluation literature in asthma

Relatively few studies have examined the costs or cost-effectiveness of strategies to improve asthma care and most of these have focused on educational interventions. Windsor et al. examined a range of alternative adherence measures, including correct inhaler use, medication adherence and inhaler adherence in a randomized controlled trial [36]. They found a 42% improvement in adherence in the intervention group compared to the control group at a total cost of $32.03/per patient/year. Other studies have assessed asthma self management [37, 38], childhood asthma management [39, 40], and medication use [41, 42]. Ross et al. considered the incremental costs of treating asthma in patients for whom cromolyn sodium was included in the routine treatment plan [43]. Rutten-van-Molken et al. [44] and Connett et al. [45] studied the cost-effectiveness of inhaled corticosteroids but came to conflicting conclusions. In a pediatric population, Connett et al. found that budesonide produced a favorable clinical response, increasing symptom-free days and reducing overall costs. In adults, Rutten-van-Molken et al. found that the addition of an inhaled corticosteroid to beta-agonist therapy produced an incremental cost-effectiveness ratio of $5 per symptom-free day gained.

No studies have specifically evaluated the NAEPP guidelines. No studies have been conducted in conformity with the 1996 recommendations of the U.S. Panel on Cost-Effectiveness in Health and Medicine [46].

Asthma may be the only major chronic condition for which independent pharmacoeconomic evaluations have yet to appear in the literature with regularity. This places patients with asthma at a distinct disadvantage in competing for scarce health care resources. Use of a policy modeling approach aims to meet that challenge by producing objective, quantifiable evidence of: 1) the social costs of clinician nonadherence to recommended practice; and 2) the comparative cost-effectiveness of increased investment in state-of-the-art asthma patient care.

26.2 DESCRIPTION OF APPLICATION

We have developed the Asthma Policy Model [47-50], a state-transition simulation of the natural history and clinical management of asthma. We have used this model to support clinical judgment in asthma by producing literature-based estimates of health outcomes, quality-of-life effects, economic costs, and the cost-effectiveness of different patient care strategies. In the sections that follow, we describe the structure of the model and illustrate its usefulness with an application to the cost-effectiveness of inhaled corticosteroid (ICS) therapy in patients with mild-to-moderate disease [50].

To briefly summarize the application, we compare two intervention strategies: 1) quick relievers (e.g., short-acting β-agonists) on an as-needed basis, versus 2) quick relievers plus ICS therapy. We employ a secondary research design to conduct clinical effectiveness and pharmacoeconomic analyses where the outcomes of interest include: symptoms (both daytime and nocturnal), acute event rates (including events requiring an urgent care visit, an ED visit, and/or hospitalization), and mortality (both asthma-related and other). We also record time spent at a given disease severity level. Patients are followed from age 18 to death, thereby permitting us to consider events along the spectrum of adult disease. In keeping with accepted methods for cost-effectiveness analysis, the baseline analysis adopts a societal perspective, considering all economic costs and consequences to be important regardless of their source or beneficiary. We report value for money on an incremental basis, measured in terms of both dollars per quality-adjusted life-year (QALY) gained [46] and dollars per symptom-free day gained [51]. We discount all outcomes at 3% per annum. Monetary values are reported in 1998 U.S. dollars, adjusted (when necessary) using the medical care component of the Consumer Price Index [52].

26.3 METHODOLOGY USED

26.3.1 Model overview

The Asthma Policy Model is a Markov, state-transition simulation of the natural history of asthma in a general patient population. It characterizes the natural history of illness as a sequence of flows into and out of a defined set of "health states." Patients in a given health state are assumed to share a similar clinical history and prognosis, a common perception of well-being, and a comparable pattern of health care utilization [53, 54]. At any point in time, a patient is assigned to one and only one health state. The model classifies the health states into three general categories (chronic, acute, and death), reflecting the observation that the clinical course of asthma is characterized by long periods of chronic illness, punctuated by episodic, acute exacerbations. Deaths can occur from either the acute or chronic health states.

Patients are assigned to an initial health state in proportions that reflect published distributions of patient age, asthma prevalence, and pulmonary function. Each month, a patient's risk profile may change, thus producing a re-assignment to a new health state. For example, at the beginning of each month, the model adds a month to the age of every member of the population;

patients who turn 35 advance from the 18-35 age stratum to the >35 category. State transitions may also result from progression to a more severe level of pulmonary dysfunction or death. Acute exacerbations also produce transitions to a new health state; the model identifies three categories of acute events — urgent care visits, ED visits, and hospitalizations — reflecting the observation that each of these involves a unique set of mortality risks, clinical consequences, quality-of-life reductions, and economic costs. The acute health states are modeled as single-cycle, transient states, meaning that patients reside in them for exactly one month before either returning to a chronic state of health or dying. This modeling convenience makes it possible to capture the rapid succession of events, the increased risks of death, the reduction in quality of life, and the increased costs that characterize the month immediately following the onset of an acute event.

26.3.2 State space

The state space in the Asthma Policy Model is defined along the following clinical dimensions:

- **Disease status** (chronic/stable; acute/hospitalization; death): Patients spend the majority of their time in a chronic, stable state of health. Acute events typically involve short bursts of intense resource consumption and transient reductions in patient perceptions of well-being. The model captures this by distinguishing between time spent in a stable state of illness and time spent in the hospital. (For computational purposes, it is also useful to define a "death" state.)

- **Lung function** (mild, moderate, severe): Airflow impairment is an imperfect but valuable predictor of patient prognosis, risk of exacerbation, quality of life, and resource use. Based on the NAEPP recommendations [1], we define three strata of lung dysfunction, using the forced expiratory volume in one second, expressed as a percent of predicted normal value (FEV1% predicted), as our severity measure: <60%, 60-80%, and >80%. Readers should note that this is a critical – and controversial – modeling assumption, one which we discuss further in Section 3.3.

- **Patient age** (18-35; >35): The current version of the model is focused on adults. The 35-year adult cut-point is chosen to reflect differences in background mortality rates between younger and older adults and to minimize the risk of misclassification of asthma and chronic obstructive pulmonary disease (COPD) among younger adults.

- **Prior hospitalizations** (0, 1, >1): A history of asthma-related hospitalizations leads to higher acute event and resource consumption rates, reduced quality of life and survival, and different clinical care patterns [55]. The decision to lump all prior hospitalizations in excess of one into a single category reflects the lack of published data linking specific numbers of hospitalizations to events rates and resource consumption. It also reflects a concern that the state space not be expanded beyond the ability of the data to support that expansion.

- **Cause of death** (non-asthma-related; asthma-related): The model records cause of death in order to reflect resource consumption patterns in the last periods of life.

Together, these dimensions define a total of 32 logically feasible health states. Ideally, of course, it would be desirable to add many more dimensions to the health state definitions. However, data collection burdens and computational complexities increase with dimensionality; a model that distinguishes among many patient types might appear more precise, but its predictive validity and credibility would be much more difficult to establish.

26.3.3 Note on modeling strategy: Measuring disease severity

In this section, we provide more detail on what was perhaps the most challenging methodological task associated with the model development – namely, the specification of an index of lung dysfunction. No formally accepted definition of asthma severity exists. Most indices incorporate multiple characteristics including asthma symptoms, objective measures of lung function and airflow impairment, medication and resource utilization [56]. The 1991 NAEPP Guidelines classified asthma severity into three categories (mild, moderate and severe) based on the frequency and severity of asthma symptoms, exacerbations and objective measures of lung function [57]. The 1997 Guidelines modified the severity classification to include four categories; mild intermittent, mild persistent, moderate persistent, and severe persistent [1].

In the context of policy modeling, the goal of disease severity classification is less to describe "what is" and more to inform "what ought to be." The challenge, therefore, is to define a severity measure that assembles a sufficient amount of clinical information and is adequately predictive of the downstream outcomes of interest but – at the same time – remains independent of those outcomes and amenable to policy intervention. With these requirements in mind, we explored FEV1% predicted as a potential asthma severity index. At first blush, FEV1% predicted has the advantage of objectivity and reproducibility [58, 59]. Moreover, it appeals to the intuition as an underlying, physiologic driver of clinical events. In addition, asthma intervention trials frequently measure FEV1% predicted as an outcome and express the efficacy of new therapies in terms of impact on FEV1% predicted. Nevertheless, there is concern that FEV1% predicted might not be sufficiently predictive of prognosis. We sought to explore this concern, conducting our own analysis of the relationship of FEV1% predicted to symptoms, acute exacerbations, costs, and quality of life. We summarize our findings from that exploration in the sub-sections that follow [47-49, 60].

Relationship of FEV1% predicted and asthma symptoms The most distinctive characteristic of asthma is a pattern of airway irritability that increases when the disease is active and decreases in response to appropriate therapy. An obvious clinical measure of the disease process is the presence of symptoms (both daytime and nocturnal). One approach to symptom measurement involves the concept of the "symptom-free day." This global measure of the number of days for which the patient has had no symptoms (including cough, wheeze, shortness of breath, or nighttime awakening) is endorsed for use in economic evaluations of asthma interventions by the National Asthma Education and Prevention Program Working Group Report on the Cost-effectiveness of Asthma Care [51].

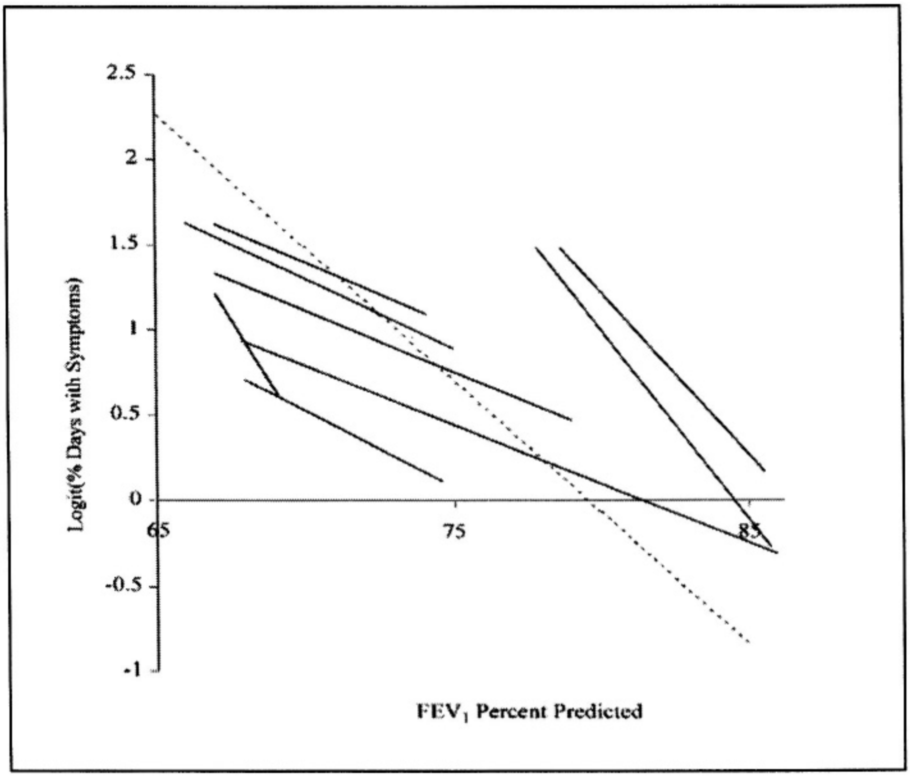
We reviewed the literature and developed a novel approach to estimating the impact of changes in FEV1% predicted on asthma-related symptoms [47]. The analysis used asthma-related clinical trials that reported estimates of mean FEV1% predicted and symptoms (symptom score or percentage of symptom days or nighttime awakenings). Using average baseline values from each study in weighted linear regression analyses, a negative association was found between lung function and symptom score (p<0.001) and the percentage of nighttime awakenings (p=0.18), but no association was found between lung function and symptom-days.

Consistent associations were identified between mean changes in lung function and symptom-days at follow-up within the studies. Accordingly, we plotted the change in FEV1% predicted versus the change in the logit of % of days with symptoms for each treatment arm, where each treatment arm was represented by a single line (see Figure 26.1). The weighted (by number of patients) average of the slopes of the lines was –0.1550. We estimated an intercept term by minimizing the vertical distances between the lines determined by each intervention and the fitted line. The resulting relationship is:

$$\textbf{\% symptom-days} = 1/(1+\exp(-12.5 + 0.1550 * \text{FEV1\% predicted})) * 100$$

This regression equation was applied in the model to translate a given increase in FEV1% predicted into a measurable reduction in the percentage of days with symptoms.

Figure 26.1 Relationship between FEV1% predicted and percentage of days with symptoms. Each solid line represents the intervention arm of a clinical trial. The dotted line represents the fitted line from the within-population analysis:
logit(% days with symptoms) = 12.5 - 0.1550 * FEV1 % predicted



Reprinted from Kuntz K.M., B.T. Kitch, A.L. Fuhlbrigge, A.D. Paltiel, P.J. Neumann, and S.T. Weiss (2001). A novel approach to defining the relationship between lung function and symptom status in asthma. *Journal of Clinical Epidemiology* 55, 11-18, with permission from Elsevier Science.

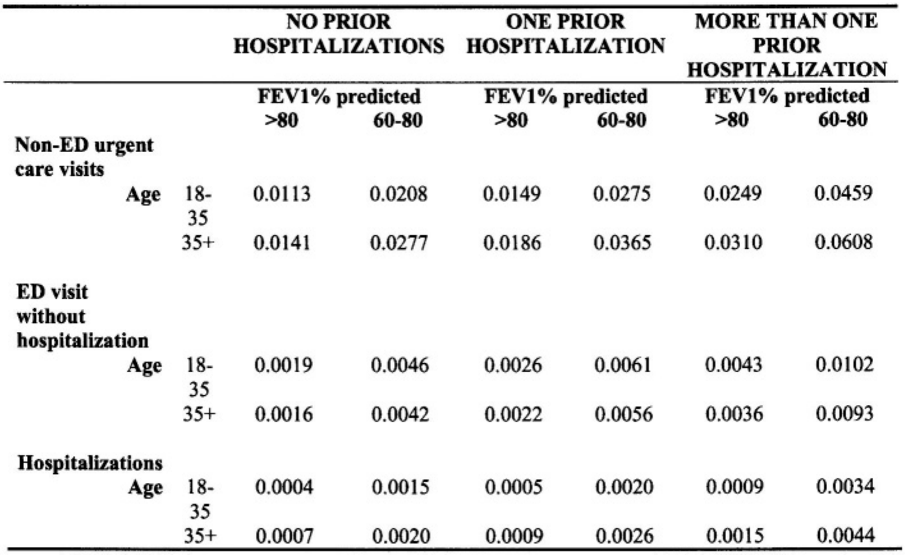
Relationship of FEV1% predicted and acute events We performed separate analyses in two adult and one pediatric cohort to explore the relationship between FEV1% predicted and reports of attacks of wheezing and shortness of breath over the subsequent year [48]. In the pediatric cohort, a progressive decrease in the proportion of individuals reporting an attack was associated with improved lung function. In multivariate models, FEV1% predicted was an independent predictor of attacks: among a parental report group, OR=2.1 (95%CI (1.3, 3.4)) and OR=1.4 (95%CI (1.2,1.6)) for FEV1% predicted <60% and 60%-80% compared with >80%, respectively. Among a self-report group, OR= 5.3 (95%CI (2.2,12.9) and OR=1.4 (95%CI (1.2,1.7) for FEV1% predicted <60% and 60%-80% compared with >80%, respectively. In the two adult cohorts, a similar relationship was observed between FEV1% predicted and risk of an asthma exacerbation over the three years following its measurement. Among subjects in one cohort with an FEV1% predicted of <60%, 58% of the observations reported at least one asthma exacerbation over the three years following measurement of lung function. For those with an FEV1% predicted >80%, only 13% had an exacerbation. In a second cohort, a similar relationship was observed: among subjects with an FEV1% predicted <60%, 65% had at least one exacerbation, compared to those with an FEV1% predicted of >80%, where 34% experienced an exacerbation. Using logistic regression analysis to control for potential confounding variables, the association of FEV1% predicted and subsequent asthma attack persisted in both populations.

For purposes of modeling acute event incidence, we focused on the relationship between FEV1% predicted and observed rates of emergency department (ED) use. We used a retrospective study that reported ED rate and mean FEV1% predicted for three severity groups: mild, moderate, and severe [61]. We assumed a logistic relationship to estimate the following logit risk function:

$$\text{ED rate} = 1/(1+\exp(-2.1872 + 0.0560 * \text{FEV1\% predicted}))$$

The overall rate estimated by this function was adjusted upward or downward depending on the history of prior hospitalizations. To then determine the number of acute exacerbations that resulted only in an urgent care visit, we made use of a local patient database that has previously been employed for other peer-reviewed utilization studies [62, 63]. Specifically, we estimated the ratio of the number of urgent care visits that do not result in an ED visit (or hospitalization) to the number of asthma-specific ED visits (assuming that all ED visits are the result of an exacerbation). These ratios were stratified by asthma severity and age group and ranged from 3.4 to 6.1. (Milder disease and older age were associated with higher ratios.) We used the same database to estimate the proportion of all asthma-related ED visits that result in an admission to the hospital, stratified by asthma severity and age group. Patients with more severe disease and older age were more likely to be admitted to the hospital from the ED (proportions ranged from 0.17 to 0.32) (see Table 26.1).

Table 26.1 Monthly acute event probabilities for adults with mild-to-moderate asthma*

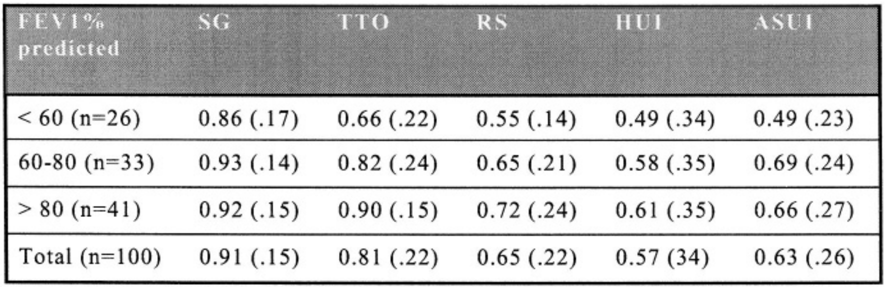


		NO PRIOR HOSPITALIZATIONS		ONE PRIOR HOSPITALIZATION		MORE THAN ONE PRIOR HOSPITALIZATION	
		FEV1% predicted		FEV1% predicted		FEV1% predicted	
		>80	60-80	>80	60-80	>80	60-80
Non-ED urgent care visits							
Age	18-35	0.0113	0.0208	0.0149	0.0275	0.0249	0.0459
	35+	0.0141	0.0277	0.0186	0.0365	0.0310	0.0608
ED visit without hospitalization							
Age	18-35	0.0019	0.0046	0.0026	0.0061	0.0043	0.0102
	35+	0.0016	0.0042	0.0022	0.0056	0.0036	0.0093
Hospitalizations							
Age	18-35	0.0004	0.0015	0.0005	0.0020	0.0009	0.0034
	35+	0.0007	0.0020	0.0009	0.0026	0.0015	0.0044

*Sources: [61, 62].

Relationship of FEV1% predicted and patient preferences We sought to examine how preference-based quality of life varied with FEV1% predicted in adults with asthma using four preference elicitation techniques [49]. In the case of the relationship between lung function and values (utilities), published evidence is limited [64, 65]. No studies have considered a sufficiently rich set of symptomatic health states to be suitable for our use. Moreover, we know of no studies that have directly collected community-based preferences in asthma. We therefore chose to collect our own preference weights in a companion, cross-sectional study of 100 adult (≥18 years) asthmatics in the Lexington, Kentucky area [66]. All patients in the study met the following inclusion criteria: 1) diagnosis of asthma as documented in the pharmacy computer system; 2) drug therapy indicative of asthma; and 3) self-report of asthma. Patients were administered several preference elicitation techniques, including time tradeoff (TTO) and standard gamble (SG) questions, the Health Utilities Index (HUI), and the Asthma Symptom Utility Index (ASUI). For each patient, information was also collected on FEV1% predicted. The relationship between FEV1% predicted and preference scores was obtained for each of the preference assessment techniques using univariate and multivariate regression techniques (see Table 26.2).

Table 26.2 FEV1% predicted and patient preferences*



FEV1% predicted	SG	TTO	RS	HUI	ASUI
< 60 (n=26)	0.86 (.17)	0.66 (.22)	0.55 (.14)	0.49 (.34)	0.49 (.23)
60-80 (n=33)	0.93 (.14)	0.82 (.24)	0.65 (.21)	0.58 (.35)	0.69 (.24)
> 80 (n=41)	0.92 (.15)	0.90 (.15)	0.72 (.24)	0.61 (.35)	0.66 (.27)
Total (n=100)	0.91 (.15)	0.81 (.22)	0.65 (.22)	0.57 (34)	0.63 (.26)

* Sources: [65, 66]; Preference scores from five different elicitation methods: SG = standard gamble; TTO = time tradeoff; RS = rating scale; HUI = Health Utilities Index; ASUI = Asthma Symptom Utility Index. Each cell reports mean (and standard deviation) values.

We found that FEV1% predicted was positively associated with preference scores for each of the instruments, though the relationship was statistically significant only for the TTO. Based upon these results, we concluded that lung function, as measured by FEV1% predicted, is an important predictor of patient preference in adults with asthma. The current version of the Asthma Policy Model therefore employs the following estimated function to adjust for health-related quality-of-life (HRQOL) effects:

$$\text{TTO} = 0.521 + 0.003958 * \text{FEV1\% predicted}$$

The U.S. Panel on Cost-effectiveness in Health and Medicine recommends that community preferences for health states be used, whenever feasible, to value morbidity consequences [46]. However, the Panel acknowledges the difficulties of adhering to this recommendation; our approach reflects one of the practical approximations they enumerate.

26.3.4 Note on modeling strategy: The impact of therapy

Having established FEV1% predicted as our marker of disease severity, we turned to the question of modeling the impact of therapy on that marker. To quantify this relationship, we performed a Medline search on the words: *randomized clinical trial; asthma; FEV1* (or any one of *forced expiratory volume, spirometry, pulmonary function, lung function,* or *respiratory function*); and *beclomethasone* (or any one of *flunisolide, triamcinalone, budesonide,* or *fluticasone.*) Of 352 English-language articles retrieved, 76 were deemed appropriate for full review based on the following criteria: 1) randomized controlled trial (excluding studies in which either the intervention arm or the "placebo" group used continuous oral steroids or other controller therapies, such as leukotriene modifiers, cromolyn sodium, nedocromil, or long-acting β-agonists); 2) intervention was an ICS (any of the five currently available in the U.S. market); 3) subjects met the individual study's criteria for asthma diagnosis; and 4) FEV1% predicted (or the data necessary to calculate it) reported as an outcome. Sixteen studies, yielding 26 active treatment arms, met these criteria [67-82].

For each treatment arm, the effect of intervention was expressed in terms of the percent change in FEV1% predicted. We subtracted the change in FEV1% predicted for the placebo group from the change in the FEV1% predicted for the intervention group, giving us the expected effect of a particular treatment relative to placebo. Because our purpose was to produce a summary efficacy estimate for the model, we calculated a mean change in FEV1% predicted weighted by the number of study subjects (both for all studies combined and stratified by baseline lung function).

We found that improvements in FEV1% predicted ranged from 1% to 22% of the baseline value. Because the dose range in these studies was relatively narrow (none of the trials compared low-dose therapy to high-dose therapy), and the absolute improvement in FEV1% predicted associated with the highest compared to the lowest dose was negligible or small (0 to 3%), we assumed equivalency across dose ranges. We estimated that ICS therapy produces baseline relative increases in FEV1% predicted of 7.6% for patients with mild disease and 11.6% for patients with moderate asthma. Mindful of the limitations of our approach, however, we explored efficacy values ranging from 1% to 22% in sensitivity analysis.

We assumed that the impact of therapy on an acute event was mediated entirely through lung function. No benefits independent of FEV1% predicted effect were assumed. Moreover, we assumed equivalency across all ICS preparations. While the model can be employed to simulate the effects of a given agent, our approach is consistent with the most recent version of the NAEPP guidelines (which acknowledge differences on a per inhalation or microgram basis, but which do not currently define any implications of these differences for purposes of clinical dosing recommendations) [1].

26.3.5 Other modeling considerations

Mortality Because no published estimates of asthma mortality could be located that were stratified by lung function, age-specific death rates were applied across severity strata. Estimated monthly probabilities of asthma-related death were 10^{-5} for patients aged 18-35 and 2.0×10^{-5} for patients over

35 years [83]. Non-asthma-related mortality rates were taken from the U.S. life tables [84].

Costs Costs were estimated from published resource utilization studies [85-91]. Baseline monthly chronic care costs (medications, routine office visits, laboratory testing) were \$35 and \$57 for mild and moderate patients, respectively. Acute event costs included \$63 for non-ED urgent care visits, \$242 for ED visits, and \$3,200 for hospitalizations. A \$38/month ICS drug cost was estimated from the manufacturer's reported number of inhalations per container, the average wholesale cost per container, and the estimated number of inhalations required to achieve the desired daily dose over 30 days [92]. Recognizing the demographic and clinical variation across studies, we conducted analyses that explored the effects of varying costs by as much as 50% to 200% of their baseline values.

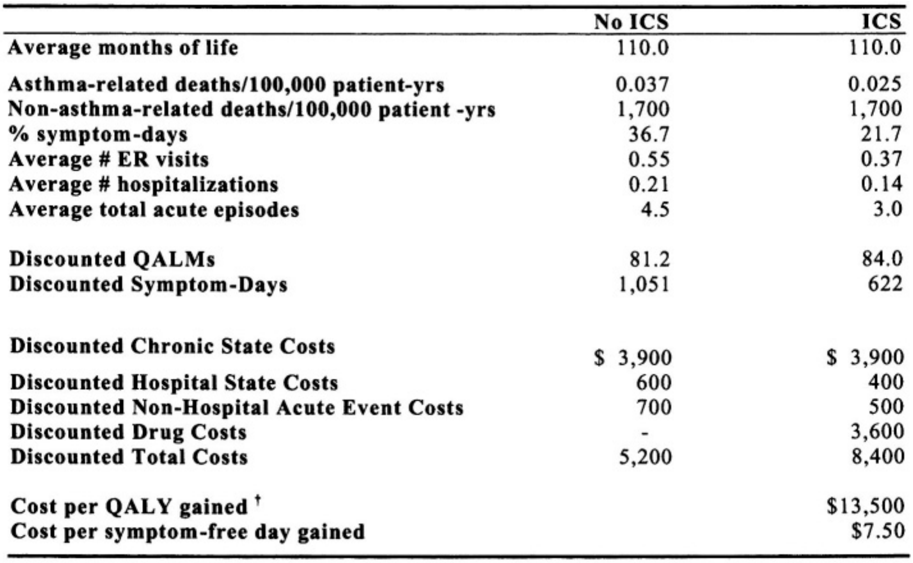
Disutility of ICS therapy Documented side-effects of ICS therapy in adults include dysphonia and thrush, cataracts and intraocular pressure, adrenal suppression, and the development of osteoporosis [93, 94]. There is concern that these weigh heavily in patients' and clinicians' decision making. In the absence of any data on the subject, we conducted a "what-if analysis, with the impact of side-effects modeled as an across-the-board, percent reduction in the HRQOL adjustment value for any patient receiving ICS therapy, regardless of dose. The baseline value for this reduction was 0%; values up to 3% were considered in sensitivity analysis.

26.4 RESULTS

26.4.1 Natural history

The model estimates that, over a 10-year horizon, patients receiving quick relievers alone will live an average of 81.2 quality-adjusted life-months (QALMs). Undiscounted and unadjusted for HRQOL effects, this equates to 110 months of life, with virtually all deaths attributable to non-asthma–related causes (see Table 26.3). The model predicts a population average of 36.7% symptom-days, 4.5 acute episodes per person over the 10-year period, and \$5,200 in per person, discounted, total direct costs. As expected, the model predicts more serious outcomes for patients with more severe disease. For example, a patient with mild asthma will experience 3.4 acute exacerbations and will incur costs of \$4,200 over a 10-year period; patients with moderate illness are predicted to experience, on average, 7.4 acute exacerbations and incur \$7,800 in costs over the same period.

Table 26.3 Baseline clinical and cost-effectiveness results for adults with mild-to-moderate asthma (10-year planning horizon)



	No ICS	ICS
Average months of life	110.0	110.0
Asthma-related deaths/100,000 patient-yrs	0.037	0.025
Non-asthma-related deaths/100,000 patient -yrs	1,700	1,700
% symptom-days	36.7	21.7
Average # ER visits	0.55	0.37
Average # hospitalizations	0.21	0.14
Average total acute episodes	4.5	3.0
Discounted QALMs	81.2	84.0
Discounted Symptom-Days	1,051	622
Discounted Chronic State Costs	$ 3,900	$ 3,900
Discounted Hospital State Costs	600	400
Discounted Non-Hospital Acute Event Costs	700	500
Discounted Drug Costs	-	3,600
Discounted Total Costs	5,200	8,400
Cost per QALY gained †		$13,500
Cost per symptom-free day gained		$7.50

† Reported cost-effectiveness ratios are not precisely equal to the ratio of costs and effects due to rounding.

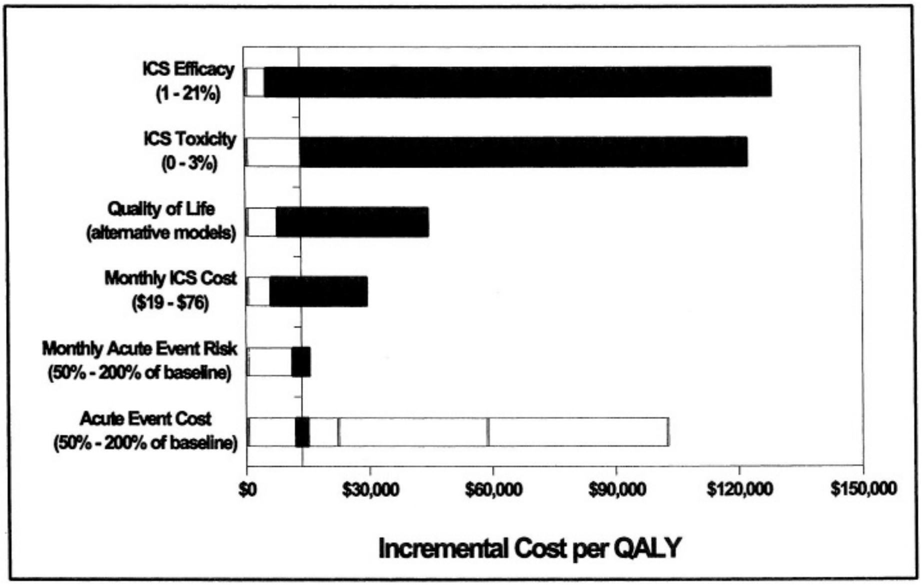
26.4.2 Cost-effectiveness of ICS

Addition of ICS increases discounted, quality-adjusted life over the 10-year planning period to 84.0 QALMs and costs to $8,400. (Additional costs are almost all attributable to drug outlays.) Symptom-days are reduced to 21.7% and the average patient experiences only 3.0 acute events. Compared with quick relievers alone, we estimate an incremental cost of $13,500 per QALY gained and $7.50 per additional symptom-free day (see Table 26.3).

26.4.3 Sensitivity analysis

To explore uncertainties in our results, we conducted a number of one-way sensitivity analyses (wherein a single input parameter was varied over the range of plausible values). A selection of output from these efforts is presented in Figure 26.2. In most instances, the policy conclusions were robust over reasonable parameter uncertainty; there were some exceptions, however. Most notable among these was the effect of ICS efficacy. As the percent change in FEV1% predicted was varied from 21% to 1%, the cost-effectiveness ratio ranged from $5,000 to $128,000 per QALY gained.

Figure 26.2 Sensitivity analysis*



* This "tornado diagram" summarizes the results of a series of one-way sensitivity analyses. Each horizontal bar represents a given model parameter. The vertical axis sits at the base case incremental cost-effectiveness estimate ($13,500/QALY). The span of a given horizontal bar denotes the range of cost-effectiveness outcomes produced by varying that specific parameter over its plausible range.

26.4.4 Targeted intervention

We considered four possible approaches to targeted intervention: 1) quick relievers alone ("No ICS"), 2) ICS targeted to mild asthmatics ("Mild Only"), 3) ICS targeted to moderate asthmatics ("Moderate Only"), and 4) the original intervention where all patients receive ICS ("Mild and Moderate"). In keeping with accepted methods [46], we arrayed results in order of increasing cost and eliminated from consideration any "dominated" strategies (i.e., strategies that were more expensive and delivered fewer benefits than some other approach or combination of approaches). As described in Table 26.4, the "Mild Only" approach was dominated. We computed cost-effectiveness ratios for the remaining strategies. Compared with use of quick relievers alone, the "Moderate Only" strategy conferred additional QALYs at an incremental cost of $10,300. Compared with the "Moderate Only" strategy, scaling up to the "Mild and Moderate" approach had a cost-effectiveness ratio of $15,000/QALY gained.

Table 26.4 Cost-effectiveness of targeted interventions

Strategy	Cost	QALYs	Incremental $/QALY
Quick relievers alone	$ 5,185	6.77	--
+ ICS in moderate disease	$ 5,961	6.84	$ 10,300
+ ICS in mild disease	$ 7,616	6.93	Dominated
+ ICS in mild/mod disease	$ 8,392	7.00	$ 15,000

26.4.5 Conclusions from the present study

We conclude from the present study that even ICS intervention in mild asthmatics compares favorably with a host of lifesaving and health promotion alternatives for chronic disease. (Examples include: Ticlopidine vs. aspirin in 65-year-olds with high risk of stroke, $48,000/QALY; chemotherapy vs. no chemotherapy in 75-year-olds with breast cancer, $58,000/QALY; and bypass surgery vs. medical management in patients with moderate angina and triple vessel disease, $30,000/QALY [95].) These findings should be interpreted with some caution, however, since they hinge upon a number of uncertainties, most notably the efficacy of ICS therapy and its impact on patient perceptions of HRQOL.

26.5 AVENUES FOR FURTHER RESEARCH AND CONCLUSIONS

The analysis we have described is only a first step. We intend to continue to refine the model to explore questions which — given both the inevitability of difficult choices and severe time, data, and financial constraints — only a simulation-based approach can currently address. In this section, we describe our current activities to expand the existing Asthma Policy Model in four areas: enlarging the set of variables describing health states to capture children and high-risk populations (26.5.1); refining the measure of efficacy to consider new pharmacologic and non-pharmacologic interventions (26.5.2); improving capability to simulate drug side-effects (26.5.3); and incorporating a measure of patient adherence to therapy (26.5.4).

26.5.1 Model improvement #1: Children and high-risk populations

We are currently working to enlarge the set of variables describing health states along the following dimensions:

Patient age (6-10; 11-18; 18-35; >35): Children and adults experience different rates of disease progression, risk of exacerbations, and response to therapy. This produces differences in patterns of care, resource use, and perceived quality of life, The impact of age on these outcomes is most pronounced when comparing children to adults. We plan to distinguish between four broad age categories: pre-adolescent (age 6-10), adolescent (11-18), young adult (18-35), and older adult (35+).

Smoking *(yes; no):* Current smoking is associated with increased risks of health care utilization [96, 97] and accelerated rates of decline in FEV1% predicted among adult asthmatics [98, 99]. The effect of passive smoking on the respiratory health of children is undisputed, with odds ratios of up to 1.6 for respiratory illness, symptoms and middle ear disease [100]. Smoking status will be categorized into current smoker vs. nonsmoker among adult subjects. We will also distinguish between passive and active tobacco exposure.

Race/Ethnicity *(Black/Hispanic; other):* The impact of asthma is felt disproportionately in historically underserved communities. Among U.S. blacks and Hispanics, asthma incidence, hospitalization, mortality, and resource consumption rates all exceed observed averages. Reflecting these observations, we are working to refine the state dimensions to create an "at-risk" population segment (Black/Hispanic). This certainly will not capture the full meaning of "disadvantaged population." Nevertheless, it will provide us with a credible, widely reported, aggregate marker for socio-economic and cultural characteristics believed to drive resource utilization and health outcome patterns in asthma.

An expanded Asthma Policy Model will permit us to focus more directly on some of the nation's highest-risk and most vulnerable populations. These modifications will cause the model to grow from the current 32 states to roughly 250 states. While this represents an order-of-magnitude increase in size, the revised model will nevertheless remain comparatively small.

26.5.2 Model improvement #2: New drug and non-drug interventions

We are working to enlarge the spectrum of interventions that the model can simulate. This includes both drug therapies (including leukotriene modifiers and long-acting beta-agonists) and non-drug interventions (including smoking cessation, allergen avoidance, disease management, patient/parent education, and the use of case managers), to be used both as stand-alone programs and in combination with each other.

Enlarging the Asthma Policy Model's capacity to simulate the effects of such interventions represents a significant improvement over our existing framework. The current version of the model conservatively assumes that the impact of intervention is mediated exclusively through lung function. Any effects on symptoms, exacerbations, quality of life, or cost are the indirect result of a change in lung function. The updated model will capture direct effects on disease processes, thus refining the notion of "efficacy" and permitting us to evaluate not only existing interventions, but also to perform "what-if' assessments, a capability that will be of particular use in priority setting for new clinical trials and estimating the value of better clinical information.

26.5.3 Model improvement #3: Therapeutic side-effects

Documented side-effects of ICS therapy include dysphonia and thrush, cataracts and intraocular pressure, adrenal suppression, the development of osteoporosis, and negative effect on growth rates in children [101-103]. There is concern that these side-effects weigh heavily in patients' and clinicians' decision making. Our goal is to ask "threshold" questions such as: What magnitude of various side-effects — physiologic, psychological, and economic — would warrant a change in current practice guidelines? What would have to be true about patient preferences regarding risks and benefits in order to justify discontinuation of effective therapy due to side-effects? What could clinicians do with better preference data on side-effects and how might that information influence their patient care choices?

Refining the model's handling of side-effects will permit us to explore not only the short-run, clinical consequences of adverse therapeutic events but also their downstream impact on outcomes, cost, and perceptions of well-being. Given the weight that many patients and clinicians currently ascribe to treatment side-effects, this represents a significant improvement in the realism of the model.

26.5.4 Model improvement #4: Imperfect adherence

Imperfect adherence to therapy remains a major obstacle to the achievement of better outcomes in asthma. A review of the medical literature reveals that no single intervention strategy will assure compliance with prescribed medication [104, 105]. Moreover, limited evidence suggests that nearly complete benefit may be derived from ICS therapies even when patients are only moderately adherent [106]. The current version of the Asthma Policy Model does not explicitly account for any of these concerns. The model has some implicit accounting for adherence as the input data for the model are drawn from clinical trials conducted with less-than-perfectly adherent subjects. However, the adherence observed in a trial differs from that which can be expected in everyday settings. By failing to adjust for this, the current model may seriously overstate the efficacy of drug therapy.

Admittedly, any approach for modeling the effect of adherence is simple. However, the existing science base cannot currently support a more nuanced, multifactorial assessment. What can (and should) be undertaken at present is a model-based, “what-if” exploration designed to establish credible performance benchmarks for further investigation.

Acknowledgments

This research is supported by the National Heart, Lung, and Blood Institute (NHLBI) via grant R01- HL68201-01A1. Dr. Fuhlbrigge is supported by an NHLBI Mentored Clinical Scientist Development Award (K08-HL03910-01).

References

[1] National Asthma Education and Prevention Program (1997). *Guidelines for the Diagnosis and Management of Asthma. Expert Panel Report 2,* Publication # 97-4051, U.S. Department of Health and Human Services, Washington, DC.

[2] Mack, H., P. Johnson, H. Abbey, and R.C. Talcerno (1982). Prevalence of asthma and health service utilization of asthmatic children in a inner city. *Journal ofAllergy and Clinical Immunology,* 70, 367-372.

[3] Halfon, N. and P.W. Newacheck (1986). Trends in the hospitalization of acute childhood asthma, 1970-1984. *American Journal of Public Health,* 76, 1308-1311.

[4] Evans, R. 3rd, D.E. Mullally, R.W. Wilson, P.J. Gergen, H.M. Rosenberg, J.S. Grauman, F.M. Chevarley, and M. Feinleib (1987). National trends in the morbidity and mortality of asthma in the US. Prevalence, hospitalization, and death from asthma over two decades: 1965-1984. *Chest,* 91, 65S-74S.

[5] Sly, R.M. (1988). Mortality from asthma, 1979-1984. *Journal of Allergy and Clinical Immunology,* 82, 705-717.

[6] Weiss, K.B., P.J. Gergen, and E.F. Crain (1992). Inner-city asthma. The epidemiology of an emerging US public health concern. *Chest,* 101, 362S-367S.

[7] National Heart, Lung, and Blood Institute (2000). *Morbidity and Mortality: 2000 Chart Book on Cardiovascular, Lung, and Blood Diseases.* National Institutes of Health, Bethesda, MD.

[8] Marder, D., P. Targonski, P. Orris, V. Persky, and W. Addington (1992). Effect of racial and socioeconomic factors on asthma mortality in Chicago. *Chest,* 101, 4265S-4429S.

[9] Carr, W., L. Zeitel, and K.B. Weiss (1992). Asthma hospitalization and mortality in New York City. *American Journal ofPublic Health,* 82, 59-65.

[10] Martinez, F.D., A.L. Wright, L.M. Taussig, C.J. Holberg, M. Halonen, and W.J. Morgan (1995). Asthma and wheezing in the first six years of life. The Group Health Medical Associates. *New England Journal of Medicine,* 332, 133-138.

[11] Kattan, M., H. Mitchell, P. Eggleston, P. Gergen, E. Crain, S. Redline, K. Weiss, R. Evans **3rd,** R. Kaslow, C. Kercsmar, F. Leickly, F. Malveaux, and H.J. Wedner (1997). Characteristics of inner-city children with asthma: The National Cooperative Inner-City Asthma Study. *Pediatric Pulmonology,* 24, 253-262.

[12] Kitch, B.T., G. Chew, H.A. Burge, M.L. Muilenberg, S.T. Weiss, T.A. Platts-Mills, G. O'Connor and D.R. Gold (2000). Socioeconomic predictors of high allergen levels in homes in the greater Boston area. *Environmental Health Perspectives,* 108, 301-307.

[13] Evans, R. (1992). Asthma among minority children. A growing problem. *Chest,* 101(Suppl), 368S-371S.

[14] Gottlieb, D.J., A.S. Beiser, and G.T. O'Connor (1995). Poverty, race, and medication use are correlates of asthma hospitalization rates. A small area analysis in Boston. *Chest,* 108, 28-35.

[15] Weiss, K.B., E.N. Grant, and T. Li (1999). The effects of asthma experience and social demographic characteristics on responses to the Chicago Community Asthma Survey-32. Chicago Asthma Surveillance Initiative Project Team. *Chest,* 116(Suppl 1), 183S-189S.

[16] Stableforth, D.E. (1987). Asthma mortality and physician competence. *Journal of Allergy and Clinical Immunology,* 80, 463-466.

[17] Kravis, L.P. (1987). An analysis of fifteen childhood asthma fatalities. *Journal of Allergy and Clinical Immunology,* 80, 467-472.

[18] Strunk, R.C., D.A. Mrazek, G.S.W. Fuhrmann, and J.F. LaBrecque (1985). Deaths from asthma in childhood. Can they be predicted? *Journal of the American Medical Association,* 254, 1193.

[19] Barnes, P.J. (1989). A new approach to the treatment of asthma. Department of Thoracic Medicine, Brompton Hospital, London, United Kingdom. *New England Journal of Medicine,* 321, 1517-1527.

[20] Barnes, P.J. and I. Adcock (1993). Anti-inflammatory actions of steroids: molecular mechanisms. *Trends in Pharmacological Sciences,* 14, 436-441.

[21] Pearlman, D.S., P. Chervinsky, C. LaForce, J.M. Seltzer, D.L. Southern, J.P. Kemp, RJ. Dockhorn, J. Grossman, R.F. Liddle, and S.W. Yancey (1992). A comparison of salmeterol with albuterol in the treatment of mild-to-moderate asthma. *New England Journal of Medicine,* 327, 1420-1425.

[22] Henderson, W.R., Jr. (1994). Role of leukotrienes in asthma. *Annals of Allergy,* 72, 272-278.

[23] Buchner, D.A., L.T. Butt, A. De Stefano, B. Edgren, A. Suarez, and R.M. Evans (1998). Effects of an asthma management program of the asthmatic member: Patient centered results of a 2 year study in a managed care organization. *American Journal ofManaged Care,* 4, 1288-1297.

[24] Hammarquist, C., M.L. Burr, and P.C. Gotzsche for the Cochrane Airways Group (2001). House dust mite control measures for asthma. *Cochrane Database of Systematic Reviews,* Issue 1.

[25] Peat, J.K. and J. Li (1999). Reversing the trend: Reducing the prevalence of asthma. *Journal ofAllergy and Clinical Immunology,* 130, 1-10.

[26] Custovic, A., S.C.O. Taggart, H.C. Francis, M.D. Chapman, and A. Woodcock (1995). Exposure to house dust mite allergens and the clinical activity of asthma. *Journal of Allergy and Clinical Immunology,* 98, 64-72.

[27] Durham, S.R. (1996). Allergen avoidance measures. British Thoracic Society. *Respiratory Medicine,* 90, 441-5.

[28] Watson, W.T.A., A.B. Becker, and F.E.R. Simons (1993). Treatment of allergic rhinitis with intranasal corticosteroids in patients with mild asthma: effect on lower airway responsiveness. *Journal of Allergy and Clinical Immunology,* 91, 97-101.

[29] Pedersen, B., R. Dahl, N. Lindqvist, and N. Mygind (1990). Nasal inhalation of the glucocorticoid budesonide from a spacer for the treatment of patients with pollen rhinitis and asthma. *Allergy,* 45, 451-456.

[30] Grant, J.A., C.F. Nicodemus, S.R. Findlay, MM. Glovsky, J. Grossman, H. Kaiser, E.O. Meltzer, D.Q. Mitchell, D. Pearlman, and J. Selner (1995). Clinical aspects of allergic disease. Cetrizine in patients with seasonal rhinitis and concomitant asthma: prospective, randomized, placebo-controlled trial. *Journal of Allergy and Clinical Immunology,* 95, 923-932.

[31] Corren, J., A. Harris, D. Aaronson, W. Beaucher, R. Berkowitz, E. Bronsky, R. Chen, P. Chervinsky, R. Cohen, J. Fourre, J. Grossman, E. Meltzer, A. Pedinoff, W. Strieker, and A. Wanderer (1997). Efficacy and safety of loratadine plus pseudoephedrine in patients with seasonal allergic rhinitis and mild asthma. *Journal of Allergy and Clinical Immunology,* 100, 781-788.

[32] Parmley, W.W. (1994). Clinical practice guidelines: does the cookbook have enough recipes? *Journal of the American Medical Association,*272, 1374-1375.

[33] Ellrodt, A.G., L. Conner, M. Riedinger, and S. Weingarten (1995). Measuring and improving physician compliance with clinical practice guidelines: a controlled interventional trial. *Annals of Internal Medicine,* 122,277-282.

[34] Grimshaw, K., N. Freemantle, S. Wallace, I. Russell, B. Herwitz, I. Watt, A. Long, and T. Sheldon (1995). Developing and implementing clinical practice guidelines. *Quality in Health Care,* 4, 55-64.

[35] Adams, R.J., A. Fuhlbrigge, T. Guilbert, P. Lozano, and F. Martinez (2002). Inadequate use of asthma medication in the United States: results of the asthma in America national population survey. *Journal of Allergy and Clinical Immunology,* 110, 58-64.

[36] Windsor, R.A., W.C. Bailey, J.M. Richards Jr., B. Manzella, S.J. Soong, and M. Brooks (1990). Evaluation of the efficacy and cost-effectiveness of health education methods to increase medication adherence among adults with asthma. *American Journal of Public Health,* 80, 1519.

[37] Gallefoss, F. and P.S. Bakke (2001). Cost-effectiveness of self-management in asthmatics: a 1-yr follow-up randomized, controlled trial. *European Respiratory Journal,* 17, 206-213.

[38] Wilson Pessano, S.R., P. Scamagas, G.M. Arsham, L. Chardon, S. Coss, D.F. German, and G.W. Hughes (1987). An evaluation of approaches to asthma self-management education for adults: the AIR/Kaiser Permanente study. *Journal of Health Education,* 14, 333-343.

[39] Clark, N.M., C.H. Feldman, D. Evans, M.J. Levison, Y. Wasilewski, and R.B. Mellins (1986). The impact of health education on frequency and cost of health care use by low income children with asthma. *Journal of Allergy and Clinical Immunology,* 78, 108-115.

[40] Lewis, C.E., G. Rachelefsky, M.A. Lewis, A. de la Sota, and M. Kaplan (1984). A randomized trial of A.C.T. (asthma care training) for kids. *Pediatrics,* 74, 478-486.

[41] Tierce, J.C., W. Meller, B. Berlow, and W.C. Gerth (1989). Assessing the cost of albuterol inhalers in the Michigan and California Medicaid programs: a total cost-of-care approach. *Clinical Therapeutics,* 11, 53.

[42] Campbell, L.M., R.J. Simpson, M.L. Turbitt, and P.D.I. Richardson (1993). A comparison of the cost-effectiveness of budesonide 400 mg/day and 800 mg/day in the management of mild-to-moderate asthma in general practice. *British Journal ofMedical Economics,* 6, 67-74.

[43] Ross, R.N., M. Morris, S.R. Sakowitz, and B.A. Berman (1988). Cost-effectiveness of including cromolyn sodium in the treatment program for asthma: a retrospective, record based study. *Clinical Therapeutics,* 10, 188-203.

[44] Rutten-van-Molken, M.P.M.H., E.K.A. Van Doorslaer, M.C.C. Jansen, H.A.M. Kerstjens, and F.F.H. Rutten (1995). Costs and effects of inhaled corticosteroids and bronchodilators in asthma and chronic obstructive pulmonary disease. *American Journal of Respiratory and Critical Care Medicine,* 151, 975-982.

[45] Connett, G.J., W. Lenney, and S.M. McConchie (1993). The cost-effectiveness of budesonide in severe asthmatics aged one to three years. *British Journal of Medical Economics,* 6, 127-270.

[46] Gold, M.R., J.E. Siegel, L.B. Russel, and M.C. Weinstein, Eds. (1996). Report of the Panel on Cost-effectiveness in Health and Medicine. *Cost-Effectiveness in Health and Medicine.* Oxford University Press, New York.

[47] Kuntz, K.M., B.T. Kitch, A.L. Fuhlbrigge, A.D. Paltiel, P.J. Neumann, and S.T. Weiss (2002). A novel approach to defining the relationship between lung function and symptom status in asthma. *Journal of Clinical Epidemiology,* 55, 11-18.

[48] Fuhlbrigge, A.L., B.T. Kitch, A.D. Paltiel, K.M. Kuntz, P.J. Neumann, D.W. Dockery, and S.T. Weiss (2001). FEV1 is associated with risk of asthma attacks in a pediatric population. *Journal of Allergy and Clinical Immunology,* 107, 61-67.

[49] Neumann, P.J., K. Blumenschein, A. Zillich, M. Johannesson, K.M. Kuntz, R.H. Chapman, S.T. Weiss, B.T. Kitch, A.L. Fuhlbrigge, and A.D. Paltiel (2000). Relationship between FEV1% predicted and utilities in adult asthma [Abstract]. *Medical Decision Making,* 20, 488.

[50] Paltiel, A.D., A.L. Fuhlbrigge, B.T. Kitch, B. Liljas, S.T. Weiss, P.J. Neumann, and K.M. Kuntz (2001). Cost-effectiveness of inhaled corticosteroids in adults with mild-to-moderate asthma: Results from

the Asthma Policy Model. *Journal of Allergy and Clinical Immunology,* 108, 39-46.

[51] Sullivan, S.D., A. Elixhauser, A.S. Buist, B.R. Luce, J. Eisenberg, and K.B. Weiss (1996). National Asthma Education and Prevention Program working group report on the cost-effectiveness of asthma care. *American Journal ofRespiratory and Critical Care Medicine,* 154, S84-S95.

[52] United States Bureau of the Census (1998). *StatisticalAbstract ofthe UnitedStates: 1998.* Washington, DC.

[53] Beck, J.R. and S.G. Pauker (1983). The Markov process in medical prognosis. *MedicalDecision Making,* 3, 419-458.

[54] Sonnenberg, F.A. and J.R. Beck (1993). Markov models in medical decision making: a practical guide. *Medical Decision Making,* 13, 322-338.

[55] Crane, J., N. Pearce, C. Burgess, K. Woodman, B. Robson, and R. Beasley (1992). Markers of risk of asthma death or readmission in the 12 months following a hospital admission for asthma. *International Journal of Epidemiology,* 21, 737-744.

[56] Ortega, A.N., K.D. Belanger, M.B. Bracken, and B.P. Leaderer (2001). A childhood asthma severity scale: symptoms, medications, and health care visits. *Annals ofAllergy, Asthma, and Immunology,*86, 405-413.

[57] National Asthma Education and Prevention Program (1991). *Guidelines for the Diagnosis and Management ofAsthma. Expert Panel Report.* U.S. Department of Health and Human Services, Washington, DC, Publication # 91-3042.

[58] Enright, P.L., M.D. Lebowitz, and D.W. Cockroft (1994). Physiologic measures: pulmonary function tests. Asthma outcomes. *American Journal ofRespiratory and Critical Care Medicine,* 149, S9-S20.

[59] Enright, P.L., L.R. Johnson, J.E. Connett, H. Voelker, and A.S. Buist (1991). Spirometry in the Lung Health Study. Methods and quality control. *American Review of Respiratory Disease,* 143, 1215-1221.

[60] Kitch, B.T., A.L. Fuhlbrigge, K.M. Kuntz, P.J. Neumann, A.D. Paltiel, B. Rijeken, B. Schouten, D.S. Postma, and S.T. Weiss (1999). Relationship between FEV1 and subsequent asthma attacks. *American Journal of Respiratory and Critical Care Medicine,* 159, A135.

[61] Thomas, P., R.N. Ross, and J.R. Farrar (1996). A retrospective assessment of cost avoidance associated with the use of nedocromil sodium metered-dose inhaler in the treatment of patients with asthma. *Clinical Therapeutics,* 18, 939-952.

[62] Donahue, J.G., S.T. Weiss, J.M. Livingston, M.A. Goetsch, D.K. Greineder, and R. Platt (1997). Inhaled steroids and the risk of hospitalization for asthma. *Journal of the American Medical Association,* 277, 887-891.

[63] Finkelstein, J.A., M.B. Barton, J.G. Donahue, P. Algatt-Bergstrom, L.E. Markson, and R. Platt (2000). Comparing asthma care for Medicaid and non-Medicaid children in an HMO. *Archives of Pediatrics and Adolescent Medicine,* 154, 563-568.

[64] Rutten-van-Molken, M.P.M.H., E.K.A. Van Doorslaer, M.C.C. Jansen, H.A.M. Kerstjens, and F.F.H. Rutten (1995). Costs and effects of inhaled corticosteroids and bronchodilators in asthma and chronic obstructive pulmonary disease. *American Journal of Respiratory and Critical Care Medicine,* 151, 975-982.

[65] Revicki, D.A., N.K. Leidy, F. Brennan-Diemer, S. Sorensen, and A. Togias (1998). Integrating patient preferences into health outcomes assessment: the multiattribute Asthma Symptom Utility Index. *Chest,* 114, 998-1007.

[66] Neumann, P.J., K. Blumenschein, A. Zillich, M. Johannesson, K.M. Kuntz, R.H. Chapman, S.T. Weiss, B.T. Kitch, A.L. Fuhlbrigge, and A.D. Paltiel (2000). Relationship between FEV1% predicted and utilities in adult asthma. *Medical Decision Making,* 20, 488.

[67] Pearlman, D.S., W. Stricker, S. Weinstein, G. Gross, P. Chervisnsky, A. Woodring, B. Prillaman, and T. Shah (1999). Inhaled salmeterol and fluticasone: a study comparing monotherapy and combination therapy in asthma. *Annals ofAllergy, Asthma, and Immunology,* 82, 257-265.

[68] Noonan, M., P. Chervinsky, W.W. Busse, S.C. Weisberg, J. Pinna, B.P. de Boisblanc, H. Boltansky, D. Pearlman, L. Rephser, D. Kellerman (1995). Fluticasone propionate reduces oral prednisone use while it improves asthma control and quality of life. *American Journal of Respiratory and Critical Care Medicine,* 152, 1467-1473.

[69] Bel, E.H., M.C. Timmers, J. Hermans, J.H. Dijkman, and P.J. Sterk (1990). The long-term effects of nedocromil sodium and beclomethasone dipropionate on bronchial responsiveness to methacholine in nonatopic asthmatic subjects. *American Review of Respiratory Disease,* 141, 21-28.

[70] Fahy, J.V. and H.A. Boushey (1998). Effect of low-dose beclomethasone dipropionate on asthma control and airway inflammation. *European Respiratory Journal,* 11(6), 1240-1247.

[71] Galant, S.P., M. Lawrence, E.O. Meltzer, M. Tomasko, K.A. Baker, and D.J. Kellerman (1996). Fluticasone propionate compared with theophylline for mild-to-moderate asthma. *Annals of Allergy, Asthma, and Immunology,* 77, 112-118.

[72] Kemp, J., A.A. Wanderer, J. Ramsdell, D. L. Southern, S. Weiss, D. Aaronson, and J. Grossman (1999). Rapid onset of control with budesonide Turbuhaler in patients with mild-to-moderate asthma. *Annals of Allergy, Asthma, and Immunology,* 82, 463-471.

[73] Li, J.T., L.B. Ford, P. Chervinsky, S.C. Weisberg, D.J. Kellerman, K.G. Faulkner, N.E. Herje, A. Hamedani, S.M. Harding, and T. Shah (1999). Fluticasone propionate powder and lack of clinically significant effects on hypothalamic-pituitary-adrenal axis and bone mineral density over 2 years in adults with mild asthma. *Journal of Allergy and Clinical Immunology,* 103, 1062-1068.

[74] McFadden, E.R., T.B. Casale, T.B. Edwards, J.P. Kemp, W.J. Metzger, H.S. Nelson, W.W. Storms, and M.J. Neidl (1999). Administration of budesonide once daily by means of turbuhaler to subjects with stable asthma. *Journal of Allergy and Clinical Immunology,* 104, 46-52.

[75] Meltzer, E.O., H.A. Orgel, E.F. Ellis, H.N. Eigen, and M.P. Hemstreet (1992). Long-term comparison of three combinations of albuterol, theophylline, and beclomethasone in children with chronic asthma. *Journal of Allergy and Clinical Immunology,* 90, 2-11.

[76] Nathan, R.A., J.L. Pinnas, H.J. Schwartz, J. Grossman, S.W. Yancey, A.H. Emmett, and K.A. Rickard (1999). A six-month, placebo-controlled comparison of the safety and efficacy of salmeterol or beclomethasone for persistent asthma. *Annals of Allergy, Asthma, and Immunology,* 82, 521-529.

[77] Osterman, K., M. Carlholm, J. Ekelund, J. Kiviloog, K. Nikander, L. Nilholm, P. Salomonsson, V. Strand, P. Venge, and O. Zetterstrom (1997). Effect of 1 year daily treatment with 400 microg budesonide (Pulmicort Turbuhaler) in newly diagnosed asthmatics. *European Respiratory Journal,* 10, 2210-2215.

[78] Peden, D.B., W.E. Berger, M.J. Noonan, M.J. Noonan, M.R. Thomas, V.L. Hendricks, A.G. Hamedani, P. Mahajan, and K.W. House (1998). Inhaled fluticasone propionate delivered by means of two different multidose powder inhalers is effective and safe in a large pediatric population with persistent asthma. *Journal of Allergy and Clinical Immunology,* 102, 32-38.

[79] Reed, C.E., K.P. Offord, H.S. Nelson, J.T. Li, and D.G. Tinkelman (1998). Aerosol beclomethasone dipropionate spray compared with theophylline as primary treatment for chronic mild-to-moderate asthma. The American Academy of Allergy, Asthma and Immunology Beclomethasone Dipropionate-Theophylline Study Group. *Journal of Allergy and Clinical Immunology,* 101, 14-23.

[80] Sheffer, A.L., C. LaForce, P. Chervinsky, D. Pearlman, and A. Schaberg (1996). Fluticasone propionate aerosol: efficacy in patients

with mild to moderate asthma. Fluticasone Propionate Asthma Study Group. *Family Practice,* 42, 369-375.

[81] Simons, F.E. (1997). A comparison of beclomethasone, salmeterol, and placebo in children with asthma. Canadian Beclomethasone Dipropionate-Salmeterol Xinafoate Study Group. *New England Journal of Medicine,* 337, 1659-1665.

[82] Tinkelman, D.G., C.E. Reed, H.S. Nelson, and K.P. Offord (1993). Aerosol beclomethasone dipropionate compared with theophylline as primary treatment of chronic, mild to moderately severe asthma in children. *Pediatrics,* 92, 64-77.

[83] Evans, R. 3rd, D.I. Mullally, R.W. Wilson, P. J. Gergen, H.M. Rosenberg, J.S. Grauman, F.M. Chevarley, and M. Feinleib (1987). National Trends in Morbidity and Mortality of Asthma in the US. *Chest,* 91, 65S-73S.

[84] U.S. Bureau of the Census (1997). *Statistical Abstract of the United States: 1997.* Washington, DC.

[85] Serra-Batlles, J., V. Plaza, E. Morejon, A. Comella, and J. Brugues (1998). Costs of asthma according to the degree of severity. *European Respiratory Journal,* 12, 1322-1326.

[86] Stanford, R., T. McLaughlin, and L.J. Okamoto (1999). The cost of asthma in the emergency department and hospital. *American Journal of Respiratory and Critical Care Medicine,* 160, 211-215.

[87] Lewis, C.E., G. Rachelefsky, M.A. Lewis, A. de la Sota, and M. Kaplan (1984). A randomized trial of A.C.T. (asthma care training) for kids. *Pediatrics,* 74, 478-486.

[88] Ross, R.N., M. Morris, S.R. Sakowitz, and B.A. Berman (1988). Cost-effectiveness of including cromolyn sodium in the treatment program for asthma: a retrospective, record-based study. *Clinical Therapeutics,* 10, 188-203.

[89] Weiss, K.B., P.J. Gergen, and T.A. Hodgson (1992). An economic evaluation of asthma in the United States. *New England Journal of Medicine,* 326, 862-866.

[90] Coventry, J.A., M.S. Weston, and P.M. Collins (1996). Emergency room encounters of pediatric patients with asthma: cost comparisons with other treatment settings. *Journal of Ambulatory Care Management,* 19, 9-21

[91] Segal, R., L.D. Ried, and J. Mackowiak (1995). Cost of asthma illness: Emergency department visits without admission. *Pharmacy Practice Management Quarterly,* 15, 72-82.

[92] Drug Topics Red Book (1998). *Medical Economics.* Montvale, NJ.

[93] Barnes, P.J., S. Pedersen, and W.W. Busse (1993). Efficacy and safety of inhaled corticosteroids. New developments. *American Journal of Respiratory and Critical Care Medicine,* 157, S1-S53.

[94] O'Byrne, P.M. and S. Pedersen (1998). Measuring efficacy and safety of different inhaled corticosteroid preparations. *Journal of Allergy and Clinical Immunology,* 102, 879-886.

[95] Chapman, R.H., P.W. Stone, E.A. Sandberg, C. Bell, and P.J. Neumann (2000). A comprehensive league table of cost-utility ratios and a sub-table of "panel-worthy" studies. *Medical Decision Making,* 20, 451-467.

[96] Sippel, J.M., K.L. Pudula, W.M. Vollmer, A.S. Buist, and M.L. Osborne (1999). Association of smoking with hospital-based care and quality of life in patients with obstructive airway disease. *Chest,*115, 691-696.

[97] Prescott, E., P. Lange, and J. Vestbo (1997). Effect of gender on hospital admission for asthma and prevalence of self-reported asthma: a prospective study based on a sample of the general population. Copenhagen City Health Study Group. *Thorax,* 52, 287-289.

[98] Lange, P., J. Parner, J. Bestbro, P. Schnohr, and G. Jensen (1998). A 15 year follow up study of ventilatory function in adults with asthma. *New England Journal of Medicine,* 339, 1194-1200.

[99] Ulrik, C.S. and P. Lange (1994). Decline of Lung Function in Adults with Bronchial Asthma. *American Journal of Respiratory and Critical Care Medicine,* 150, 629-634.

[100] Cook, D.G. and D.P. Strachan (1999). Health effects of passive smoking-10: Summary of effects of parental smoking on the respiratory health of children and implications for research. *Thorax,* 54, 357-366.

[101] Toogood, J.H., A.E. Markov, J. Baskerville, and C. Dyson (1993). Association of ocular cataracts with inhaled and oral steroid therapy during long-term treatment of asthma. *Journal of Allergy and Clinical Immunology,* 91, 571-579.

[102] Toogood, J.H. (1998). Inhaled steroid asthma treatment: ‘Primum non nocere’. *Canadian Respiratory Journal,* 5, 50A-53A.

[103] Niewoehner, C.B. and D.E. Niewoehner (1999). Steroid-induced osteoporosis. Are your asthmatic patients at risk? *Postgraduate Medicine,* 105, 79-83, 87-8, 91.

[104] Christensen-Szalanski, J.J. and G.B. Northcraft (1985). Patient compliance behavior: the effect of time on patients’ valued of treatment regiments. *Social Science and Medicine,* 21, 263-273.

[105] Greenburg, R.N. (1984). Overview of patient compliance with medication dosing: a literature review. *Clinical Therapeutics,* 6, 592-599.

[106] Donahue, J.G., S.T. Weiss, J.M. Livingston, M.A. Goetsch, D.K. Greineder, and R. Platt (1997). Inhaled steroids and the risk of hospitalization for asthma. *Journal of the American Medical Association,* 277, 887-891.

27 A BAYESIAN NETWORK TO ASSIST MAMMOGRAPHY INTERPRETATION

Daniel L. Rubin[1], Elizabeth S. Burnside[2] and Ross Shachter[3]

[1] Stanford Medical Informatics
Stanford University
Stanford, CA 94305

[2] Department of Radiology
University of Wisconsin
Madison, WI 53792

[3] Department of Management Science and Engineering
Stanford University
Stanford, CA 94305

SUMMARY

Mammography is a vital screening test for breast cancer because early diagnosis is the most effective means of decreasing the death rate from this disease. However, interpreting the mammographic images and rendering the correct diagnosis is challenging. The diagnostic accuracy of mammography varies with the expertise of the radiologist interpreting the images, resulting in significant variability in screening performance. Radiologists interpreting mammograms must manage uncertainties arising from a multitude of findings. We believe that much of the variability in mammography diagnostic performance arises from heuristic errors that radiologists make in managing these uncertainties. We developed a Bayesian network that models the probabilistic relationships between breast diseases, mammographic findings and patient risk factors. We have performed some preliminary evaluations in test cases from a mammography atlas and in a prospective series of patients who had biopsy confirmation of the diagnosis. The model appears useful for clarifying the decision about whether to biopsy abnormalities seen on mammography, and also can help the radiologist correlate histopathologic findings with the mammographic abnormalities observed. Our preliminary experience suggests that this model may help reduce variability and improve overall interpretive performance in mammography.

KEY WORDS

Mammography, Diagnosis, Breast cancer, Bayesian networks

27.1 INTRODUCTION

Breast cancer is the most frequently diagnosed malignancy among American women. It is the second leading cause of cancer death (after lung cancer) among women of all ages and the leading cause of cancer death among women aged 40 to 59 years [1]. Mammography has been shown to be effective in detecting breast cancer before it becomes clinically evident [2]; consequently, routine screening with mammography is now generally accepted as a valuable tool for decreasing mortality from breast cancer.

The benefits of screening mammography are limited by the quality of the image acquired and by the accuracy of image interpretation. Image acquisition quality depends on the quality and operation of the imaging equipment, while interpretation depends on the training and expertise of a human reader (the radiologist). In recent years, standards relating to the imaging equipment such as the Mammography Quality Standards Act (MQSA) [3] have improved the quality of mammogram images at many facilities [4]. However, overall accuracy of mammographic interpretation, in terms of sensitivity and specificity, is a problem because of variability in the training and experience of radiologists interpreting the images [5]. Several studies have reported substantial mammogram interpretation inconsistencies among different radiologists [6-8], which would lead to different followup testing and treatment decisions.

False-negative and false-positive interpretations have been called "risks" of screening mammography and have been cited as arguments against routine screening of various populations of women [9-11]. False negative interpretations are risks because patients having cancer are not detected (reduced efficacy of screening), thus delaying cancer treatment and leading to higher morbidity and mortality. On the other hand, false positive interpretations are risks because patients without cancer undergo unnecessary biopsy (causing anxiety and increased medical costs). Variability in interpretive accuracy among radiologists lowers the average positive predictive value for mammography, which makes it a less effective tool for the early detection of breast cancer. Therefore, strategies to reduce variability in mammographic interpretation are essential to improve patient care.

Some of the variability in interpretative accuracy among radiologists is likely related to training and experience. Some radiologists have subspeciality training in mammography and read these studies exclusively. These individuals are generally considered "experts" in the field. On the other hand, community radiologists read the majority of mammograms in the context of diverse general practice. Community radiologists have higher biopsy rates and thus lower positive predictive value of malignant disease [5, 12].

One approach to reduce the variation in interpretations among radiologists is to standardize the vocabulary used in mammography reports. The American College of Radiology (ACR) developed BI-RADS, a lexicon of mammogram findings (or "features") and the distinctions that describe them. [13]. The developers of BI-RADS tried to identify those features of mammograms that are most useful for discriminating diseases of the breast. To accomplish this, they performed statistical analysis of the terms ("descriptors") used to describe imaging findings to determine which descriptors best discriminate between a benign or malignant diagnosis [14].

While BI-RADS is an important step in reducing the variation in mammography reporting, it does not solve the problem of how radiologists relate a set of findings they observe on the mammogram to a diagnosis. Specifically, how does the radiologist determine the probability of malignancy given a set of observed findings so as to choose followup tests and treatments? The quality of this determination is the essential difference between an expert and a non-expert, and likely accounts for much of the interpretive variation among radiologists.

Because the BI-RADS findings observed on mammography were selected to discriminate between benign and malignant diseases, they contain precisely the information we want to obtain for diagnosis. Our hypothesis is that we can build a probabilistic model relating diseases to the BI-RADS descriptor findings seen on mammography, and that, given the BI-RADS findings, this model can be used to compute posterior probabilities for the possible breast diseases. Such probabilities can guide the radiologist's decision making.

Our goal has been to build a model that represents these probabilistic relationships among BI-RADS findings and includes other pertinent information (patient risk factors) to standardize how combinations of BI-RADS findings are interpreted. Such a model could also be used to determine the likelihood of breast diseases and to evaluate the agreement (concordance) between biopsy results and the mammographic findings. Our hope is to bring clarity to decision making and reduce suboptimal variability in patient management based on a normative approach.

27.2 METHODOLOGY

27.2.1 Building a model for mammography diagnosis

To represent the probabilistic relationships among findings and diseases, we built a Bayesian belief network. Bayesian networks are graphical probabilistic models of the conditional dependencies among variables of interest [15]. In our application, we are interested in breast diseases that are diagnosed on mammography, the radiological findings that are observed on mammography (in terms of BI-RADS descriptors), and patient risk factors (age, history of hormone treatment, and history of prior breast cancer). Because a mammogram may contain more than one abnormality ("lesion"), we built a lesion-centric model; if a patient has more than one lesion, the model can be applied to each lesion independently. For the time being, however, we will assume that each patient has at most one lesion.

Diseases From a review of the literature and with the assistance of an expert in mammography, we identified 23 diseases of the breast. These diseases, in addition to a "normal" diagnosis and two combined diagnoses, were selected as the distinctions for a DISEASE node (having 26 states) in the model (Table 27.1).

In order for us to define mutually exclusive disease distinctions, we assume that a given lesion on the mammogram represents a single specific disease process. Because of the pathophysiology of breast cancer, it is possible to see two diseases simultaneously within a single malignant lesion. Simultaneous appearance of two diseases occurs when atypical cells transform into malignant cells. For example, "ductal carcinoma *in situ*" (DCIS) contains non-invasive neoplastic cells that may undergo transformation into "ductal carcinoma, not otherwise specified" (DCNOS). Thus, some breast lesions may contain both diseases if only some of the cells have transformed. For this reason, the DISEASE node includes two combined diagnoses, "LC+LCIS" and "DCNOS+DCIS" (Table 27.1).

In order for all 26 states in DISEASE to be collectively exhaustive, we assume that we have modeled all possible diseases (or disease combinations) that may be diagnosed on mammography, including a benign state of no disease (Normal).

Findings and Patient Risk Factors We compiled a list of findings (abnormalities) observed on mammography from the BI-RADS descriptors [13]. BI-RADS consists of 43 descriptors, some of which are organized in a hierarchical taxonomy (Figure 27.1). The hierarchical structure of the descriptors helps the user navigate and select descriptors and their modifiers

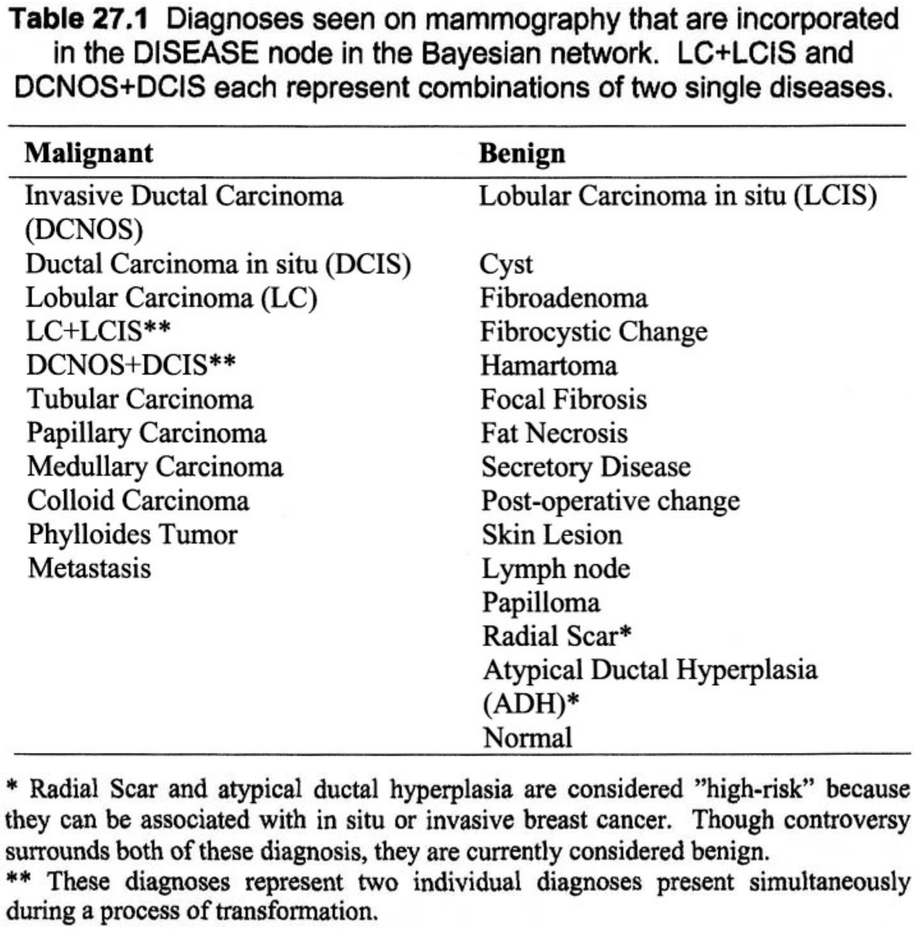


Table 27.1 Diagnoses seen on mammography that are incorporated in the DISEASE node in the Bayesian network. LC+LCIS and DCNOS+DCIS each represent combinations of two single diseases.

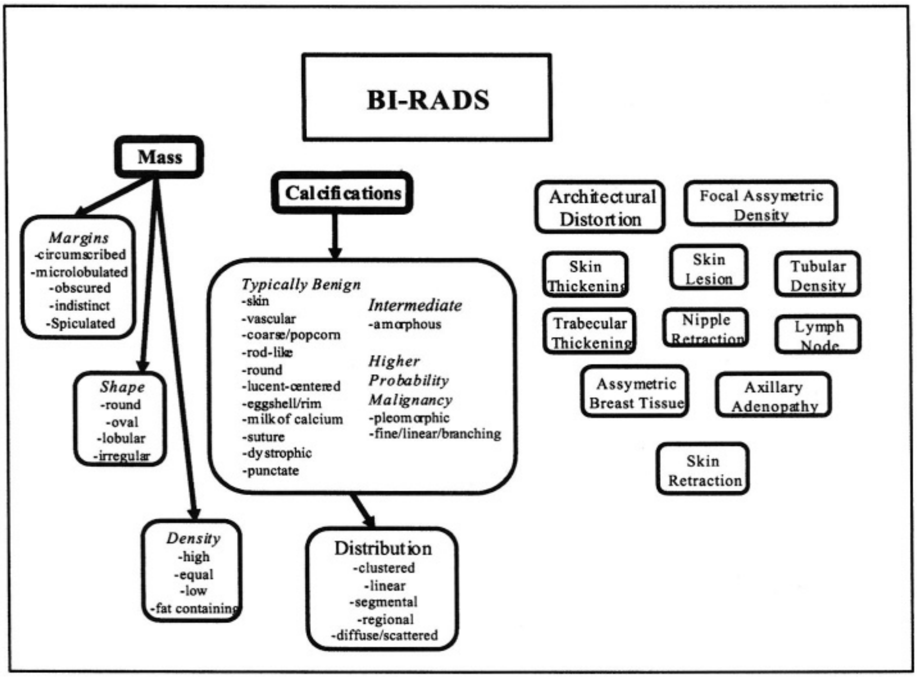
Malignant	Benign
Invasive Ductal Carcinoma (DCNOS)	Lobular Carcinoma in situ (LCIS)
Ductal Carcinoma in situ (DCIS)	Cyst
Lobular Carcinoma (LC)	Fibroadenoma
LC+LCIS**	Fibrocystic Change
DCNOS+DCIS**	Hamartoma
Tubular Carcinoma	Focal Fibrosis
Papillary Carcinoma	Fat Necrosis
Medullary Carcinoma	Secretory Disease
Colloid Carcinoma	Post-operative change
Phylloides Tumor	Skin Lesion
Metastasis	Lymph node
	Papilloma
	Radial Scar*
	Atypical Ductal Hyperplasia (ADH)*
	Normal

* Radial Scar and atypical ductal hyperplasia are considered ”high-risk” because they can be associated with in situ or invasive breast cancer. Though controversy surrounds both of these diagnosis, they are currently considered benign.
** These diagnoses represent two individual diagnoses present simultaneously during a process of transformation.

efficiently. For example, once a mass is identified, the user can describe the margins, shape, and density. The mass shape can be characterized by “detailed descriptors” such as round, oval, lobular, or irregular (Figure 27.1). Other examples of detailed descriptors are the modifiers of mass density: high, equal, low, and fat-containing (Figure 27.1).

We incorporated 38 of the BI-RADS descriptors into the model. We excluded five descriptors (skin thickening, trabecular thickening, nipple retraction, skin retraction, and asymmetric breast tissue) because they are rare, late, or non-contributory findings on screening mammography, and because they would have increased the complexity of the model without significantly improving its diagnostic effectiveness. Each descriptor we selected became a node in our Bayesian network. If that descriptor had detailed descriptors, they became distinctions for that node; otherwise the states for the node were “present” and “not-present.” For example, the

Figure 27.1 The BI-RADS terminology

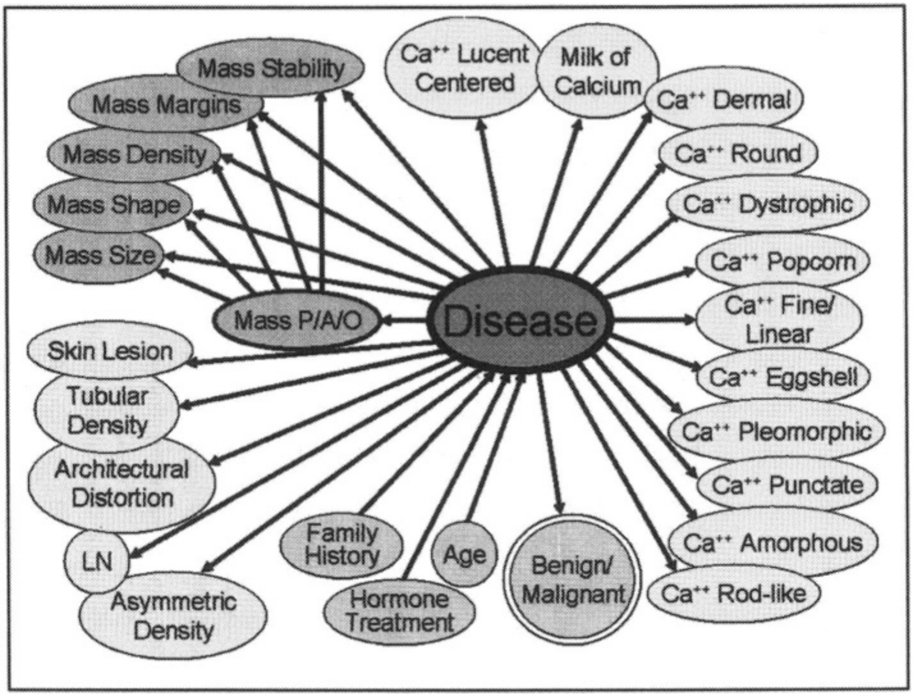


DENSITY node has states called “high,” “equal,” and “low”; the ROUND CALCIFICATION node has states “clustered,” “linear,” “segmental,” “regional,” and “diffuse/scattered”; the ARCHITECTUAL DISTORTION node has states “present” and “not present” (Figure 27.1).

We incorporated the following patient risk factors into our model: age (40-44, 45-50, 51-54, 55-60, 61-64, 65-70), history of breast cancer (strong, minor, or no history), and history of hormone treatment (less than 5 years, more than 5 years, or no history).

Bayesian Network To implement the model and perform inference given particular observations, we used the GeNIe modeling environment developed by the Decision Systems Laboratory of the University of Pittsburgh (http://www2.sis.pitt.edu/~genie/). In defining the structure of the model, we consulted with two experts in mammography to define the conditional dependencies among findings and diseases (Figure 27.2). The prior probability of disease is dependent on the patient risk factors, and these factors were believed to be conditionally independent of disease; thus disease has a separate parent for each risk factor. All BI-RADS descriptors except for those relating to a mass were believed to be conditionally

Figure 27.2 Bayesian network model of mammography diagnosis



independent manifestations of disease, so each descriptor has disease as a parent (Figure 27.2). The descriptors relating to a mass all depend on the presence of a mass, and we assume they become conditionally independent given the disease and whether the mass is present.

Normal structures of the breast can obscure masses on a mammographic image. This obscuration is more common in younger women and women on estrogen replacement therapy because they tend to have relatively dense breast tissue. While obscuration of the finding does not change the probability of disease given findings about the mass, it does decrease the probability that the mass and its features will be recognized. To model obscuration, we added an "obscured" state to joint distribution of the "mass" descriptor. Thus, a mass on the mammogram may be present, absent, or obscured, represented by the node "MASS P/A/O" which depends on the disease (Figure 27.2). (In later versions of the model, MASS P/A/O also depends on the patient's age and hormone treatment.)

The Bayesian network includes a deterministic node labeled "Benign/Malignant" which categorizes the diseases into these two distinct categories (Figure 27.2 and Table 27.1). This is useful because knowing the type of disease is important in determining correct management. For example, standard practice requires that all malignant diseases receive definitive therapy including excision. Some diseases such as radial scar and papilloma are controversial. These benign disease entities are sometimes associated with malignancy and therefore management depends on many factors. We have classified these diseases in the "benign" category pending farther data (Table 27.1).

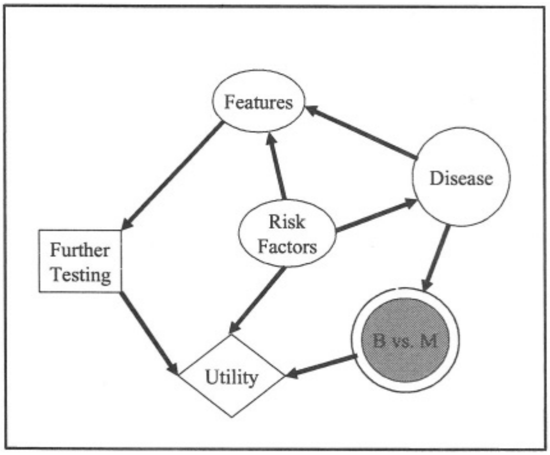
The initial values for the joint probability distributions were defined by consulting the experts and by reviewing the literature. We obtained the prior probabilities, age-specific and risk factor-specific distributions of diseases from census data and large randomized trials [16-18]. We derived many of the joint probabilities from studies of radiological/pathological correlation of individual breast diseases [19, 20]. Some of these initial probabilities evolved over time as we evaluated the model with test cases in which the correct diagnosis was known.

The Bayesian network we built calculates the probability of disease given these findings. The calculation depends on the pre-test probability of disease *d* (prevalence), given patient risk factors, and the joint probability distributions associated with the patient risk factors, disease, and the BI-RADS descriptor nodes in the Bayesian network (Figure 27.2). For a particular patient, values for risk factors and mammography findings are entered as observed evidence in the Bayesian network. If a particular finding or patient risk factor is not reported, then the corresponding node is unobserved. The joint probability distributions can then be updated using Bayes rule, giving a posterior probability distribution over the diseases, *P{d/f}*. This posterior probability can be used in several ways. Below we describe two ways we have used this information thus far: (1) to make a diagnosis on mammography, and (2) in evaluating concordance of mammography findings with biopsy results.

27.2.2 Using the model to make a diagnosis on mammography

Mammography is a screening test used to recognize breast cancer as early as possible when therapies are most effective and least debilitating. In the screening setting, observations are made to differentiate "benign" or "malignant" disease, and this diagnosis affects patient management decision making (Figure 27.3). If the screening mammogram has features suspicious for malignant disease, then the patient is called back for a more detailed diagnostic mammogram. If the probability of malignancy given all of the information available is high enough, then the radiologist will perform a biopsy for histological diagnosis. The influence diagram [21, 22] shown in Figure 27.3 represents these decisions. The critical distinction, represented by the deterministic node, B vs. M, is whether the disease is benign or malignant.

Figure 27.3 Influence diagram showing the radiologist's management decision whether to perform further tests which may lead to treatments



At the time of the "Further Testing" choice (whether to recall the patient for a diagnostic x-ray or to biopsy) the radiologist has observed features visible on the available mammogram(s). We assume that the patient's utility only depends on the malignancy of the underlying disease and whether the patient receives sufficient testing to confirm the diagnosis and initiate subsequent treatment. It is therefore critical that the set of findings observed on mammography be correctly translated into a probability of malignancy so that the correct decision about patient management can be made. Integrating the findings into a "benign vs. malignant" diagnosis is likely responsible for much of the variation among radiologist practice effectiveness in mammography.

Our Bayesian network model can be used to formulate a differential diagnosis (a ranked list of diagnoses in decreasing order of $P\{d|f\})$ by entering the patient risk factors and findings seen on mammography and calculating the posterior probability distribution over diseases. If the model puts a very high probability on a particular disease, this indicates that the model believes this is the mammographic diagnosis. If more than one disease shares similar probability mass, then the model suggests more than one diagnosis should be considered in the mammographic diagnosis. The model can also give the probability of benign or malignant disease from the probability distribution in the BENIGN/MALIGNANT node (Figure 27.2) which corresponds to the "B vs. M" node in the influence diagram (Figure 27.3).

To simplify the process of entering test cases into the network, we created a web-based data entry form (Figure 27.4). The web form corresponds to all observations that might be made for a particular lesion in a patient. The user enters the observations that apply to a patient and then submits the web form. We make a distinction between a feature that is observed to be present, a feature that is observed to be absent, and the lack of an observation about the feature. The evidence is submitted to our model and the posterior probability distribution is reported back to the user as a ranked differential diagnosis list, with the most probable diagnosis at the top of the list (Figure 27.5).

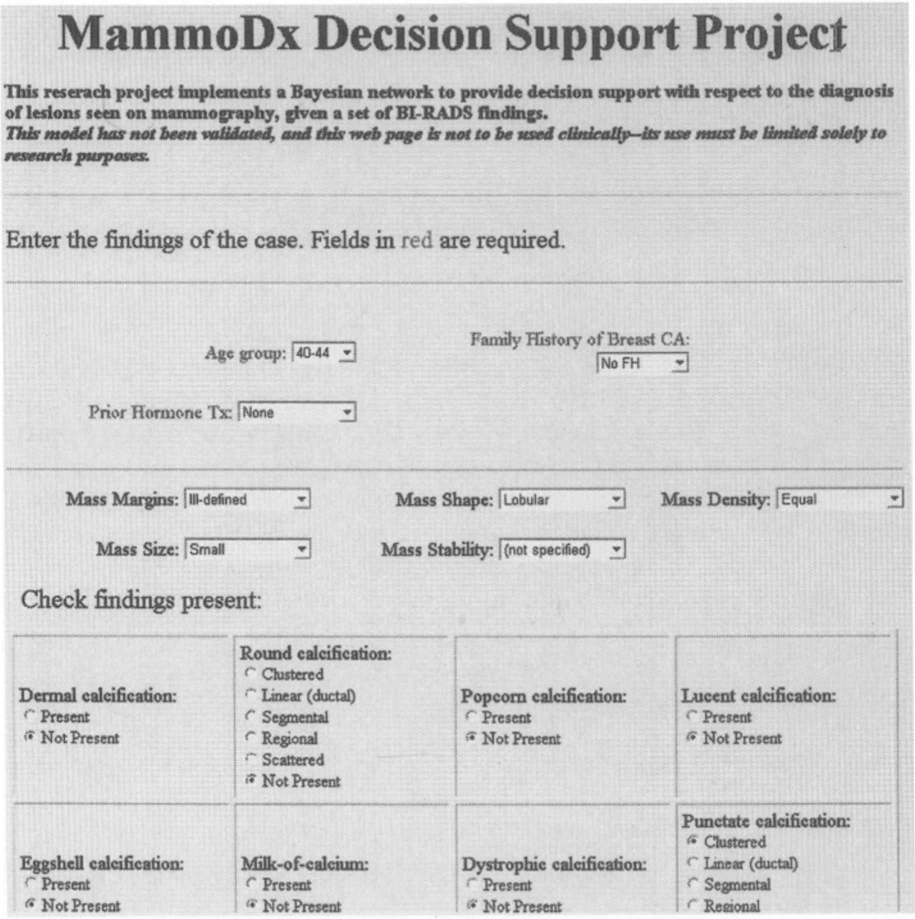
We evaluated the quality of mammographic diagnoses made by the model by entering several mammography cases in which the actual diagnosis was known (established by biopsy). In addition, we entered 105 cases from a teaching atlas of mammography [23] that contains sufficient clinical information and mammographic descriptors to enter these cases into the Bayesian network. To summarize the varying sensitivity and specificity at different probability thresholds, we built a receiver operating characteristic (ROC) curve using the ROCKIT 0.9B software (http://www-radiology.uchicago.edu/krl/toppage11.htm).

27.2.3 Using the model to evaluate concordance of mammography with biopsy results

Once a patient has a mammogram and the findings have been recognized by the radiologist, we can compute a post-test probability of malignancy. If that probability is high enough, then the radiologist will perform a biopsy so that a pathologist can make a more definitive, histological diagnosis.

Unfortunately, the biopsy test is imperfect and sampling error might occur. If the biopsy does not contain a sample of the lesion or the pathologist fails to observe the lesion cells, then the pathologist might fail to recognize malignant disease. Therefore, it is very important to correlate the histologic results from breast biopsy with the mammography findings [24-26]. The error rate can be as high as 3.3-6.2% in 14-gauge large-core needle biopsy, and 70% of these errors can be recognized immediately through careful correlation of the gross and/or histologic data with the mammography imaging findings [27-29]. Another tissue sampling technique, 11-gauge stereotactic vacuum-assisted biopsy, is associated with a lower but still significant sampling error rate, 0.8-1.7%. These sampling errors are also immediately detectable with careful imaging-histologic correlation [29-31].

Figure 27.4 A web form used to submit data on a patient to the Bayesian network model of mammography diagnosis (only a portion of the form is shown). Any finding that is not completed is treated as unobserved evidence.



MammoDx Decision Support Project

This reserach project implements a Bayesian network to provide decision support with respect to the diagnosis of lesions seen on mammography, given a set of BI-RADS findings.
This model has not been validated, and this web page is not to be used clinically—its use must be limited solely to research purposes.

Enter the findings of the case. Fields in red are required.

Age group: 40-44

Family History of Breast CA: No FH

Prior Hormone Tx: None

Mass Margins: Ill-defined **Mass Shape:** Lobular **Mass Density:** Equal

Mass Size: Small **Mass Stability:** (not specified)

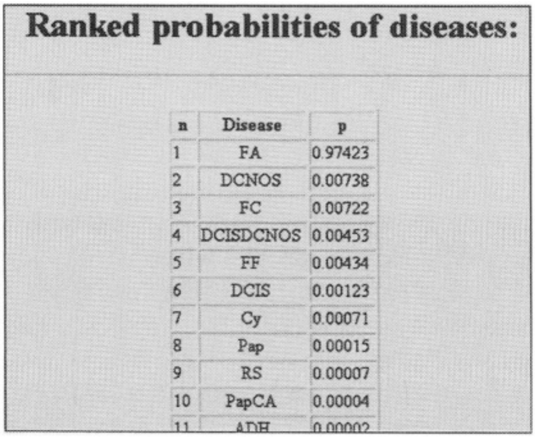
Check findings present:

Dermal calcification: ○ Present ● Not Present	**Round calcification:** ○ Clustered ○ Linear (ductal) ○ Segmental ○ Regional ○ Scattered ● Not Present	**Popcorn calcification:** ○ Present ● Not Present	**Lucent calcification:** ○ Present ● Not Present
Eggshell calcification: ○ Present ● Not Present	**Milk-of-calcium:** ○ Present ● Not Present	**Dystrophic calcification:** ○ Present ● Not Present	**Punctate calcification:** ● Clustered ○ Linear (ductal) ○ Segmental ○ Regional

Thus, breast imaging experts recommend that mammography images should be correlated with the pathology results [24-26]. This can be onerous in high volume settings, so an automated method to correlate biopsy results with mammography findings would be highly desirable.

Our model can assess concordance of mammographic image findings with histologic results from biopsy by recognizing that there can be a biopsy sampling error, which we denote by *miss*. Our model then uses the radiological findings, *f*, and the pathologist's reported disease, *Rd*, as evidence, and computes a posterior probability, $P\{miss|Rd,f\}$. This prob-

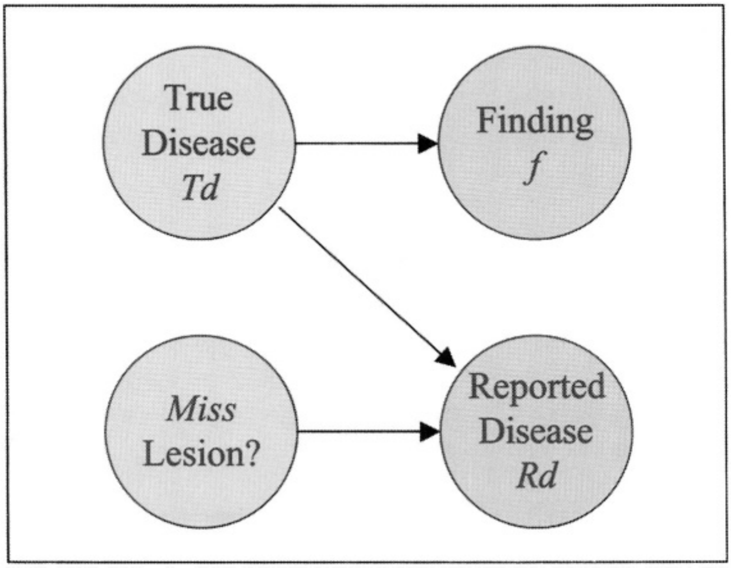
Figure 27.5 A list of diseases and the posterior probabilities. The diseases are ranked with the most likely disease first (an incomplete list is shown). These results are generated from the observations entered into the form in Figure 27.4.



Ranked probabilities of diseases:

n	Disease	p
1	FA	0.97423
2	DCNOS	0.00738
3	FC	0.00722
4	DCISDCNOS	0.00453
5	FF	0.00434
6	DCIS	0.00123
7	Cy	0.00071
8	Pap	0.00015
9	RS	0.00007
10	PapCA	0.00004
11	ADH	0.00002

Figure 27.6 Belief network showing a constant chance of biopsy sampling error, "Miss Lesion?", and the relationship between the true disease and the observed radiological findings and the pathologist's disease report. The radiologic finding and the disease report are observed, while the other nodes are not observed.



ability is based on the same relationships between diseases and findings that we discussed earlier. The radiologist should consider performing another biopsy when *Rd* for a benign disease has a high enough $P\{miss|Rd,f\}$, but if that probability is low enough or *Rd* is malignant then the radiologist can be confident that malignant disease has not been overlooked due a sampling error.

To obtain the formula for $P\{miss|Rd,f\}$, we need to make several assumptions, some of which are shown in the belief network (Figure 27.6): (1) the finding, *f*, and disease report, *Rd,* are observed while the true disease, *Td,* and whether there was a sampling error, *miss*, are not; (2) the finding is independent of the disease report and error if we knew the true disease; the sampling error is equally likely to happen with any patient and any findings, denoted P{*miss*}; and (4) if there is no sampling error, then *Rd*=*Td* and otherwise the disease will be observed at the prevalence rate, P{*Rd*}=P*{Td}*. In that case, given any finding, *f*, there are two possibilities: a sampling error which produces report *Rd* with probability P{*miss*}P{*d*} or a concordant diagnosis which yields report *Rd*=*Td* with probability $(1\text{-}P\{miss\})P\{d|f\}$.

$$P\{miss \mid Rd,f\} = \frac{P\{miss\}P\{d\}}{P\{miss\}P\{d\}+(1-P\{miss\})P\{d \mid f\}} = \frac{1}{1+\frac{(1-P\{miss\})P\{d \mid f\}}{P\{miss\}P\{d\}}}$$

Thus, our model can produce a probability indicating how likely a biopsy samples a lesion seen on mammography (i.e., whether the biopsy is concordant with mammography findings). This can be very useful to the radiologist to help identify those cases that are likely not to be concordant, and thus require further evaluation.

We evaluated the ability of our model to assess the concordance of breast biopsy results with mammography by entering cases into our model that had breast biopsy. We included 92 consecutive cases having 14-gauge and 11-gauge biopsies. A panel of expert radiologists reviewed each case and determined the concordance between the pathology and the mammography findings as concordant (“C”) or non-concordant (“N”). The experts used the following guidelines that are generally used in assessing concordance: (1) histologic documentation of microcalcifications when the mammographic abnormality contained microcalcifications; (2) histologic explanation for the imaging pattern (e.g., histologic explanation for a mass such as fibroadenoma or focal fibrosis in contrast to benign breast tissue); and (3) histologic explanation for abnormalities with a high pre-test probability of cancer (either a diagnosis of cancer or specific histology explaining the suspicious mammography findings) [26]. In our series of 92 cases, condition (1) was satisfied in all cases; thus, concordance hinged on agreement between the pathology report and the mammographic findings.

We used the concordance determination from this expert panel as our gold standard for testing our model, and compared these assessments with the probability ***P{miss|d,f}*** produced by the Bayesian network. Since different values of ***P{miss|d,f}*** can be used as a threshold for categorizing a case as "C" or "N," we constructed an ROC curve to quantify the performance of the Bayesian network across different thresholds in the concordance assessment task.

27.3 RESULTS

27.3.1 Using the model to make a diagnosis on mammography

We tested several cases (in which the diagnosis was known) to evaluate the behavior of the model. Table 27.2 shows the probability distribution for the following cases as well as the probability for the categorized diagnosis of "benign" and "malignant" disease. No probability is truly zero but many are rounded to zero when we only display four decimal places.

Case 1 A 40 year old female with no family history or hormone use has a mammogram which demonstrates a spiculated mass with associated linear and branching calcifications. According to literature and expert opinion, a spiculated mass is typical for ductal carcinoma. The branching calcifications suggest an intraductal component. In this case our model generates the following probabilities: DCNOS+DCIS diagnosis is most likely with a 95% post-test probability. DCNOS alone has a post-test probability of 4.5%, and DCIS alone is unlikely. Variations of this scenario illustrate how the probabilities change as features differ.

Case 2 A patient with similar demographic characteristics has a spiculated mass without calcifications detected on her mammogram. This finding elicits an increased post-test probability of DCNOS to 88%. DCNOS+DCIS decreases to 2.9%, and again DCIS is unlikely.

Table 27.2 Differential diagnosis as well as summation into management categories with associated post-test probabilities for example cases. Boldface type indicates the most likely diagnoses.

Disease	Pre-test	Case 1	Case 2	Case 3	Case 4
DCNOS	0.0090	**0.0451**	**0.8851**	0.0011	0.0047
DCIS	0.0019	0.0003	0.0000	**0.7053**	0.0000
DCNOS+ DCIS	0.0012	**0.9502**	**0.0285**	**0.0892**	0.0001
LC	0.0009	0.0000	0.0000	0.0007	0.0000
LCIS	0.0008	0.0000	0.0000	0.0007	0.0000
LC/ LCIS	0.0001	0.0000	0.0000	0.0001	0.0000
TubCA	0.0001	0.0005	0.0093	0.0000	0.0000
PapCA	0.0002	0.0000	0.0009	0.0014	0.0004
MedCA	0.0001	0.0000	0.0002	0.0000	0.0040
CollCA	0.0001	0.0000	0.0002	0.0000	0.0040
Phy	0.0010	0.0000	0.0000	0.0000	0.0005
Mets	0.0010	0.0024	0.0474	0.0001	0.0001
RS	0.0010	0.0000	0.0000	0.0008	0.0000
Cy	0.0700	0.0000	0.0002	0.0122	**0.7749**
FA	**0.1200**	0.0014	0.0274	0.0218	**0.1815**
FC	**0.1300**	0.0000	0.0003	0.1098	0.0000
Ham	0.0001	0.0000	0.0000	0.0000	0.0000
FF	0.0050	0.0000	0.0000	0.0014	0.0137
FN	0.0050	0.0000	0.0000	0.0042	0.0000
SecDis	0.0010	0.0000	0.0000	0.0008	0.0000
POC	0.0010	0.0000	0.0000	0.0008	0.0000
SL	0.0230	0.0000	0.0000	0.0181	0.0130
LN	0.0300	0.0000	0.0000	0.0253	0.0026
Pap	0.0020	0.0000	0.0002	0.0000	0.0004
Normal	**0.5945**	0.0000	0.0002	0.0051	0.0000
Benign	**0.9738**	**0.0014**	**0.0284**	**0.2007**	**0.9867**
Malignant	**0.0262**	**0.9986**	**0.9716**	**0.7973**	**0.0133**

Case 3 The only finding, in a similar patient, is linear calcifications in a clustered distribution. The post-test probability for DCIS increases to 70%. DCNOS+DCIS has a post-test probability of 8.9% and DCNOS alone is .001%. These probabilities are consistent with the pathophysiology of the disease as described above.

Case 4 A 50 year old patient has a mammogram that demonstrates a round, circumscribed mass. This is our example of a "probably benign" finding. Our system reveals that with these findings there is a 1.3% chance of malignancy in this setting. This is consistent with the radiology literature [32].

We also conducted an evaluation of the Bayesian network in a larger number of test cases selected from a teaching atlas (Section 27.3.2). The performance of the model on these test cases is summarized by the ROC curve shown in Figure 27.7. The area under the ROC curve is 0.95, which compares favorably to that of an earlier Bayesian network model, 0.88 [33]. In fact, our system compares favorably with several other computer diagnostic aids developed in the domain of screening mammography in which a similar area under the ROC curve methodology was used to evaluate these systems. Two different studies tested neural network models, reporting area under the ROC curve (Az) values of 0.85 [34] and 0.76 [35]. Finally, a survey of US radiologists evaluating performance used a test set containing 79 cases of which 45 were malignant (56% malignant). The average Az value for these radiologists was .85 [36]. We realize that the Az value can be influenced by the composition of the test set used to generate the ROC curve, but this was the benchmark used for the other procedures and allows a first-order comparison of the different methodologies. We believe that our results compared with the prior studies are encouraging and suggest that our model can assist in making a mammographic diagnosis.

27.3.2 Using the model to evaluate concordance of mammography with biopsy results

Of the 92 total cases evaluated for concordance, 3 were non-concordant. Thus, the non-concordance rate was 3.3%, which is comparable to that reported previously [24-29]. In most of the cases that the expert panel determined to be concordant, the model generated an extremely low $P\{miss|d,f\}$, strongly predicting concordance (Figure 27.8). The model's assessment of concordance between mammography and pathology, $P\{miss|d,f\}$, was extremely high ($P\{miss|d,f\}$ less than 0.02) in 75 of the 92 cases (all concordant cases). $P\{miss|d,f\}$ was 2-7% in 5 cases, and 23-28% in 3 cases; all of these cases were also concordant. In the remaining 9 cases, $P\{miss|d,f\}$ was 41% and greater; 3 of these cases were the ones considered

Figure 27.7 ROC analysis of 105 teaching cases. TPF: true positive fraction; FPF: false positive fraction.

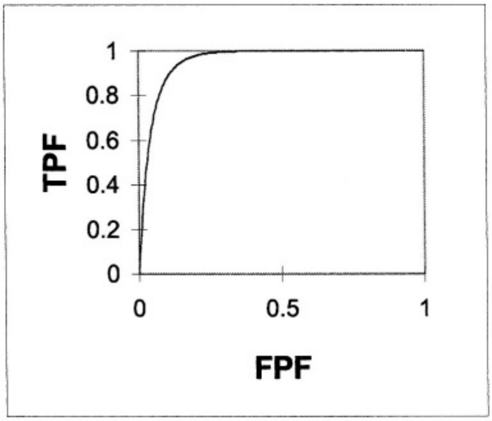
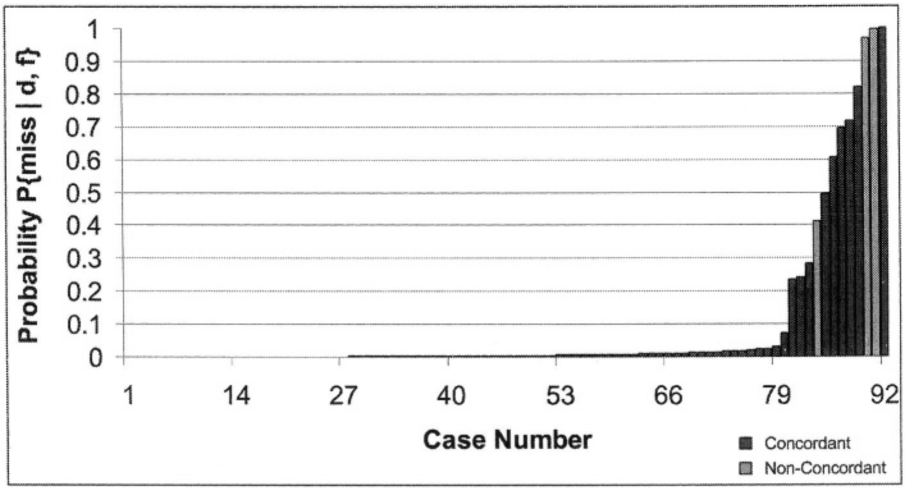
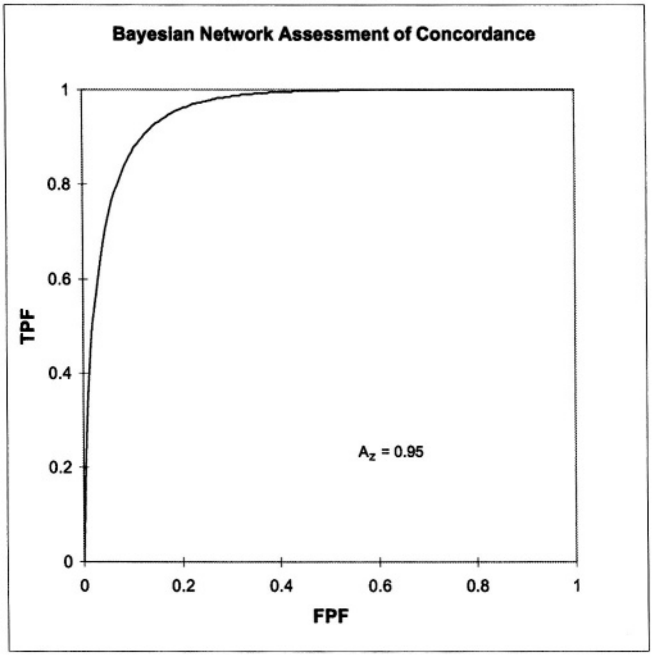


Figure 27.8 Histogram of 92 cases of mammography-biopsy correlation. In most of the cases that are concordant, the model predicts the probability of sampling error (*P{miss|d,f}*) is extremely low.



non-concordant. Using a threshold on *P{miss|d,f}* of 40% and assuming that detecting a non-concordant case is a "positive" case, there were 3 true positives, 83 true negatives, 6 false positives, and 0 false negatives (or 100% sensitivity and 93% specificity). Actually, it is more likely that a radiologist would prefer a lower threshold on *P{miss|d,f}*, such as 7%, for predicting non-concordance, which would lower the specificity to 90% with 100% sensitivity.

Figure 27.9 ROC analysis of performance of the model in assessing concordance of mammography findings with pathology. TPF: true positive fraction; FPF: false positive fraction.



An alternative summary of the performance of the model for assessing concordance is ROC analysis (Figure 27.9). The area under the curve on this graph is 0.95.

These results suggest that our model can discern those patients whose biopsy results are concordant with mammography findings. Consequently, many mammography-histologic correlations can be accurately assessed using the model, reserving only those cases where the model is uncertain (e.g., $P\{miss|d,f\}$ > 7%) for review by the radiologist. With the threshold of 7%, 80 of the 92 cases (87%) would not have required the radiologist to manually correlate mammography findings with histopathology. This is an important benefit because these correlations are labor-intensive and consume a considerable amount of radiologist time. In some busy practices, such correlation is not even routinely done. Even if the radiologist were very conservative with the model's predictions and reviewed any case for which the probability of sampling error given by the model is 1% or greater, more than half of the cases (55%) would not require manual review.

The probability value, $P\{miss|d,f\}$, produced by our system can be useful beyond establishing a threshold. First, it may encourage the pathologist to be more specific in benign disease diagnosis, to allow more concordance with the radiological findings. Second, a high probability suggests discordance and possible sampling error, prompting further review by the radiologist. We reviewed the cases that the model predicted to be discordant with high probability, but were called concordant by the expert panel. These cases were particularly complicated clinical scenarios, such as a diagnosis with an uncommon imaging presentation. These cases require expert evaluation, so a discordant assessment by the model is actually desirable behavior. Thus, our model can be useful in categorizing cases where concordance evaluation by a physician is needed.

27.4 AVENUES FOR FUTURE RESEARCH AND CONCLUSION

Our preliminary experience using a Bayesian network to model the uncertainties associated with mammography diagnosis appear promising. The model may provide several benefits. First, the model provides a normative approach to integrating findings observed by the radiologist into a ranked list of diagnoses. Second, the probabilities of disease given observed findings can be integrated with biopsy results to predict which cases are likely to be discordant, assisting patient management. Third, the model makes the mammographic decision making process explicit, providing radiologists of all levels of expertise a basis for communication and practice improvement.

The benefit of a normative approach in mammography diagnosis is greater consistency in mammography interpretation and subsequent improved health outcomes. Theoretically, if two patients have the same risk factors and the same abnormalities on mammography, they should have the same differential diagnosis and subsequent workup. However, previous studies have shown great variation in mammography practice [6-8]. Much of this variation can be attributed to how the patient risk factors and mammography findings are integrated into a differential diagnosis. Our model will produce consistent results with consistent inputs, so we would expect this to reduce some of the variation in mammography practice currently observed. Of course, this assumes that the radiologist is able to consistently detect the pertinent abnormalities on the mammogram in the first place, a fundamental task in radiology. There will still likely remain variation among radiologists with respect to identifying abnormalities on mammography images and assigning BI-RADS descriptors, but at least the variation in decisions based on these findings can be minimized using our Bayesian network.

Beyond assisting mammography diagnosis, the model's predictions and other information such as biopsy results can be integrated to assist with concordance assessment. We have shown that our model can identify those patients whose results are so likely to be concordant that they do not need to be manually reviewed by the radiologist. This would be particularly helpful in practices that currently do not do imaging-histologic correlation due to time constraints.

Our probabilistic approach is designed to support, rather than supplant, physician decision making. The radiologist is the ultimate decision maker regarding imaging-histologic correlation in difficult cases; our model can help identify those cases that are most suspicious and would benefit most from the expertise of the radiologist.

We have shown that our model performs well using the gold standard of a panel of expert radiologists. The gold standard for detecting breast biopsy sampling error is long term imaging and clinical follow-up to ensure that patients do not subsequently develop breast cancer. We plan to conduct studies evaluating our model using this preferred gold standard, which will allow us to ascertain better how well our model performs in assessing mammographic concordance. In the future, with refinement, our system may be able to improve the radiologist's ability to detect sampling error, using the gold standard of long term follow-up.

The radiology community has only incorporated a small portion of the BI-RADS descriptors into the decision-making process in this field. The entire lexicon, with its probabilistic underpinnings, when coupled with our Bayesian model has great potential to communicate quantitative probabilistic information that will aid management decisions. Our model relates benign and malignant breast diseases to BI-RADS descriptors and allows us to integrate radiological observations in a principled fashion. In addition, BI-RADS descriptors are crisply defined, with atlases showing examples of their proper usage, helping to reduce variability.

We are pursuing several other directions. First, we are collecting a larger series of confirmed cases to validate the model. This is important because we are continuing to improve our conditional probability distributions. Second, we are embarking on a large prospective data collection project in which routine mammogram cases will be compiled and their findings recorded to establish more accurately the probabilities of particular findings given disease. This information may be used to update conditional probability distributions in our Bayesian network that had little supporting data when it was originally built. Third, we will be evaluating the value of information of BI-RADS descriptors in the model which can suggest to the radiologist particular features on the mammogram that should be checked.

Finally, we wish to compare the diagnostic performance of the model directly with experts and non-experts to determine how well it can elevate performance of mammographers. Ultimately, our goal is to refine our system as an aid in normative decision making and education. We hope to demonstrate that the accuracy and quality of medical practice is elevated among practitioners of varying experience using this approach. We believe that with further testing and use our model will help to elevate the standard of all mammography practice and improve the quality of patient care.

References

[1] Greenlee, R.T., M.B. Hill-Harmon, T. Murray, and M. Thun (2001). Cancer statistics, 2001. *Cancer Journal for Clinicians,* 51, 15-36.

[2] Baker, L.H. (1982). Breast Cancer Detection Demonstration Project: five-year summary report. *Cancer Journal for Clinicians,* 32, 194-225.

[3] Houn, F., M.L. Elliott, and J.L. McCrohan (1995). The Mammography Quality Standards Act of 1992. History and philosophy. *Radiology Clinics of North America,* 33, 1059-1065.

[4] Pisano, E.D., *et al.* (2000). Has the Mammography Quality Standards Act affected the mammography quality in North Carolina? *American Journal of Roentgenology,* 174, 1089-1091.

[5] Sickles, E.A., D.E. Wolverton, and K.E. Dee (2002). Performance parameters for screening and diagnostic mammography: specialist and general radiologists. *Radiology,* 224, 861-869.

[6] Ciccone, G., P. Vineis, A. Frigerio, and N. Segnan (1992). Inter-observer and intra-observer variability of mammogram interpretation: a field study. *European Journal ofCancer,* 28A, 1054-1058.

[7] Elmore, J.G., et al. (2002). Screening mammograms by community radiologists: variability in false-positive rates. *Journal ofthe National Cancer Institute,* 94, 1373-1380.

[8] Elmore, J.G., C.K. Wells, C.H. Lee, D.H. Howard, and A.R. Feinstein (1994). Variability in radiologists' interpretations of mammograms. *New England Journal of Medicine,* 331, 1493-1499.

[9] Sirovich, B.E. and H.C. Sox, Jr. (1999). Breast cancer screening. *Surgery Clinics of North America,* 79, 961-990.

[10] Harris, R. (1997). Variation of benefits and harms of breast cancer screening with age. *Journal of the National Cancer Institute Monographs,* 139-143.

[11] Christiansen, C.L., et al. (2000). Predicting the cumulative risk of false-positive mammograms. *Journal of the National Cancer Institute,* 92, 1657-1666.

[12] Brown, M.L., F. Houn, E.A. Sickles, and L.G. Kessler (1995). Screening mammography in community practice: positive predictive value of abnormal findings and yield of follow-up diagnostic procedures. *American Journal of Roentgenology,* 165, 1373-1377.

[13] American College of Radiology (1998). *Breast Imaging Reporting and Data System (BI-RADS).* American College of Radiology, Reston, VA.

[14] Swets, J.A., et al. (1991). Enhancing and evaluating diagnostic accuracy. *Medical Decision Making,* 11, 9-18.

[15] Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference.* Morgan Kaufmann Publishers, San Mateo, CA.

[16] Colditz, G.A., et al. (1995). The use of estrogens and progestins and the risk of breast cancer in postmenopausal women. *New England Journal of Medicine,* 332, 1589-1593.

[17] Ries, L.A.G. and National Cancer Institute (U.S.). Division of Cancer Prevention and Control. (1997). *SEER Cancer Statistics Review, 1973-1996.* National Cancer Institute, Bethesda, MD.

[18] Slattery, M.L. and R.A. Kerber (1993). A comprehensive evaluation of family history and breast cancer risk. The Utah Population Database. *Journal of the American Medical Association,* 270, 1563-1568.

[19] Monsees, B.S. (1995). Evaluation of breast microcalcifications. *Radiology Clinics of North America,* 33, 1109-1121.

[20] Evans, W.P. (1995). Breast masses. Appropriate evaluation. *Radiology Clinics of North America,* 33, 1085-1108.

[21] Howard, R.A. and J.E. Matheson (1984). Influence diagrams. In *The Principles and Applications ofDecision Analysis,* R.A. Howard and J.E. Matheson, Eds., Strategic Decisions Group, Menlo Park, CA.

[22] Shachter, R.D. (1986). Evaluating influence diagrams. *Operations Research,* 34, 871-882.

[23] Tabár, L. and P.B. Dean (1983). *Teaching Atlas of Mammography.* Thieme Medical Publishers, New York.

[24] Berg, W.A., et al. (1996). Lessons from mammographic-histopathologic correlation of large-core needle breast biopsy. *Radiographics,* 16, 1111-1130.

[25] Ioffe, O.B., W.A. Berg, S.G. Silverberg, and D. Kumar (1998). Mammographic-histopathologic correlation of large-core needle biopsies of the breast. *Modern Pathology,* 11, 721-727.

[26] Liberman, L., et al. (2000). Imaging-histologic discordance at percutaneous breast biopsy. *Cancer,* 89, 2538-2546.

[27] Jackman, R.J., et al. (1999). Stereotactic, automated, large-core needle biopsy of nonpalpable breast lesions: false-negative and histologic underestimation rates after long-term follow-up. *Radiology,* 210, 799-805.

[28] Lee, C.H., L.E. Philpotts, L.J. Horvath, and I. Tocino (1999). Follow-up of breast lesions diagnosed as benign with stereotactic core-needle biopsy: frequency of mammographic change and false-negative rate. *Radiology,* 212, 189-194.

[29] Liberman, L. (2000). Centennial dissertation. Percutaneous imaging-guided core breast biopsy: state of the art at the millennium. *American Journal of Roentgenology,* 174, 1191-1199.

[30] Philpotts, L.E., N.A. Shaheen, D. Carter, R.C. Lange, and C.H. Lee (1999). Comparison of rebiopsy rates after stereotactic core needle biopsy of the breast with 11-gauge vacuum suction probe versus 14-gauge needle and automatic gun. *American Journal of Roentgenology,* 172, 683-687.

[31] Burbank, F. (1997). Stereotactic breast biopsy: comparison of 14- and 11-gauge Mammotome probe performance and complication rates. *American Surgeon,* 63, 988-995.

[32] Sickles, E.A. (1995). Management of probably benign breast lesions. *Radiology Clinics of North America,* 33, 1123-1130.

[33] Kahn, C.E., Jr., L.M. Roberts, K. Wang, D. Jenks, and P. Haddawy (1995). Preliminary investigation of a Bayesian network for mammographic diagnosis of breast cancer. *Proceedings of the Annual Symposium on Computing Applied to Medical Care,* 208-212.

[34] Baker, J.A., P.J. Kornguth, J.Y. Lo, M.E. Williford, and C.E. Floyd, Jr. (1995). Breast cancer: prediction with artificial neural network based on BI-RADS standardized lexicon. *Radiology,* 196, 817-822.

[35] Jiang, Y., et al. (1999). Improving breast cancer diagnosis with computer-aided diagnosis. *Academic Radiology,* 6, 22-33.

[36] Beam, C.A., P.M. Layde, and D.C. Sullivan (1996). Variability in the interpretation of screening mammograms by US radiologists. Findings from a national sample. *Archives ofInternal Medicine,* 156, 209-213.

28 OPTIMIZATION AND DECISION SUPPORT IN BRACHYTHERAPY TREATMENT PLANNING

Eva K. Lee[1,2] and Marco Zaider[3]

[1] Department of Industrial and Systems Engineering
Georgia Institute of Technology
Atlanta, GA 30332

[2] Department of Radiation Oncology
Emory University
Atlanta, GA 30322

[3] Department of Medical Physics
Memorial Sloan-Kettering Cancer Center
New York, NY 10021

SUMMARY

This chapter describes treatment planning optimization in brachytherapy and the design of a clinical decision support system. Brachytherapy refers to the placement of radioactive sources (seeds) inside a tumor site. The fundamental problem in treatment planning for brachytherapy is to determine where to place sources so as to deliver a sufficient radiation dose to kill the cancer, while limiting exposure of healthy tissue. We first present the sequence of steps that are involved in brachytherapy treatment planning. State-of-the-art mixed integer programming models are then described and some algorithmic approaches are outlined. The automated clinical decision support system allows for real-time generation of optimal seed configurations using ultrasound images acquired prior to seed implantation, and dynamic dose correction during the implantation process.

KEY WORDS

Integer programming, Decision support system, Radiation treatment, Cancer

28.1 INTRODUCTION

In recent years, technical advances in medical devices have led to a resurgence in the use of radioactive implants as an alternative or supplement to external beam radiation for treating a variety of cancers. This treatment modality, known as brachytherapy, involves the placement of encapsulated radionuclides ("seeds") either within or near a tumor [1]. In the case of prostate cancer, seed implantation is typically performed with the aid of a transrectal ultrasound transducer attached to a template consisting of a plastic slab with a rectangular grid of holes in it. The transducer is inserted into the rectum and the template rests against the patient's perineum. A series of transverse images are taken through the prostate, and the ultrasound unit displays the template grid superimposed on the anatomy of the prostate. Needles inserted in the template at appropriate grid positions enable seed placement in the target at planned locations.

Despite the advances in devices that assist in accurate placement of seeds, deciding *where* to place the seeds remains a difficult problem. A treatment plan must be designed so that it achieves an appropriate radiation dose distribution to the target volume, while keeping the dose to surrounding normal tissues at a minimum.

Traditionally, to design a treatment plan, several days (or weeks) prior to implantation the patient undergoes a simulation ultrasound scan. Based on the resulting images, an iterative process is performed to find a pattern of needle positions and seed coordinates along each needle that will yield an acceptable dose distribution. Adjustments are typically guided by repeated visual inspection of isodose curves overlaid on the target contours. Since the process requires manual inspection at each iteration, the process is not only lengthy – sometimes taking up to four hours to complete – but it also means that only a small fraction of possible configurations can actually be examined. More importantly, by the time the implantation is performed several days later, the prostate volume will have changed in both shape and size, making the pre-plan invalid.

In recent years, computer-aided iterative approaches and automated methods have been developed to aid in brachytherapy treatment planning in the operating room [2-10]. At Memorial Sloan-Kettering Cancer Center (MSKCC), we have developed a state-of-the-art intra-operative plan optimization system for permanent prostate implants [6, 9, 10, 15]. This chapter discusses these topics as applied to brachytherapy treatment for prostate cancer. The methodology described below reflects (unavoidably) accepted practice at MSKCC. In Section 28.2, we outline the sequence of steps that are involved in brachytherapy treatment planning. Mixed integer programming models used for designing treatment planning models are described in Section 28.3. Section 28.3 also outlines some algorithmic approaches that could be applied in solving the models and an illustrative example of solving an integer program. An integrated clinical decision support system is briefly summarized in Section 28.4.

28.2 BRACHYTHERAPY TREATMENT FOR PROSTATE CANCER

Treatment planning in brachytherapy consists of a sequence of steps that include the following: a) selection of appropriate sources, b) localization of potential source positions, c) dose prescription, d) treatment plan design and verification.

In brachytherapy, radioactive isotopes are selected based on two criteria: a) the energy of the ionizing particle, and b) the decay rate of the radionuclide. Low-energy sources are preferred because of their evident advantage in terms of radiation protection. They also offer better flexibility in designing conformal plans, as well as avoiding excess irradiation of healthy tissues that surround the target. The main benefit of high-energy sources is that (dosimetrically) they cover a larger volume and thus fewer sources may be needed [16-22].

In permanent prostate implants, potential source positions are localized with respect to a template that is placed in a fixed position relative to the treatment region (the prostate gland). The template, shown in Figure 28.1, has a rectangular pattern of holes; needles are inserted through the template grid, and seeds are placed along each needle at positions (typically, in multiples of 0.5 cm) determined by the treatment plan. A series of parallel ultrasound (US) images is taken through the prostate, and firmware in the ultrasound unit overlays a grid of dots onto these images that correspond to the template holes (Figure 28.2). The grid coordinates on the template and the distance of the US image away from the template uniquely identify the three-dimensional coordinates of each potential seed position relative to the gland anatomy. It is often the case that, as they penetrate into the prostate, inserted needles deviate from the initial grid coordinates. However modern planning systems have provisions for taking this into account in dosimetric calculations.

The dose prescription is made relative to the clinical target. Specifically, the recommended procedure is to use the minimum peripheral dose (mPD), which is the largest-dose isodose surface that completely surrounds the target (Figure 28.3) [11]. Often this is found to be too restrictive (and perhaps unrealistic in terms of being able to implement the plan). A different type of prescription makes use of D90, which means that one stipulates that a dose equal to or larger than the prescription dose be delivered to at least 90% of the target volume. The determination that the

Figure 28.1 Template used in a prostate permanent implant

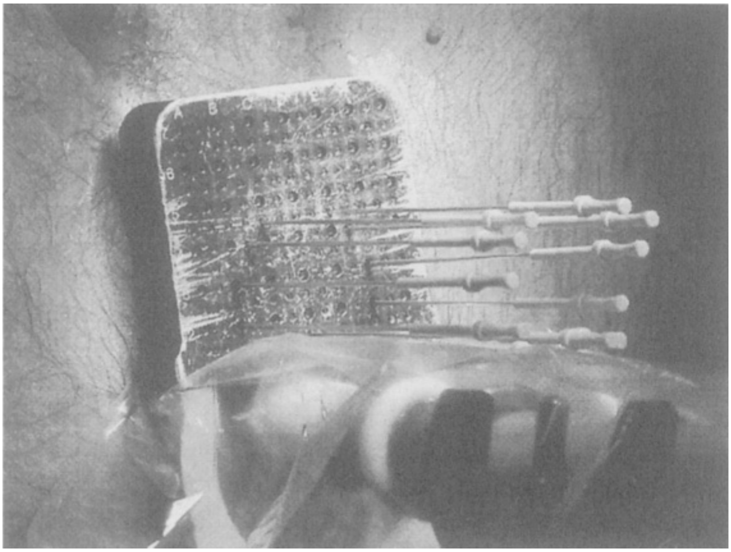
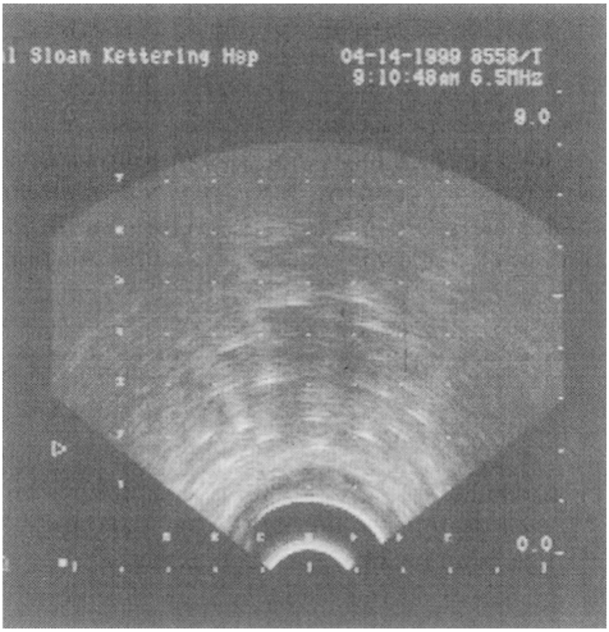


Figure 28.2 The grid superimposed on the ultrasound image provides the (x,y) coordinates of the inserted needles. Potential seed positions (x,y,z) are along these needles. The third coordinate (z) is determined by the position of the ultrasound probe (equivalently, by the image number).



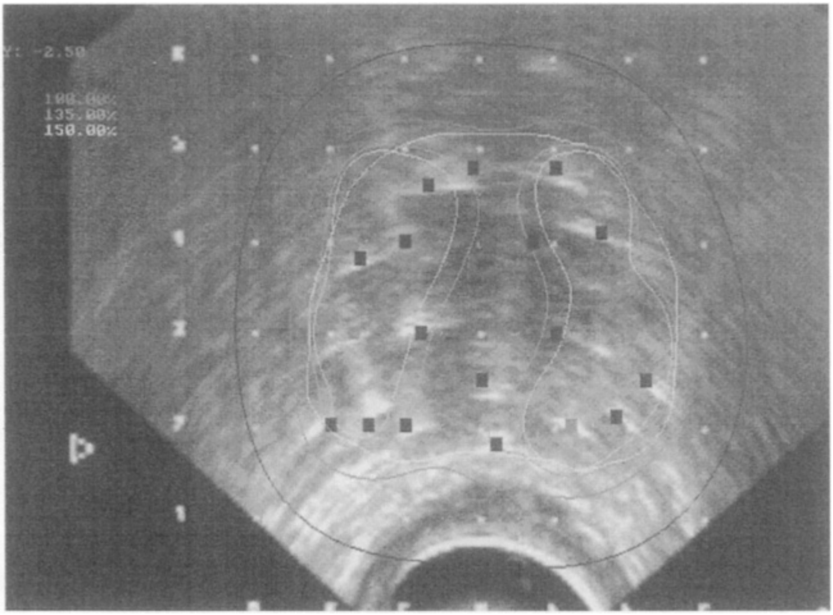
treatment plan achieves the required mPD or D90 can be made with the aid of dose-volume histograms (DVH), which plot as a function of dose, D, the probability that a randomly-selected voxel volume receives a dose of at least D (this is, of course, the cumulative probability distribution of dose in the target volume).

28.3 TREATMENT PLAN MODEL DESIGN AND ALGORITHMS

In brachytherapy, planning means finding a pattern of sources (of given strength) that is consistent with dosimetric constraints – typically, a minimum dose for the target and a maximum dose for the healthy tissues adjacent to the target. The search for this optimal source distribution may be performed using iterative (trial and error) methodology or – as described here – using computer-based optimization. For the latter, a mathematical model is usually developed which includes the essential dosimetric constraints, and an objective function (often user-specific). The objective function is a mathematical expression that measures the quality of the dose distribution. This metric can be selected according to the desire of the planner regarding the characteristics of the resulting plan. The model is then solved by some algorithms.

Search algorithms used in brachytherapy treatment planning include exact algorithms such as branch-and-bound, or heuristic approaches such as simulated annealing and genetic algorithms. A branch-and-bound algorithm is a tree search approach that works by searching through the set of all feasible plans (those that satisfy all the input constraints in the model) and returns an optimal plan that provides the best objective value. When allowed to run to completion, this approach will return a proven-optimal plan. Heuristic procedures work somewhat differently. Depending on the design, the search does not necessarily return plans that satisfy all of the imposed constraints; rather, the search attempts to obtain seed patterns that provide the least violation to the imposed constraints. Furthermore, there is no information on whether one obtains an optimal seed configuration or not. Instead, termination for heuristic algorithms is often based on the number of iterations that the user specifies. At each iteration, the algorithm obtains a plan and evaluates its associated objective function value. If the objective function is better than the incumbent plan, the incumbent plan will be updated.

Figure 28.3 Dose prescription for a permanent prostate implarlth this 2D ultrasound image the white line delineates the prostate and the 100% isodose line (mPD=144 Gy) is shown in green. Green dots indicate seeds and red dots show unused seed locations along needles.



28.3.1 Mixed integer programming based treatment planning models

The planning problem consists of determining if each possible source location should be implanted with a radioactive source or not. Hence the decision variables represent the locations of the grid position – a mathematical model that consists of integer decision variables [6, 9, 10]. In our case, it is the "yes" or "no" decision of placing a seed or not at each possible location. Mathematically, let x_j be a 0/1 decision variable for recording the placement of a source at grid position j. The total dose, D_P, at point P is given by:

$$D_P = \sum_j D\left(\left\|P - X_j\right\|\right) x_j \qquad (1)$$

where X_j is a vector that gives the coordinates of grid position *j*, $\|P - X_j\|$ is the Euclidean distance between P and X_j, and *D(r)* is the dose contribution to P from a source at distance r away (within the point source approximation). For each point of interest one can define upper (U_P) and lower (L_P) dose limits that D_P must satisfy:

$$\sum_j D\left(\left\|P-X_j\right\|\right)x_j \geq L_P \qquad (2)$$

$$\sum_j D\left(\left\|P-X_j\right\|\right)x_j \leq U_P. \qquad (3)$$

Generally, it is not possible to have all points P satisfy these constraints, in which case there will be no feasible solution to this linear system. Instead, one attempts to maximize the number of points that satisfy these inequalities. This is achieved in the following manner. Let LP + M_P denote the absolute maximum acceptable dose at point P, and similarly let LP - N_P denote the absolute minimum dose. Further, let v_P^L , v_P^U be binary (0/1) variables that indicate whether equation (2) or (3) is satisfied (when equal to 1) or not (when equal to 0). With this, the constraints, equations (2) and (3), can be replaced by:

$$\sum_j D\left(\left\|P-X_j\right\|\right)x_j + N_P(1-v_p^L) \geq L_P \qquad (4)$$

$$\sum_j D\left(\left\|P-X_j\right\|\right)x_j - M_P(1-v_P^U) \leq U_P , \qquad (5)$$

and the sum:

$$F\left(x, v^L, v^U\right) = \sum_P \left(v_P^L + v_P^U\right), \qquad (6)$$

which depends on the configuration x_j's, gives the total number of points that satisfy the original constraints, equations (2) and (3). The optimization problem consists of maximizing the objective function, F.

All points P need not have the same clinical importance: for instance, avoiding urethral toxicity (a common side effect) may be more important than satisfying the condition of dose uniformity across the target. This is addressed by assigning different weights, α_P and β_P, to v_P^L and v_P^U , respectively, and maximizing instead:

$$F\left(x, v^L, v^U\right)=\sum_P\left(\alpha_p v_P^L+\alpha_p v_P^U\right). \tag{7}$$

The problem described by objective (7) and constraints (4) and (5) is known as a linear integer programming (IP) problem because the objective function is linear in the unknown variables and these variables can take only integer (here 0 or 1) values. Other objectives can also be employed. For example, instead of maximizing the number of points that satisfy the original constraints (2) and (3), one can employ nonnegative continuous variables y_P^L and y_P^U to capture the deviations of the dose level at a given point from its target lower and upper bounds, respectively. In this case, equations (2) and (3) become

$$\sum_j D\left(\left\|P-X_j\right\|\right)x_j+y_P^L \geq L_P \tag{8}$$

$$\sum_j D\left(\left\|P-X_j\right\|\right)x_j-y_P^U \leq U_P, \tag{9}$$

and the sum F, to be minimized, is

$$F\left(x, y^L, y^U\right)=\sum_P\left(\alpha_p y_P^L+\alpha_p y_P^U\right). \tag{10}$$

When the target bounds L_p and U_p are expressed as multiples of a target prescription dose, T_p , another natural approach is to capture the deviations from T_p directly [12]. In our mixed-integer programming (MIP) model, this can be achieved by replacing constraints (8) and (9) with

$$\sum_j D\left(\left\|P-X_j\right\|\right)x_j+y_P=T_P, \tag{11}$$

where y_p is a continuous variable, unrestricted in sign. In the objective, one can then minimize the q norm of the vector y of all deviations; i.e.,

$$min\, F(x,y)=\left(\sum_P \left|y_P\right|^q\right)^{1/q}. \tag{12}$$

In this case, the problem becomes a *quadratic* 0/1 integer program. Details of these models, and variations that take into account other planning parameters and possible enhancements, can be found in [9, 10, 13, 14]. Algorithms commonly used for solving this problem are now described.

28.3.2 Computational algorithms

Branch-and-Bound The classical approach to solving linear 0/1 mixed integer programs is branch-and-bound. This is a tree search approach where, at each node of the tree, certain binary variables are fixed to zero or one, and the remaining binary variables are relaxed (i.e., allowed to assume any value between zero and one). This results in a linear program (LP) being associated with each node of the tree. The LP at the root node is simply the original 0/1 MIP instance with all of the binary variables relaxed. The tree is constructed such that the binary variables fixed in a parent node will be fixed identically in any of its children, and each child will have an additional binary variable fixed to zero or one. Typically, children are formed in pairs as follows. Assume that the LP at a given node is solved, and one or more of the relaxed binary variables is fractional in the optimal solution. One selects such a fractional binary variable and branches on it. That is, two child nodes are formed; one with the selected binary variable fixed to zero, and the other with the selected binary variable fixed to one. Of course, each child also inherits all of the fixed binary variables of its parent. Note that the objective value of a child node can be no greater (in the case of maximization) than the objective value of its parent.

If the linear program at a given node is solved and the optimal solution happens to have integral values for all the relaxed binary variables, then this solution is feasible for the original 0/1 mixed integer program. Once a feasible solution for the original problem is found, the associated objective value can be used as a lower bound (in the case of maximization) for the objective values of LP's at other nodes. In particular, if an LP at another node is solved, and its objective value is less than or equal to the lower bound, then none of its children could yield a feasible solution for the original MIP with a greater objective value than the one already obtained. Hence, no further exploration of this other node is needed, and the node is said to be fathomed. Two other criteria for fathoming a node are obvious: if the associated LP is infeasible, or if the optimal solution of the LP has integral values for all relaxed binary variables, then no further exploration of the node is required. In the latter case, the optimal objective value of the LP will be compared with the current lower bound, and the lower bound will be updated if needed. The tree search ends when all nodes are fathomed.

A variety of strategies have been proposed for intelligently selecting branching variables and nodes to process. However, no strategy stands out as being best in all cases. What has become clear from recent research in computational MIP is that branch-and-bound is most effective when coupled with other computational devices, such as problem preprocessing, primal heuristics, global and local reduced-cost fixing, and cutting planes. The reader can refer to the article by Lee [23] for a concise description of branch-and-bound methods for integer programming. The books by Schrijver [24], Nemhauser and Wolsey [25] and Parker and Rardin [26] contain detailed expositions of integer programming and related computational issues. Branch-and-bound algorithms designed for determining optimal seed locations for prostate implants can be found in [9, 10].

Genetic Algorithms were first proposed in connection with the optimal allocation of trials in 1973 [27, 28]. A genetic algorithm is a heuristic optimization method modeled on the biological mechanisms of evolution and natural selection (e.g., see [29, 30]).

In nature the characteristics of an organism are encoded in streams of DNA known as chromosomes. Likewise, in a genetic algorithm a potential solution to a problem is encoded as a stream of symbols over a given alphabet. Given an initial population of individuals (i.e., potential solutions encoded as symbol streams), a subset of the population is selected to parent offspring for the next generation. The parent selection process is stochastic, but biased towards selecting those individuals that are most fit, as measured by a pre-selected fitness function (e.g., the objective function that one is trying to optimize).

After the parents are selected, they are paired off and mated. That is, subsections of two parent symbol streams are interchanged, forming two new members for the next generation. This is analogous to cross-over in biological reproduction, where a child's genetic composition is a combination of its parents. Mutations are also possible. This is typically implemented by randomly selecting a child symbol stream and randomly altering one of its symbols. In order to ensure that the current best solution is not lost, the strategy of *elitism* can be employed; that is, the data stream with the highest fitness value is passed on unchanged to the next generation. This is implemented by simply overwriting one of the newly created children.

The algorithm can be terminated after a specified number of generations have been created (usually several thousands), or by examining when the difference between the maximum and minimum fitness values between consecutive generations remains less than a specified threshold for a number of generations. Upon termination, the individual in the final generation with the largest fitness value is selected as the operative solution to the problem at hand. Readers are referred to [4, 8, 9] for implementation of genetic algorithms for prostate implants.

Simulated Annealing, also referred to as Monte Carlo annealing, probabilistic hill climbing, statistical cooling, and stochastic relaxation [31], was first described as a heuristic for solving computer design problems [32] and the traveling salesman problem [33]. Simulated annealing is the application of statistical mechanics principles to combinatorial optimization. It has proven effective in generating near-optimal solutions for certain large problems.

Annealing is a process in which a solid is heated beyond its melting point and then cooled slowly and carefully into a perfect lattice. The crystalline structure of the perfect lattice represents a minimization of free energy for the solid. The cooling process determines if the ground state is achieved or if the solid retains a locally optimal lattice structure with crystal imperfections. The Metropolis algorithm [34] was developed to characterize cooling schedules that would produce favorable results. The central feature of the algorithm is the Metropolis condition: as the solid is cooled, the current configuration of the atoms is accepted with a certain probability and rejected otherwise. At nonzero temperatures, transitions out of local optima are always possible. Thus, the free energy is not monotonically decreased.

Simulated annealing applies these concepts to a combinatorial optimization problem. The cost function, or objective, assumes the role of the free energy function. The set of feasible solutions is analogous to the states of a solid. Let *f (i)* be the value of the cost function for solution *i.* Suppose the objective is to minimize *f.* A transition from state *i* to state *j* is accepted according to the following distribution:

$$P\{\text{accept j from i}\} = \begin{cases} 1 & \text{if f(j)} \leq \text{f(i)} \\ e^{\frac{f(i)-f(j)}{c}} & \text{if f(j)} > \text{f(i)} \end{cases}$$

The parameter *c* is an artificial "temperature" that is usually reduced as the number of iterations increases. At large values of *c,* large increases in the objective are accepted while for smaller values only small increases are accepted. Note that decreases in the objective are always accepted. The cooling schedule is the method by which *c* is decreased. A simple cooling schedule specifies c_0 as the initial temperature and a parameter ß such that c_k=ß^k c_0 [32]. Other cooling schedules have been proposed [31].

The implementation of simulated annealing requires the following: 1) a concise representation of the state space, 2) a method for randomly generating state transitions, 3) an objective function measuring the cost/benefit of transitions, and 4) cooling schedule parameters and a stop criterion [32].

Asymptotic convergence of the algorithm under various conditions on the generation and acceptance distributions has been proven [35]. The sequence of state transitions produces a discrete-time Markov chain. The method can be generalized to problems producing continuous time and continuous state space Markov chains [35, 36]. Solutions arbitrarily close to optimal usually require exponential run times, but the asymptotic behavior can be approximated in polynomial time [31]. The neighborhood structure of a problem determines which solutions are accessible in one transition from the current solution. For small problems, the neighborhood structure can have a large impact on the time to find good solutions. For problems with a large number of solutions and a relatively uniform distribution of values of the cost function, structure plays a lesser role [2, 3, 31, 37].

28.3.3 Illustrative example

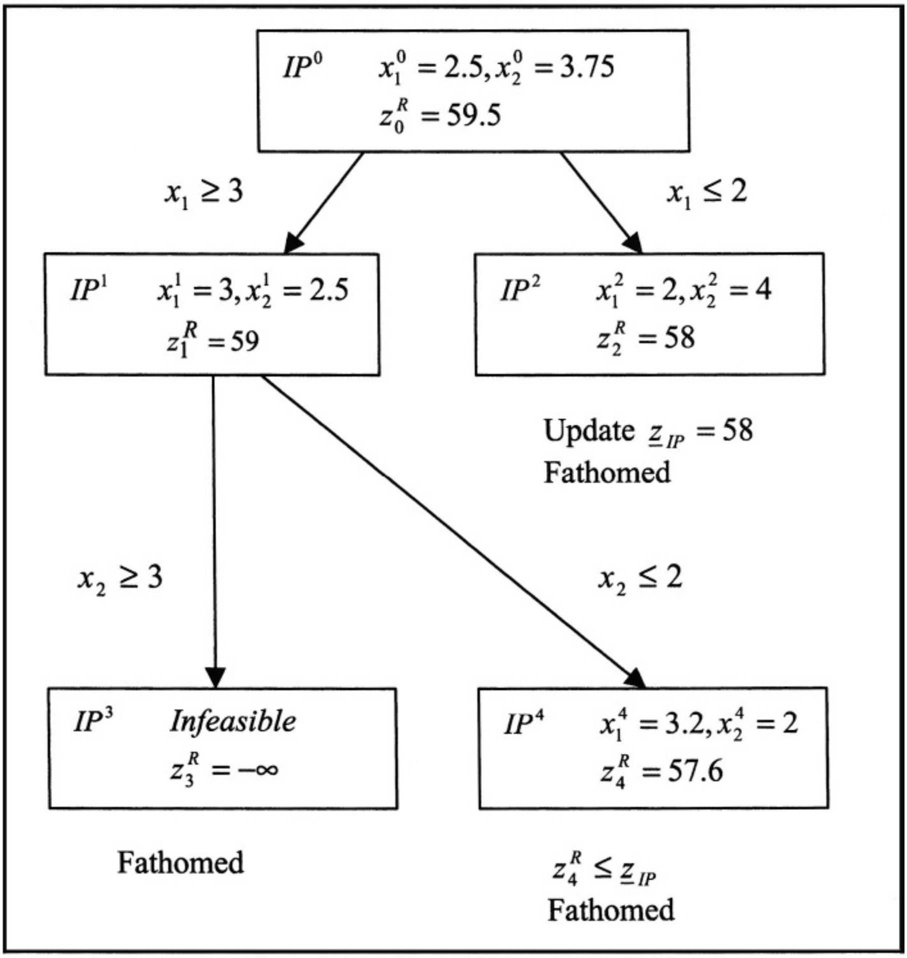
In this section, a two-variable integer program is solved using branch-and-bound [23]. The most infeasible integer variable is used as the branching variable, and best-bound is used for node selection. Consider the problem

$$\begin{array}{lll} \textbf{Maximize} & 13x_1 + 8x_2 & \\ \textbf{Subject to} & x_1 + 2x_2 \leq 10 & \\ & 5x_1 + 2x_2 \leq 20 & (IP^0) \\ & (13) & \\ & x_1 \geq 0, x_2 \geq 0 & \\ & x_1, x_2 \ \ \textbf{integer} & \end{array}$$

The progress of the algorithm is indicated in Figure 28.4. Each box contains the name of the subproblem, the solution to the LP relaxation, and its associated objective value.

Initially, the set of active problems, *L,* consists of just this problem IP^0. The solution to the LP relaxation is $x_1^0 = 2.5, x_2^0 = 3.75$, with value $z_0^R = 59.5$. The most infeasible integer variable is x_1, so two new subproblems are created, IP^1 where $x_1 \geq 3$ and IP^2 where $x_1 \leq 2$, and $L = \{IP^1, IP^2\}$.

Figure 28.4 Branch-and-bound example


IP^0 $x_1^0 = 2.5, x_2^0 = 3.75$
$z_0^R = 59.5$
$x_1 \geq 3$
$x_1 \leq 2$
IP^1 $x_1^1 = 3, x_2^1 = 2.5$
$z_1^R = 59$
IP^2 $x_1^2 = 2, x_2^2 = 4$
$z_2^R = 58$
Update $\underline{z}_{IP} = 58$
Fathomed
$x_2 \geq 3$
$x_2 \leq 2$
IP^3 *Infeasible*
$z_3^R = -\infty$
IP^4 $x_1^4 = 3.2, x_2^4 = 2$
$z_4^R = 57.6$
Fathomed
$z_4^R \leq \underline{z}_{IP}$
Fathomed

Both problems in L have the same bound, 59.5, so assume the algorithm arbitrarily selects IP^1. The optimal solution to the LP relaxation of IP^1 is $x_1^1 = 3, x_2^1 = 2.5$, with value $z_1^R = 59$. The most infeasible integer variable is x_2, so two new subproblems of IP^1 are created, IP^3 where $x_2 \geq 3$ and IP^4 where $x_2 \leq 2$, and now $L = \{IP^2, IP^3, IP^4\}$.

The algorithm next examines IP^2, since this is the problem with the best bound. The optimal solution to the LP-relaxation is $x_1^2 = 2, x_2^2 = 4$, with value $z_2^R = 58$. Since x^2 is integral feasible, $\underline{z}_{IP}$ is then updated to 58 and IP^2 is fathomed.

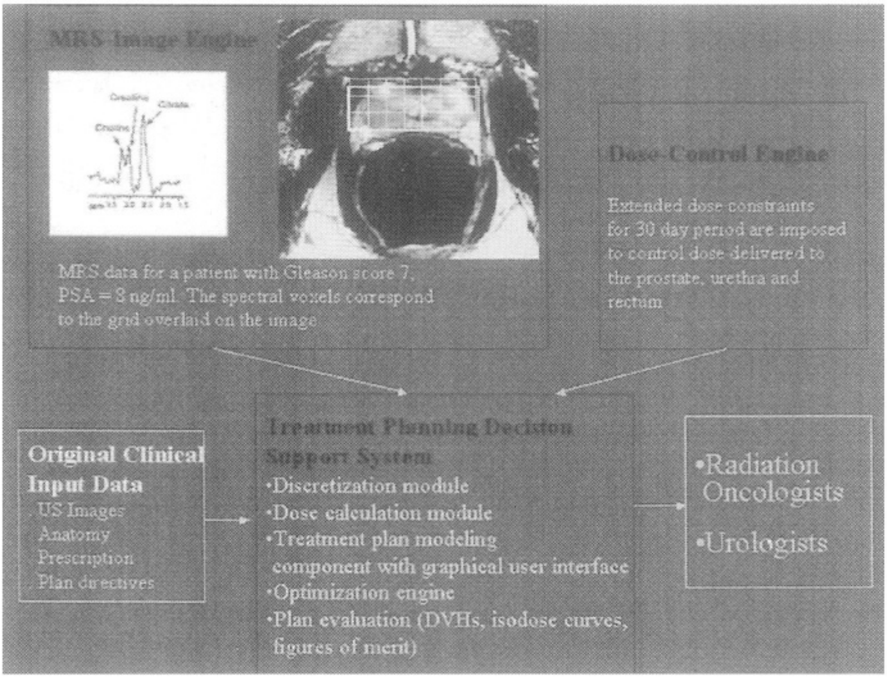
Both of the two subproblems remaining in L have the best bound greater than 58, so neither can yet be fathomed. Since these two subproblems have the same bound 59, assume the algorithm arbitrarily selects IP^3 to examine next. LP relaxation to this problem is infeasible, since it requires that x satisfy $x_1 \geq 3, x_2 \geq 3$ and $5x_1 + 2x_2 \leq 2$ simultaneously. Therefore, $z_3^R = -\infty$, and this node can be fathomed by bounds since $z_3^R \leq \underline{z}_{IP}$. That leaves the single problem IP^4 in L. The solution to the LP relaxation of this problem is $x_1^4 = 3.2, x_2^4 = 2$, with value $z_4^R = 57.6$. Since $z_4^R \leq \underline{z}_{IP}$, this subproblem can also be fathomed by bounds. The set L is now empty, hence x^2 is an optimal solution for the integer-programming problem IP^0.

28.4 SUMMARY – A CLINICAL DECISION SUPPORT SYSTEM FOR BRACHYTHERAPY TREATMENT PLANNING

A dose calculation engine, the MIP-based treatment modeling module, a branch-and-bound based optimization engine, and user evaluation tools are integrated into a computerized decision support system for optimal treatment planning design in the operating room. The decision support system takes into account the discretized US images, clinicians' planning prescriptions, radioactive source and activity, and dose information, and returns an optimal treatment plan prior to seed implantation. The system allows real-time dose correction, and allows physicians to handle unforeseen problems arising during the seed implantation process by enabling immediate re-optimization.

Figure 28.5 illustrates the flow diagram of the clinical decision support system. The flow within the basic system, with our MlP-based optimization module and solver incorporated, is shown in the lower portion of the figure. Information about two new modules is shown in the upper portion. Shown in the upper left is a module that allows for incorporation of MRS-images to identify high cell-proliferation tumor pockets within the gland [13]. In the upper right is a module that accounts for edema shrinkage and continuous dose absorption over a period of 30 days to assist in treatment and dose control over time [14]. Each module is designed and implemented in a manner that is conducive to re-engineering and future modification and expansion as experience with the system is gained. Thus, this in-house automated treatment-planning system remains amenable to modifications that reflect newly acquired clinical knowledge [13-15].

Figure 28.5 The flow design of the computerized decision support system used for brachytherapy



MRS Image Engine

MRS data for a patient with Gleason score 7, PSA = 8 ng/ml. The spectral voxels correspond to the grid overlaid on the image

Dose-Control Engine

Extended dose constraints for 30 day period are imposed to control dose delivered to the prostate, urethra and rectum

Original Clinical Input Data

US Images
Anatomy
Prescription
Plan directives

Treatment Planning Decision Support System

- Discretization module
- Dose calculation module
- Treatment plan modeling component with graphical user interface
- Optimization engine
- Plan evaluation (DVHs, isodose curves, figures of merit)

- Radiation Oncologists
- Urologists

References

[1] Interstitial Collaborative Working Group. (1990). *Interstitial Brachytherapy. Physical, Biological, and Clinical Considerations.* Raven Press, New York, 1990.

[2] Sloboda, R.S. (1992). Optimization of brachytherapy dose distribution by simulated annealing. *Medical Physics,* 19, 964.

[3] Pouliot, J., D. Tremblay, J. Roy, and S. Filice (1996). Optimization of permanent I-125 prostate implants using fast simulated annealing. International *Journal of Radiation Oncology Biology Physics,* 36, 711-720.

[4] Yu, Y. and M.C. Schell (1996). A genetic algorithm for the optimization of prostate implants. *Medical Physics,* 23, 2085-2091.

[5] Silvern, D.A., E.K. Lee, RJ. Gallagher, L.G. Stabile, R.D. Ennis, C.R. Moorthy, and M. Zaider (1997). Treatment planning for permanent prostate implants. Genetic algorithms versus integer programming. *Medical and Biological Engineering Computing,* 35, Suppl, Part 2, 850.

[6] Gallagher, R.J. and E.K. Lee. (1997). Mixed integer programming optimization models for brachytherapy treatment planning. *Proceedings of the American Medical Imaging Association Annual Fall Symposium,* 278-282.

[7] Silvern, D.A (1998). Automated OR prostate brachytherapy treatment planning using genetic optimization. PhD Dissertation, Columbia University, New York, NY.

[8] Yang, G., L.E. Reinstein, S. Pai, Z. Xu, and D.L. Carroll. (1998). A new genetic algorithm technique in optimization of permanent I-125 prostate implants. *Medical Physics,* 25, 2308-2315.

[9] Lee, E.K., RJ. Gallagher, D. Silvern, C.S. Wu, and M. Zaider (1999). Treatment planning for brachytherapy. An integer programming model, two computational approaches and experiments with permanent prostate implant planning. *Physics in Medicine and Biology,* 44, 145-165.

[10] Lee, E.K. and M. Zaider (2003). Mixed integer programming approaches to treatment planning for brachytherapy. *Annals of Operations Research, Optimization in Medicine,* 119, 147-163.

[11] Anderson, L.L., R. Nath, A.J. Olch, et al. (1991). American Endocurietherapy Society recommendations for dose specifications in brachytherapy. *Endocurietherapy Hypertherm Oncolology,* 7, 1.

[12] Brahme, A. (1995). Optimization of the 3-dimensional dose delivery and tomotherapy. *International Journal of Imaging Systems and Technology,* 6, 1.

[13] Zaider, M., M. Zelefsky, E.K. Lee, K. Zakian, H.A. Amols, J. Dyke, and J. Koutcher (2000). Treatment planning for prostate implants using MR spectroscopy imaging. *International Journal of Radiation Oncology Biology Physics,* 47, 1085-96.

[14] Lee, E.K. and M. Zaider (2001). Determining an effective planning volume for permanent prostate implants. *International Journal of Radiation Oncology Biology Physics,* 49, 1197-1206.

[15] Lee, E.K. and M. Zaider (2003). Intra-operative dynamic dose optimization in permanent prostate implants. *International Journal of Radiation Oncology Biology Physics,* 56, 854-861.

[16] Wuu, C.S. and M. Zaider (1998). A calculation of the relative biological effectiveness of 125I and 103Pd brachytherapy sources using the concept of proximity function. *Medical Physics,* 25, 2186-2189.

[17] Wuu, C.S., P. Kliauga, M. Zaider, and H.I. Amols (1996). Microdosimetric evaluation of relative biological effectiveness for 103Pd, 125I, 241Am, and 192Ir brachytherapy sources. *International Journal of Radiation Oncology, Biology, Physics,* 36, 689-697.

[18] Ling, C.C., W.X. Li, and L.L. Anderson (1995). The relative biological effectiveness of I-125 and Pd-103. *International Journal of Radiation Oncology, Biology, Physics,* 32, 373-378.

[19] Nath, R., A.S. Meigooni, and A. Melillo (1992). Some treatment planning considerations for pd-103 and I-125 permanent interstitial implants. *International Journal of Radiation Oncology, Biology, Physics,* 22, 1131-1138.

[20] Zellmer, D.L., J.D. Shadley, and M.T. Gillin (1994). Comparisons of measured biological response and predictions from microdosimetric data applicable to brachytherapy. *Radiation Protection Dosimetry,* 52, 395-403.

[21] Zellmer, D.L., M.T. Gillin, and J.F. Wilson (1992). Microdosimetric single event spectra of yb-169 compared with commonly used brachytherapy sources and teletherapy beams. *International Journal of Radiation Oncology, Biology, Physics,* 23, 627-632.

[22] International Commission on Radiation Units and Measurements (1980). *Radiation Quantities and Units.* International Commission on Radiation Units and Measurements, Washington, DC.

[23] Lee, E.K. (2001). Branch-and-bound methods. In Mauricio, G.C., Resende, Pardalos, P.M., Eds., *Handbook of Applied Optimization.* Oxford University Press, New York.

[24] Schrijver, A. (1986) *Theory of Linear and Integer Programming.* Wiley, Chichester, UK.

[25] Nemhauser, G.L. and L.A. Wolsey (1988). *Integer and Combinatorial Optimization.* Wiley, New York.

[26] Parker, R.G. and R.L. Rardin (1988). *Discrete Optimization.* Academic Press, Boston, MA.

[27] Holland, J.H. (1974). Erratum. Genetic algorithms and the optimal allocation of trials. *SIAM Journal on Computing,* 3, 326.

[28] Holland, J.H. (1973). Genetic algorithms and the optimal allocation of trials. *SIAM Journal on Computing,* 2, 88-105.

[29] Buckles, B.P. and F. Petry (1992). *Genetic Algorithms.* IEEE Computer Society Press, Los Alamitos, CA.

[30] Wasserman, P.D. (1993). *Advanced Methods in Neural Computing.* Van Nostrand Reinhold, New York.

[31] Aarts, E.H.L. and J. Korst (1989). *Simulated Annealing and Boltzmann Machines: A Stochastic Approach to Combinatorial Optimization and Neural Computing.* Wiley, Chichester, UK.

[32] Kirkpatrick, S., C.D. Gelatt, and M.P. Vecchi (1983). Optimization by simulated annealing. *Science,* 220, 671-680.

[33] Cerny, V. (1985). Thermodynamical approach to the traveling salesman problem. An efficient simulation algorithm. *Journal of Optimization Theory and Applications,* 45, 41-51.

[34] Metropolis, N.A., A. Rosenbluth, M. Rosenbluth, A. Teller, and E. Teller (1953). Equation of state calculations by fast computing machines. *Journal of Chemical Physics,* 21, 1087-1092.

[35] Aarts, E.H.L., J.H.M. Korst, and J.K. Lenstra (1997). Simulating annealing 8. In Aarts, E.H.L. and J.K. Lenstra, Eds., *Local Search in Combinatorial Optimization.* Wiley, Chichester, UK, 91-120.

[36] Hajek, B. (1985). A tutorial of theory and applications of simulated annealing. *Proceedings of the 24th Conference on Decision and Control,* 755-759.

[37] Anderson, LL. (1993). Plan optimization and dose evaluation in brachytherapy. *Seminars in Radiation Oncology,* 3, 290-300.

29 RADIOTHERAPY TREATMENT DESIGN AND LINEAR PROGRAMMING

Allen Holder

Department of Mathematics
Trinity University
San Antonio, TX 78212

SUMMARY

Intensity modulated radiotherapy treatment (IMRT) design is the process of choosing how beams of radiation will travel through a cancer patient to treat the disease, and although optimization techniques have been suggested since the 1960s, they are still not widely used. Instead, the vast majority of treatment plans are designed by clinicians through trial-and-error. Modern treatment facilities have the technology to treat patients with extremely complicated plans, and designing plans that take full advantage of the technology is tedious. The increased technology found in modern treatment facilities makes the use of optimization paramount in the design of successful treatment plans. The goals of this work are to 1) present a concise description of the linear models that are under current investigation, 2) develop the analysis certificates that these models allow, and 3) suggest future research avenues.

KEY WORDS

Mathematical programming, Intensity modulated radiotherapy treatment

29.1 INTRODUCTION

Fast proliferating cells, such as those found in cancerous and displasiac tissue, are more sensitive to radiation than healthy cells, and this fact has allowed tremendous strides in the fight against cancer. Chemotherapy exploits this property by injecting radioactive substances into the blood stream, with the goal being to administer enough radiation to kill the cancerous cells but not enough to kill the healthy cells. Because the substances are injected into the blood stream, chemotherapy affects the entire anatomy, and hence, all fast proliferating cells are attacked (such as hair cells). Intensity modulated radiotherapy treatment (IMRT) is a similar cancer treatment where external beams of radiation are focused on the cancerous regions. Since the radiation is not injected into the blood stream, IMRT is a local treatment. In fact, the radiation beams can be focused with sub-millimeter precision, giving a medical physicist precise control of how the radiation travels through the anatomy.

IMRT design is the process of deciding how beams of radiation will travel through a patient so that they deliver a tumoricidal dose of radiation to the cancerous region. At the same time, the critical structures surrounding the cancer are to receive a limited dose of radiation so that they can survive the treatment. The effect of IMRT on the tumor, which we want to receive a high level of radiation, and the surrounding critical structures, which we do not want to receive a high level of radiation, makes IMRT design a complicated process. Moreover, modern treatment facilities have the technology to deliver extremely complicated treatment plans, which in turn allows a great amount of flexibility in the design process. Indeed, the amount of flexibility that is permitted makes optimizing treatment plans beyond the scope of human comprehension, and future technology will only increase the degree of complication. If the IMRT planning process does not advance with the technology patients will not receive the added benefits of the advanced technology. So, the development of optimization models that take full advantage of emerging technology is critical.

Three groups of specialists are important to the success of improved treatment design: 1) oncologists, who tend to the needs of the patients, 2) medical physicists, who know how to model the deposition of radiation, and 3) the operations researchers, who are experts in the field of applied optimization. One of the difficulties of working in this field as an operations researcher is to find an oncologist and/or a medical physicist to work with, for a continued dialog between these three groups is important. Historically, the bulk of the research on radiotherapy treatment design was accomplished by oncologists and medical physicists. Only in the last few years have operations researchers become interested in these problems. Because optimization is playing an increasing role in treatment design, it makes sense that operations researchers be included, as they are accustomed to investigating algorithms and performing solution analysis. The goal of this chapter is to encapsulate the current research directions so that operations researchers can quickly become familiar with the interdisciplinary field of designing IMRT plans. Before we continue, we mention two Internet resources. The *Operations Research & Radiation Oncology* web site at http://www.trinity.edu/aholder/HealthApp/oncology/ has a depository of recent papers and a list of interested researchers. The other resource is *PubMed,* the index to the National Library of Medicine, located at http://www.ncbi.nlm.nih.gov/. A recent search on "optimization" and "radiotherapy" found 905 related citations, so the medical literature is immense. Most of these articles are case studies about specific types of cancer and are not directly related to operations research. The bibliography at the end of this chapter is designed to help those who are interested get started in the field.

29.2 MODELING DOSE DEPOSITION

To understand how and why linear optimization models appropriately model the planning of IMRT design, one needs to have a basic understanding of how radiation is deposited into the anatomy. The basic question is how does a focused beam of radiation deposit energy as it travels through a patient. The question has two perspectives. A forward problem is one in which we know the amount of energy being transmitted along the beam, and we want to know how much energy is deposited at a point in the anatomy. An inverse problem is one where we know how much energy is to be deposited in the anatomy, and we want to know what beam energies attain the desired amounts. IMRT planning is an inverse problem because we limit the amount of radiation received by certain tissues and find a collection of beam intensities that adhere to these bounds. While a complete discussion of the physics describing the dose deposition is beyond the scope of this article, we briefly explain a continuous model and its discrete counterpart (see [1] and [2] for more complete details).

We begin with a description of the equipment found in a standard treatment facility (see Figure 29.1). The beams of radiation are formed by a linear accelerator, and once formed, they travel through a gantry that is capable of rotating around the patient (the center of the rotation is called the *isocenter*). The fact that the gantry can rotate around the patient is important because this allows the beam of radiation to be directed at the patient from any angle. The head of the gantry is designed to accommodate one of several focusing apparatuses, with most modern facilities using a *multileaf collimator* (see Figure 29.2). This device can "shape" the beam of energy by blocking portions of the beam. Shaping the beam has tremendous benefits because it allows sensitive regions to receive relatively low levels of radiation, while nearby tumorous regions receive a higher amount of radiation. (We point out that a tumor is not well defined because microscopic extensions, which may or may not exist, do not appear in an image. As such, treatment planners define a “tumorous region” that they believe includes the tumor and the possible microscopic extensions.) A planning model must take into consideration the fact that the beams can be focused on the patient from any angle and in almost any shape.

Figure 29.1 A gantry is capable of rotating around the patient as he or she lies on the couch. The head of the gantry contains a multileaf collimator (see Figure 29.2).

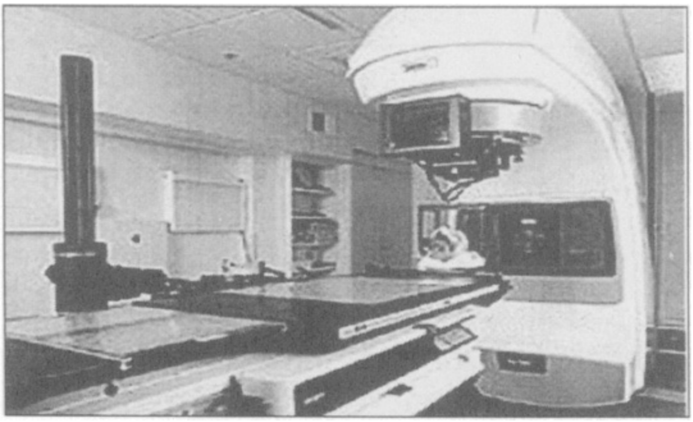


Figure 29.2 A multileaf collimator is used to ‘shape’ the beam of radiation so that surrounding tissues are shielded.

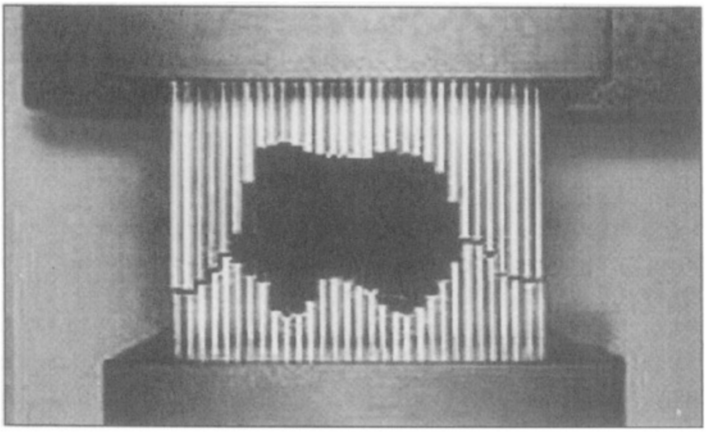
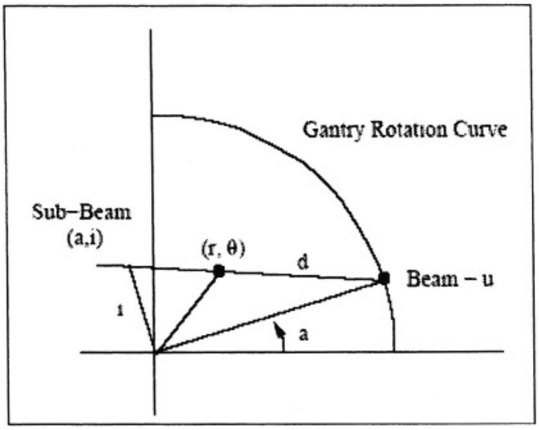


Figure 29.3 The geometry of a continuous dose deposition calculation



Consider the geometry depicted in Figure 29.3, where we want to calculate the dose at (r,θ). The gantry in this figure is located at angle *a*, and the sub-beam from angle *a* that passes through (r,θ) is *i*. The amount of energy that is to be transmitted along sub-beam *(a,i)* is *p(a,i)*. The distance from the cell to the surface of the body, denoted by *d* in the diagram, affects the radiation received by the cell. This is because the beam *attenuates* as it travels through the body, meaning that it deposits more radiation when it first enters the body and 'decays' as it travels through the tissue. This attenuation is modeled as exponential decay -i.e. by $e^{-\mu d}$ where μ depends on the particular beam of energy formed by the linear accelerator. Thus, the radiation deposited at location (r,θ) by sub-beam *(a,i)* is *p(a,i)* $e^{-\mu d}$. To calculate the total, or *integral,* dose at point (r,θ) we need to accumulate the amount deposited from every possible sub-beam that passes through (r,θ). Letting $L = \{f(a,i)$: sub-beam *(a,i)* passes through $(r,\theta)\}$, we find that the integral dose is:

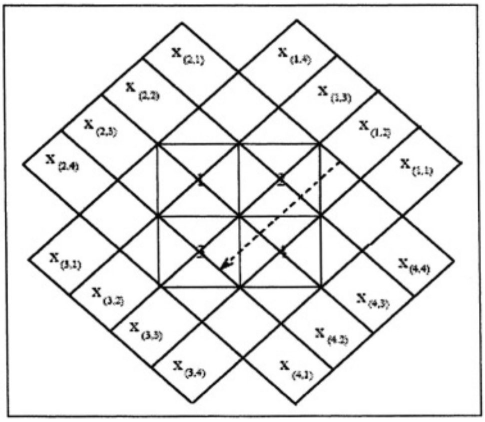
$$D(r,\theta) = \int_L p(a,i)e^{-\mu d}\,da\ .$$

Again, we point out that if we know *p(a, i)* and want to calculate $D(r,\ \mu)$, then we are dealing with a forward problem. However, IMRT planning is an inverse problem because we limit the amount of radiation to be deposited into the tissue and use these limits to calculate the amount of energy to deliver along each sub-beam li.e. we bound $D(r,\ \mu)$ and want to calculate *p(a, i)* to satisfy the bound. So, for continuous IMRT planning we need to invert an integral transformation. There are a variety of techniques to accomplish this task, but the difficulty lies in the fact that the calculation of *p(a, i)* must keep it non-negative, which is not guaranteed by these techniques. The hidden assumptions on the integrand are what often make inverse problems more difficult than forward problems.

In the continuous model we are integrating with respect to the angle *a,* so in the discrete model there are a finite number of angles, denoted $a_1, a_2, \cdots, a_\Theta$. We assume that each angle is comprised of sub-beams, which may be *elementary beams* or *pencils,* the difference being that pencils radiate from a point source and elementary beams run parallel to each other. Our development does not depend on whether pencils or elementary beams are chosen, only that there are a finite number of them. The patient image is divided into $N \times M$ pixels, and we want to measure the amount of radiation that is deposited into each pixel.

We let $x_{(a,i)}$ be the dose along the *i*th sub-beam of angle *a,* and $d_{(p,a,i)}$ be the distance from where sub-beam $x_{(a,i)}$ enters the image to where it reaches the center of pixel *p.* We further define $A_{(p,a,i)}$ to be the product of $e^{-\mu d(p,a,i)}$ and the geometric area common to both the sub-beam $x_{(a,i)}$ and pixel *p.* For example, in Figure 29.4 we have a 2×2 patient image surrounded by 4 angles, each with 4 sub-beams (in this case they are elementary beams). The elementary beam corresponding to $x_{(1,2)}$ intersects half of pixel 3, and the distance to the center of this pixel along this elementary beam is $3\sqrt{2}/2$ (assuming that each pixel has a width of one). Hence, $A_{(3,1,2)} = \frac{1}{2}e^{-3\sqrt{2}\mu/2}$.

Figure 29.4 A discretized approximation to the continuous dose deposition calculation



The components of the *dose deposition matrix,* denoted by *A,* are $A_{(p,a,i)}$, where the rows of *A* are indexed by *p* and the columns are indexed by *(a, i).*

Similarly, a *treatment plan,* or more succinctly a plan, is a non-negative vector x whose components are $x_{(a,i)}$, where the order corresponds to the columns of A. So, x is the vector of energies at the gantry, and the linear transformation $x \rightarrow Ax$ deposits the radiation into the anatomy.

Let pixel p contain point (r,θ) from the continuous model. The dose calculation at (r,θ) is approximated by

$$D(r,\theta)=\int_L p(a,i)e^{-\mu d}da \approx \sum_{(a,i)} A_{(p,a,i)}x_{(a,i)}=[Ax]_p$$

where the last notation indicates that the integral dose to pixel p is the pth component of Ax. Both the continuous and the discrete models are linear in the energy transmitted along the sub-beams |i.e. the continuous model is linear in p and the discrete model is linear in x. Physical measurements show that the integral dose to a cell is a linear function of the amount of energy transmitted along the sub-beams. So, the linear models are not crude approximations, but rather they accurately measure how radiation is deposited into the anatomy. With that said, the linear operator $x \rightarrow Ax$ only approximates the dose deposition because it does not take into account the effects of scattering. The problem here is that some radiation "bounces" off cells and scatters into areas where it was not intended. There are non-linear models that measure scattering [2, 3], but once the scattering for a particular patient is taken into account, the dose to a cell is linear in the energy transmitted along the sub-beams. Because each patient is unique, the linear coefficients depend on the patient, and in a clinical setting these linear coefficients are decided during an initial planning appointment. For the purposes of this article, we use the technique discussed above to calculate the dose deposition matrix.

The rows of A are partitioned into the rows that represent the cancerous regions, the critical structures, and the remaining healthy tissue. This reordering is represented by the submatrices A_T, A_C, and A_G, as indicated below:

$$A=\begin{array}{cl} \left[\begin{array}{c} A_T \\ A_C \\ A_G \end{array}\right] & \begin{array}{l} \leftarrow \text{Tumor} \\ \leftarrow \text{Critical Structures} \\ \leftarrow \text{Remaining HealthyTissue.} \end{array} \end{array}$$

Sub-beams that do not intersect the tumor are removed from consideration by eliminating the columns of A that have a corresponding zero column in A_T. For notational brevity, we keep the A notation for the sub-matrix with these columns removed. In what follows, $A \in \Re^{m \times n}$, $A_T \in \Re^{m_T \times n}$, $A_C \in \Re^{m_C \times n}$ and $A_G \in \Re^{m_G \times n}$. Because radiation is measured in Grays (Gy), the right-hand sides of the constraints are given in units of Gy. Allowing e to be the vector of ones, where length is decided by the context of its use, we can guarantee that the tumor pixels receive 80Gy, that the critical structures receive no more than 40 Gy, and that the remaining tissue receives no more than 90Gy by finding a nonnegative x that satisfies

$$A_T x \geq 80e, \quad A_c x \leq 40e \quad \text{and } A_G x \leq 90e.$$

If these are the only treatment goals, the design process is a feasibility problem, meaning that any nonnegative vector satisfying these constraints forms a suitable plan [4, 5].

29.3 TREATMENT CONCERNS

The primary goal of IMRT design is to construct a treatment plan that delivers a tumoricidal dose to the cancerous region and at the same time delivers low enough radiation levels to the surrounding tissues so that they maintain functionality. However, several factors make this overriding objective difficult to translate into an optimization model. When first presented with the problem, most operations researchers believe that the objective should be to deliver as much radiation as possible to the tumor. This naively makes sense because killing the cancerous cells is the purpose of the treatment. There are two reasons why maximizing the amount of radiation deposited into the tumor is not an appropriate objective. First, healthy and cancerous cells are often interspersed, and there is a limited range of radiation that will kill a cancerous cell and allow a healthy cell to survive. So, it is important to deliver enough radiation to kill the cancerous cells, but not enough to kill the healthy cells within the tumor. This is usually accomplished by the dosimetrist stating that he or she wants the cancerous regions to attain a specified amount of radiation plus or minus some percentage. For example, the tumor should receive 80 Gy ±2% means that the tumor should receive between 78:4Gy and 81:6Gy. Second, if any region of the anatomy receives an unreasonably high amount of radiation, the cells within this region are killed. If the area is large enough, the human physiology is disrupted, causing a condition known as necrosis. For these two reasons, it is paramount for the tumor to receive a uniform, tumoricidal dose of radiation and not simply as much as possible.

Treatment planning is further complicated by the fact that different organs react to radiation in different ways. For example, the liver can receive a large amount of radiation over a substantial portion of its tissue and maintain its functionality. However, if the entire liver receives a relatively low dose of radiation, the organ will fail. The colon is different because it can handle a relatively low, uniform dose but will fail if a small region receives a high dose. Organs like the liver that can successfully receive a high level of radiation over a portion of their tissue, but fail under a relatively low, uniform dose, are called *rope* organs. A *chain* organ is one that can handle a relatively low amount of radiation over its entirety but will fail when a small amount of the tissue is destroyed (see [6-9] for more complete details on rope and chain organs). So, in addition to making sure that the tumor receives a uniform, tumoricidal dose, the dosimetrist must make sure that the treatment plan delivers radiation to the critical structures in a suitable manner.

The hope that every patient receiving IMRT is cured of cancer is unrealistic, and because of this, patients and physicians routinely make difficult decisions about a course of treatment. The best of all situations is when the type and stage of cancer being treated has a high probability of cure with standard treatments. The "best" treatment plan in such a case is one that delivers a tumoricidal dose to the cancerous regions and as little radiation as possible to the critical structures. The treatment goals change if a patient's illness is terminal or the standard treatments are not promising. For terminally ill patients, destroying the cancer is not the primary objective, but rather treatments are often designed to increase the patient's quality of life. In some instances, this means that some nearby regions should receive no radiation. For example, in the case of a brain tumor it may be best to minimize the radiation deposited into adjoining regions that control speech and memory. However, minimizing the amount of radiation that is received by these regions can mean that the tumor is not treated with a uniform, tumoricidal dose, but because the patient's illness is terminal, this is not detrimental to the treatment plan. In cases where the standard treatments do not provide a high probability of success, the question becomes to what degree are the patient and physician willing to risk the nearby regions to treat the tumor with higher amounts of radiation.

The point of highlighting these situations is that the objective of treatment is not the same for all patients and is decided by the ethics and values of the patient and physician. This ethical perspective of the objective is different from the modeling perspective of the objective, which is concerned with how we measure and penalize deviation from the physician's demands. The fact that ethical concerns often make it difficult to clearly state a primary objective means that the optimization model needs to be flexible enough to accommodate several scenarios.

Before continuing with the mathematical models, we discuss two treatment concerns that are not readily discussed in the literature, but that are beginning to receive some attention. First, radiation therapy is not delivered in a single session but is instead fractionated into several treatments (usually 20 to 30). Once a plan is developed, it is divided into a number of treatments, and the patient receives these fractionated treatments daily. The idea here is to accumulate the radiation slow enough so that the healthy tissue has an increased chance of survival. This fractionization is the difference between radiotherapy and radiosurgery, with the latter being delivered all at once. A natural, but virtually unexplored, question is whether it is beneficial to deliver the overall dose in non-uniform increments. The optimization model associated with this question is a challenging optimal control problem, with the only work being the recent paper of Ferris and Voelker [10]. While the computational burden of solving their model makes it impossible to develop a patient-specific course of treatment, their work clearly indicates that delivering a uniform, fractionated dose is not typically optimal. There are many related questions that are open for investigation, such as deciding the number of fractionated treatments that maximize the success of the treatment.

The second relatively new treatment question is how to move the gantry and adjust the multileaf collimator so that the plan is delivered in as little time as possible (see [11]). This is an extremely important question because typical treatments last about 15 minutes, and if treatment plans can be delivered more efficiently, more complicated plans are possible. In general, plans with more than 5 to 7 angles are considered complicated because of the time it takes to administer them. Hence, the number of angles in a treatment plan is restricted, and the flexibility of the design is limited. The restriction on the number of angles becomes less of a concern as we find more efficient ways to move the gantry and adjust the multileaf collimator. This increased efficiency allows more flexibility in the design process and translates into a benefit for the patient.

29.4 OPTIMIZATION MODELS

In this section we develop a class of linear programs that aid IMRT design [12, 13]. The first optimization model that was developed to aid IMRT design was linear and appeared in the literature in 1968 [14]. Since then, many researchers have experimented with linear models [15-22].

While linear models are natural because dose deposition is experimentally linear, these models have been the focus of several complaints, and many other researchers have investigated nonlinear models [18, 19, 21]. The first complaint about linear models is that the physician's demands often produce an empty feasible region. For example, the physician may desire that the cancerous tissue receives 80Gy±2% and that the surrounding critical structures receive no more than 20Gy. This translates into the following constraints,

$$78.4e \le A_T x \le 81.6e, \quad A_C x \le 20e, \quad x \ge 0, \tag{1}$$

which may or may not be consistent. If the physician's demands are not possible, the optimization routine simply states that the underlying optimization problem is infeasible and provides no information about how to adjust the physician's desires. Finding and explaining a source of infeasibility is a difficult question, and there is a substantial amount of literature that deals with this issue (e.g., [24-27]). Since we cannot ask the physicians or the physicists to become experts in mathematical programming, this is a problem that needs to be addressed. The linear models that we develop overcome this difficulty by using elastic constraints.

The second major complaint about linear models has nothing to do with the linearity of the problem but rather the solution technique. The problem here is that the simplex algorithm terminates with an extreme point solution, which means that some of the inequality constraints are guaranteed to hold with equality at the solution. For example, in (1) we are guaranteed that either some of the cancerous regions are going to receive their upper bound of 81.6Gy, or that some of the cancerous regions are going to receive their lower bound of 78.4Gy, or that some of the critical structures are going to receive their upper bound of 20Gy. The problem here is that we are guaranteeing that some regions are going to attain the limits placed on them, and this is alarming because these limits are rules-of-thumb. We address this issue in two ways. First, we use a path-following interior point algorithm to solve our problems. This algorithm terminates with a solution that strictly satisfies as many inequalities as possible. So, we find an optimal plan that does not attain the limits placed on the regions, provided that such a plan is possible. Second, the elastic constraints that we use allow the physician's desires to 'float' during the optimization process, and the objective is to better them as much as possible.

From the dimensions of A, A_T, A_C, and A_G we have that m is the total number of pixels, m_T is the number of tumorous pixels, m_S is the number of critical structure pixels, and $m_G = m - m_T - m_C$ is the number of remaining pixels. A *prescription* comprises a physician's aspirations for the tumor, usually a tumoricidal dose, and upper bounds for the non-tumorous tissue. Specifically, a prescription is the 4-tuple *(TUB, TLB, CUB, GUB),* where

- *TUB* is a m_T vector of upper bounds for the tumor,
- *TLB* is a m_T vector of lower bounds for the tumor,
- *CUB* is a m_C vector of upper bounds for the critical structures, and
- *GUB* is a m_G vector of upper bounds for the remaining good tissue.

We make the realistic assumptions that $0<TLB \leq TUB$, $0 \leq CUB$, and $0 \leq GUB$. Because a uniform, tumoricidal dose is to be delivered to the tumor, the lower and upper bounds for the tumor pixels are a fixed percentage of the physician's goal for the tumor. So, if the physician's goal for a tumorous cell is *TG,* values for TUB_i and TLB_i are (1+*tol*)*TG* and (1 - *tol*)*TG,* respectively. Here, *tol* is the percentage of variation permitted over the cancerous region and is called the *tumor uniformity level.* Typical values of *tol* found in the literature range from 0.02 to 0.15. The vector *GUB* describes the highest amount of radiation that any single pixel is allowed, and in general no tissue should receive more than 10% of the tumor's desired dose. Hence, we set *GUB* = (1.1) *TG.*

The model that we use separates how we measure and penalize any deviation from the physician's goals. This generality is permitted through the use of *semimonotone* matrices, which are matrices whose Moore-Penrose generalized inverse is nonnegative (see [28] for more information). For the remainder of this chapter, the following semimonotone matrices are assumed to have full column rank: $l \in \Re^{q_T}, u_c \in \Re^{q_C}, u_G \in \Re^{q_G}, L \in \Re^{m_T \times q_T}$, $U_C \in \Re^{m_C \times q_c}$ and $U_G \in \Re^{m_G \times q_G}$. We further assume that l, u_C and u_G are positive, and that L, U_C and U_G are nonnegative with no row sum being zero -i.e. $Le>0$, $U_c e > 0$ and $U_{Ge} > 0$. Any collection of l, u_C u_G L, U_C and U_G satisfying these assumptions defines a set of *elastic functions.* The feasible region, denoted by *F,* is the collection of $x \in \Re^n$, $\alpha \in \Re^{q_T}$, $\beta \in \Re^{q_C}$ and $\gamma \in \Re^{q_G}$ that satisfy the following constraints:

$$
\begin{aligned}
& TLB - L\alpha \leq A_T x \leq TUB \\
& A_C x \leq CUB + U_C \beta \\
& A_G x \leq GUB + U_G \gamma \\
& 0 \leq L\alpha \leq TLB \qquad (2) \\
& -CUB \leq U_C \beta
\end{aligned}
$$

$$0 \leq U_G\gamma$$

$$0 \leq x$$

We note that because L and U_G are semimonotone, α and γ are nonnegative.

The constraints $TLB - L\alpha \leq A_T x$, $A_C x \leq CUB + U_C\beta$ and $A_G x \leq GUB + U_G\gamma$ are called *elastic* because the bounds are allowed to vary with the vectors α, β and γ respectively. The matrices L, U_C and U_G define how we measure the amount of elasticity, and with this in mind, we see that the assumption that $Le > 0$, $U_c e > 0$, and $U_G e > 0$ makes sure that each constraint is elastic. The elastic constraints are incorporated for two reasons. First, Lemma 1 shows that F is not empty for any collection of L, U_C, and U_G. Hence, the complaint that linear models are often infeasible does not apply to this model. Second, the different lower bounds on the elastic functions allow us to embody different treatment aspirations.

Each of L, U_C, and U_G correspond with a vector, denoted by l, u_C, and u_G, that decides how discrepancies are penalized. For example, $L\alpha$ measures how deficient a plan is with regards to meeting the minimum tumor dose, and $l^T\alpha$ is the total penalty assigned to these discrepancies. Similarly, $U_C\beta$ and $U_G\gamma$ measure a plans deviation from CUB and GUB, and u_C^T and u_G^T are the aggregated penalties assigned to these deviations. The separation of how we measure and penalize deviation is convenient because it allows us to consider one set of constraints, decided by L, U_C, and U_G, and at the same time we can manipulate the objective function to address different situations. So we can design a patient-specific objective function that takes into account their ethical desires.

The objective functions that we consider comprise the three penalty functions $l^T\alpha$, $U_C\beta$ and $U_G\gamma$, and we consider variants of the following three optimization problems:

$$LP_\omega : \min\{\omega \cdot l^T\alpha + u_C^T\beta + u_G^T\gamma : (x,\alpha,\beta,\gamma) \in F\}$$

$$MOLP : \min\{(l^T\alpha,\, u_C^T\beta + u_G^T\gamma)^T : (x,\alpha,\beta,\gamma) \in F\} \text{ and}$$

$$MOLP'_{(T,C,G)} : \min\{(l^T\alpha,\, u_C^T\beta,\, u_G^T\gamma)^T : (x,\alpha,\beta,\gamma) \in F\}.$$

The first math program, LP_ω is a linear program that has three penalization functions aggregated into a single objective, with the weight ω deciding the importance of tumor uniformity. If ω is small, we are indicating that it is not important to satisfy the tumor's lower bound. As ω increases, we increase the emphasis on finding a plan that achieves a uniform, tumoricidal dose. The second math program is a multiple objective linear program, where the two objectives are 1) to attain a uniform, tumoricidal dose and 2) to minimize the radiation deposited into other structures. The third optimization problem is another multiple objective linear program, where the three objectives are to 1) minimize any deficiencies in the cancerous regions, 2) make sure that the critical structures receive as little radiation as possible, and 3) eliminate hot spots by minimizing the amount of radiation deposited into the remaining tissue.

We point out that each of the mathematical programs is capable of addressing different ethical situations. In LP_ω, we adjust the relative importance of the tumor receiving its desired amount of radiation by adjusting the value of ω. In Section 29.5.1 we show that the minimum amount of tumor deficiency is uniformly bounded by the inverse of ω, and we use this result to construct an ω that guarantees that the tumor receives a uniform, tumoricidal dose.

Because the other two mathematical programs have multiple objectives, we need to define the sense of optimization that we are going to use. For *MOLP* we are interested in the set of *pareto* optimal, or *efficient,* solutions. We say that (x,α,β,γ) is pareto optimal if there exists a θ strictly between 0 and 1 such that (x,α,β,γ) is optimal to

$$\min\{(1-\theta)l^T\alpha+\theta(u_C^T\beta+u_G^T\gamma):(x,\alpha,\beta,\gamma)\in F\} \tag{3}$$

Since θ is positive, we have that dividing the objective function by θ transforms this problem into LP_ω, where $\omega=(1-\theta)/\theta$. So, the set of pareto optimal solutions to *MOLP* is the same as the collection of all optimal solutions to LP_ω, where ω is positive. While this means that LP_ω and *MOLP* are variants of each other, in the results that follow we use LP_ω and *MOLP* differently. So, we consider these separate, but related problems.

We use *lexicographic* optimization, instead of pareto optimization, for the third optimization problem. For example, in $MOLP'_{(T,C,G)}$ we minimize $l^T\alpha$ first, and then minimize $u_C^T\beta$ over $\operatorname{argmin}\{l^T\alpha:(x,\alpha,\beta,\gamma)\in F\}$. Similarly, the third objective of minimizing $u_G^T\gamma$ is undertaken with only the solutions of the second problem. Because the three objectives are treated individually, we can easily alter their importance. In the case of a terminally ill patient, we may decide that attaining a uniform, tumoricidal dose is the least important objective, and hence we might order the objectives by 1) guaranteeing that the critical structures are underneath their bounds, 2) making sure that there are no unusually high depositions of radiation, and 3) attempting to deliver a uniform, tumoricidal dose. In such a case, we use the subscript *(C,G,T)* to indicate that the critical structures have the highest priority, that the normal (good) tissue has the second highest priority, and that the tumor has the lowest priority. So, the first objective is to minimize $u_C^T\beta$, the second objective is to minimize $u_G^T\gamma$, and the third objective is to minimize $l^T\alpha$. To make sure that the consequences of lexicographic optimization are understood, suppose for $MOLP_{(T,C,G)}$ that there are treatment plans that achieve a uniform, tumoricidal dose. This means that the set of optimal solutions found by minimizing $l^T\alpha$ are exactly those plans that achieve a uniform, tumoricidal dose. When the second objective is minimized, we are only going to consider those treatment plans that achieve a uniform, tumoricidal dose, and it is possible that none of these plans adhere to the bounds placed on the critical structures. However, there may be a plan that delivers sufficiently low levels to the critical structures and has only the slightest tumor deficiency, but we would not find such a plan because it would not be optimal to the first problem. This is the nature of lexicographic optimization, and this type of optimization is appropriate only if a hierarchy of the objectives is clear.

Different elastic functions lead to different interpretations of the solution, and the following two collections are of particular interest.

Average Analysis

$$l = \frac{1}{m_T}e \quad u_C = \frac{1}{m_c}e \quad u_G = \frac{1}{m_{Gc}}e \quad L{=}I \quad u_C = I \quad u_G = I$$

Absolute Analysis

$$l = 1 \quad u_C = 1 \quad u_G = 1 \quad L{=}e \quad u_C = e \quad u_G = e$$

Suppose that average analysis is chosen. Then $(L\alpha)_p = \alpha_p$ tells us how deficient a plan is with regards to meeting the minimum tumor dose for pixel *p*, and $L^T\alpha = (1/m_T)e^T\alpha$ is the average amount of such deficiencies. The interpretation of $U_C\beta = \beta$ depends on the sign of the component. If $(U_C\beta)_p = \beta_p > 0$, pixel p, which is contained in some critical structure, is receiving more radiation than the physician intended. However, if $\beta_p < 0$, pixel p is receiving less radiation than is allowed. We now see that the objective term $U_C^T\beta = (1/m_c)e^T\beta$ expresses the desire to decrease the average dose to the critical structures; in fact the desire is to have the critical structures receive no radiation. Similarly, $(U_G\gamma)_p = \gamma_p > 0$ indicates how much pixel p is over its allotted upper bound, and $U_G^T\gamma = (1/m_G)e^T\gamma$ is the average amount of radiation the normal tissue is over its prescribed dose. The roles of β and γ differ because of the different lower bounds. Since $0 \leq \gamma$, any plan satisfying $A_Gx \leq GUB$ contributes zero to the objective function. However, the lower bound of $-CUB$ on β means that plans with a low integral dose to the critical structures are preferred. So, for the average analysis case we see that the objective function is three tiered in its goals:

- minimize the average amount that the tumor is under its prescribed dose,
- minimize the average amount of radiation that the critical structures receive, and
- minimize the average amount that the remaining pixels are over their upper bounds.

The interpretation is similar if absolute analysis is chosen, with the difference being that the elastic functions are each controlled by a single parameter. So, instead of minimizing an average discrepancy, the goal is to minimize the maximum amount of discrepancy. Hence, when absolute analysis is chosen, the three goals of the objective function are to

- minimize the maximum amount that the tumor is under its prescribed dose,
- minimize the maximum amount of radiation that the critical structures receive, and
- minimize the maximum amount any remaining pixel is over its upper bound.

The literature contains several models, ranging from linear to mixed integer to nonlinear models. While our concentration is on the linear models developed above, it is important for operations researchers working in the field to have an awareness of these other models. We direct researchers to the works listed in Table 29.1.

Table 29.1 Research papers that investigate and review optimization models that are used to aid IMRT design

Citations	Models Investigated
Rosen et al. [21], Shepard et al. [29]	reviews linear and nonlinear models
Langer [16], Langer et al. [17]	mixed integer models
Morrill et al. [23], Raphael [7]	probabilistic models

29.5 MATHEMATICAL AND COMPUTATIONAL RESULTS

As mentioned in the introduction, the medical literature associated with IMRT design and optimization is immense. The history of the medical research is to design a specific treatment plan for a specific type of cancer, and then show that it is appropriate through several examples. So in the medical literature, the methodology of treatment is verified through examples – i.e. showing that a technique works because it has favorable properties on these examples. While this verification approach is important, and indeed the great strides we have made in medicine are a direct result of such work, this type of research is foreign to a mathematician. The field of mathematics is concerned with statements that can be universally proved and not ones that can simply be shown to hold for a few examples. This means that an applied mathematician's perspective of a problem is different from the perspective held by a practitioner. An applied mathematician, such as an operations researcher, approaches a problem by finding the mathematical language needed to describe the essence of the problem and then proceeds to prove statements about the situation at hand. The proofs provide a theoretical certificate for what can and cannot be stated about the problem. So, instead of stating that a technique works because we can show that it does on a few examples, we can guarantee that a technique does or does not work because the proofs establish that they will, or will not. The benefit of this theoretical approach is that it is not example dependent, so we do not have to wonder if there are examples where the technique fails.

In this section we are interested in developing theoretical statements about the linear models presented in Section 29.4. Unfortunately, mathematical rigor is scarce in the literature. Consequently, this field of research has not benefited from a sound mathematical development. This author hopes that the operations researchers working in this field feel a sense of responsibility to provide the needed theoretical basis (areas in telecommunications and geosciences have benefited greatly from a similar mathematical foundation). The results of this section are divided into three subsections, each related to one of the three models presented in Section 29.4. The proofs of the mathematical results are excluded for brevity, but a citation is provided where a proof can be located.

The numerical results in each of the following sections rely on the *optimal partition* of a linear program. Consider the standard form linear program, $\min\{c^T x : Ax = b, x \geq 0\}$. Allowing P^* to be the optimality set, we have that the optimal partition $(B \mid N)$ is defined by

$$N = \{i : x_{i=0} \text{ for all } x \in P^*\} \quad \text{and} \quad B = \{1, 2, \cdots, n\} \setminus N.$$

The reason that the optimal partition is important is that it provides an algebraic characterization of the optimal set. We do not have the space in this chapter to rigorously develop this representation, but it is well known that $P^* = \{x : Ax = b, x \geq 0, x_i = 0, i \in N\}$ [30].

The optimal partition was not easily computed until path-following interior point algorithms became viable alternatives to the simplex method. Path-following interior point algorithms terminate with a solution that induces the optimal partition, meaning that if x^* is an optimal solution found by such an algorithm, then $B = \{i : x_i^* > 0\}$ and $N = \{i : x_i^* = 0\}$. Having this type of solution is important because it strictly satisfies as many inequalities as possible: the B set indexes the *entire* collection of inequalities that can be strictly satisfied by an optimal solution. So for IMRT design, the plan found by a path-following interior point algorithm strictly satisfies as much of the prescription as possible.

A path-following interior point algorithm is appropriate only if the strict interiors of the primal and dual feasibility sets are non-empty. This means that there must be an $(x, \alpha, \beta, \gamma)$ that *strictly* satisfies the inequalities in (2). Fortunately, Lemma 1 states that we are guaranteed that a path-following interior point algorithm is applicable.

Lemma 1 [13] *We have for any collection of elastic functions that the primal and dual strict interiors of* LP_ω*, MOLP, and MOLP' are non-empty.*

29.5.1 Analysis certificates for LP_ω

The objective function of LP_ω is a weighted sum of three goals. While a common criticism of such objective functions is that the weights are difficult to understand, we show that choosing ω appropriately provides a meaningful interpretation. The positive scalar ω weights the importance of a plan achieving the minimum tumor dose – i.e. large values of ω encourage $l^T\alpha$ to be as small as possible. We would like to have the property that there exists a finite $\omega > 0$ such that the optimal value of $l^T\alpha$ is zero. This follows because the tumor is then guaranteed to receive its minimum radiation level. Such an ω would serve as a certificate of a tumoricidal dose. The bad news is that there are simple examples where the optimal value of $l^T\alpha$ is not zero for all $\omega > 0$. However, the good news is that we can calculate an ω that certifies that the discrepancy between the amount delivered to the tumor and the tumor's lower bound is sufficiently small.

We say that a prescription *allows tumor uniformity* if there is a treatment plan *x* such that $TLB \le A_T x \le TUB$. Moreover, a prescription is *attainable* if there is an (x,α,β,γ) in *F* such that $l^T\alpha = 0$, $U_C\beta \le 0$ and $U_G\gamma = 0$. Obviously every attainable prescription allows tumor uniformity, but not every prescription that allows tumor uniformity is attainable. Theorem 1 shows that if the prescription allows tumor uniformity, then the tumor deficiency is uniformly bounded above by the inverse of ω. In what follows, we let $\underline{rs}(M)$ be the minimum row sum of the matrix *M,* and we use the standard big-O order notation – i.e. *f(x) = O(g(x))* if, for the nonnegative functions *f* and *g*, there exists a positive constant κ, such that $f(x) \le \kappa g(x)$. Also, we use the standard notations for the 1-norm, $\|x\|_1 = \sum_i^n |x_i|$ and the infinity-norm $\|x\|_1 = \max\{|x_i| : i = 1,2,\cdots,n\}$.

Theorem 1 [12] *Let* $(x^*(\omega),\alpha^*(\omega),\beta^*(\omega),\gamma^*(\omega))$ *be an optimal solution to* LP_ω. *For any collection of elastic functions we have that* $l^T\alpha^*(\omega) = O(\frac{1}{\omega})$, *provided that prescription allows tumor uniformity.*

From Theorem 1 there is positive scalar κ such that $l^T\alpha(\omega) = \frac{\kappa}{\omega}$, which is useful because an upper bound on κ is easily found if either average or absolute analysis is used. In particular, setting

$$u = \frac{\|TUB\|_\infty}{\min\{A_{(p,a,j)} : A_{(p,a,j)} \neq 0, p\, is\ a\ tumor\ pixel\}},$$

we have from [12] that κ is no greater than

$$\frac{\|A_C(ue) - CUB\|_\infty \|u_C^T\|_1}{\underline{rs}(U_C)} + \frac{\|A_G(ue) - GUB\|_\infty \|u_G^T\|_1}{\underline{rs}(U_G)} + u_C^T CUB$$

if average analysis is used, and no greater than

$$\frac{\|A_C(ue) - CUB\|_\infty \|u_C^T\|_1}{\underline{rs}(U_C)} + \frac{\|A_G(ue) - GUB\|_\infty \|u_G^T\|_1}{\underline{rs}(U_G)} + \frac{u_C^T U_C^T CUB}{m_C}$$

if absolute analysis is used. We let κ' be the greater of these two bounds, so that regardless of the type of analysis we have $l^T\alpha^*(\omega) \le \kappa'/\omega$.

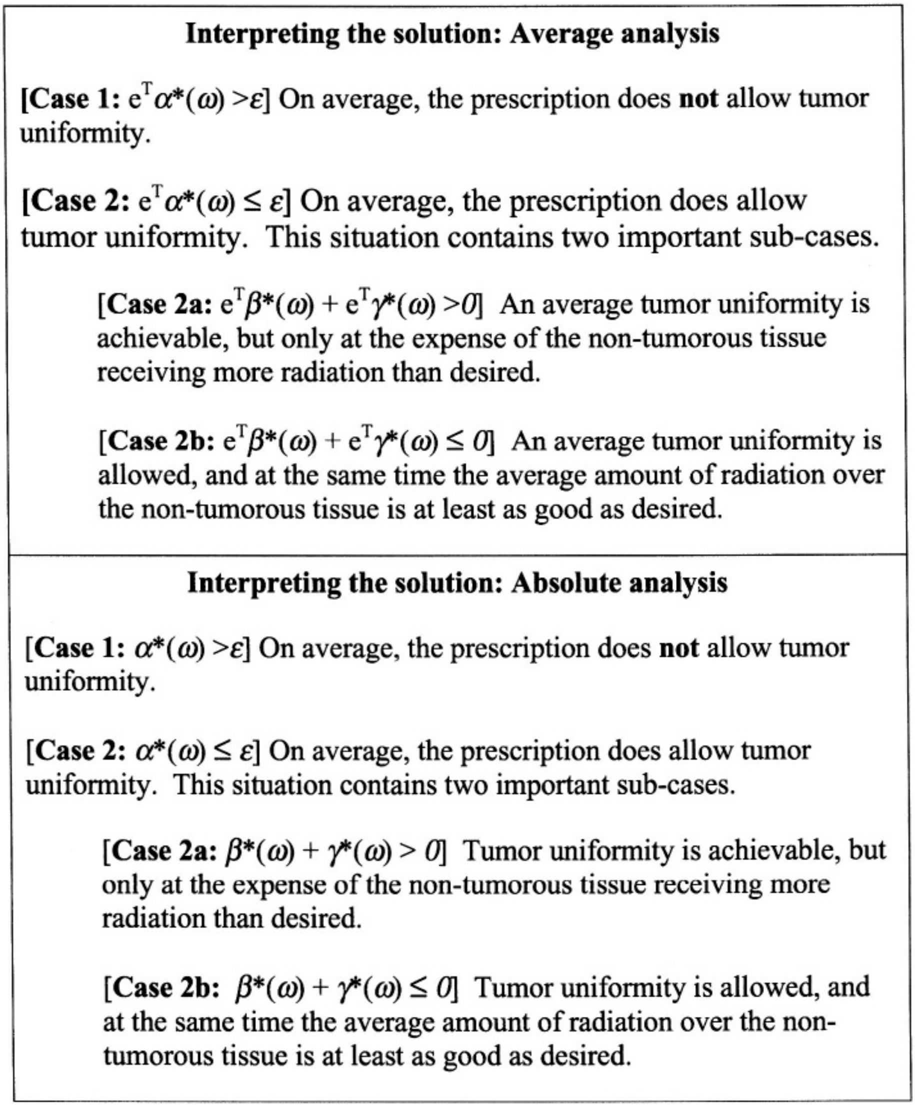
Recall that *TG* was the goal dose for the tumorous region and that we originally set *TLB*=(1-*tol*)*TG e*. To utilize the upper bound provided by κ', we slightly increase each component of *TLB* by $\varepsilon > 0$ – i.e. we instead let $TLB = (1\text{-}tol)TG\ e + \varepsilon e$. After calculating κ, we choose $\omega = \kappa'/\varepsilon$ and solve LP_ω. Theorem 1 now implies that the optimal value of $l^T\alpha$ is less than ε, and hence the sought after uniformity is guaranteed. So using only the optimal objective value, we have from Theorem 1 the analysis found in Figure 29.5. Of course a more detailed interpretation of the solution is possible by examining the individual components of $(\alpha^*(\omega), \beta^*(\omega), \gamma^*(\omega))$.

A prototype treatment system called *R*adiotherapy optim*A*l *D*esign, or *RAD,* has been developed using MATLAB©. This system is available at http://www.trinity.edu/aholder/research/oncology/. *RAD* uses a 64×64 grid, and allows angles evenly spaced at every 15, 5, or 1 degree(s), with each beam comprising 10, 32, or 32 pencils, respectively. In addition to allowing the user to choose from different angle geometries, *RAD* has the following features.

- Either absolute or average analysis can be used.

- A prescription window allows the user to easily set the tissue type, the prescription levels, and the tumor uniformity level.

Figure 29.5 Interpreting the solution of average and absolute analysis



Interpreting the solution: Average analysis

[Case 1: $e^T\alpha^*(\omega) > \varepsilon$] On average, the prescription does **not** allow tumor uniformity.

[Case 2: $e^T\alpha^*(\omega) \leq \varepsilon$] On average, the prescription does allow tumor uniformity. This situation contains two important sub-cases.

> **[Case 2a:** $e^T\beta^*(\omega) + e^T\gamma^*(\omega) > 0$] An average tumor uniformity is achievable, but only at the expense of the non-tumorous tissue receiving more radiation than desired.
>
> **[Case 2b:** $e^T\beta^*(\omega) + e^T\gamma^*(\omega) \leq 0$] An average tumor uniformity is allowed, and at the same time the average amount of radiation over the non-tumorous tissue is at least as good as desired.

Interpreting the solution: Absolute analysis

[Case 1: $\alpha^*(\omega) > \varepsilon$] On average, the prescription does **not** allow tumor uniformity.

[Case 2: $\alpha^*(\omega) \leq \varepsilon$] On average, the prescription does allow tumor uniformity. This situation contains two important sub-cases.

> **[Case 2a:** $\beta^*(\omega) + \gamma^*(\omega) > 0$] Tumor uniformity is achievable, but only at the expense of the non-tumorous tissue receiving more radiation than desired.
>
> **[Case 2b:** $\beta^*(\omega) + \gamma^*(\omega) \leq 0$] Tumor uniformity is allowed, and at the same time the average amount of radiation over the non-tumorous tissue is at least as good as desired.

- A simplex-based solver is available.

- After the optimization routine is complete, three figures are presented. The first and second figures are a contour plot and a 3-D image of the radiation levels delivered by the plan. The third figure provides an explanation of the solution that depends on whether absolute or average analysis was chosen.

In the examples that follow there are 360 equally spaced beams, and each beam contains 32 sub-beams. The amount by which *TLB* is increased is internally set to 10^{-4}, $TLB = (1 - tol)\ TG + 10^{-4}\ e$. The problems were solved on a 1.5 GHz PC with 1G of RAM.

The first example is found in Figures 29.6 and 29.7. In this example a tumor has grown half-way around a critical structure. The tumoricidal dose was 80Gy, and the critical structure was restricted to no more than 30Gy. The tumor uniformity level was 2%, and an absolute analysis was used. The value of $l^T\alpha^*$ was less than 10^{-4}, from which we conclude that the prescription allows tumor uniformity. Indeed, the maximum and minimum doses were 78.42Gy and 81.57Gy, which are within the 80Gy±2%. Not only does this plan strictly satisfy the tumor uniformity bounds, but it also does not deliver any radiation to the critical structure.

So, we have designed a plan that delivers a uniform, tumoricidal dose to the tumor and does not deposit any radiation in the critical structure.

The example in Figure 29.8 is significantly more complicated because the tumor is nearly surrounded by critical structures. The tumoricidal dose is 78Gy, with a uniformity level of 4%. The plan depicted in Figures 29.9 and 29.10 attained the tumor uniformity with the minimum and maximum doses inside the cancerous region being 75.17Gy and 81.1Gy. However, there is no plan that achieves a uniform, tumoricidal dose that does not violate the bounds placed on the critical structure. We know this because the value of $l^T\alpha^*$, which is simply α^* for the absolute analysis, is greater than 10^{-4}, and the value of $\mu_C^T\beta^*$, which is simply β^8, is 6.62. So, we know that some part of a critical structure must receive 6.62Gy over its prescribed bound to attain a uniform, tumoricidal dose. Depending on the type of critical structure, this may or may not be acceptable, and if not, the planners need to reconsider their desires for the tumor.

29.5.2 *Evaluating an angle's value with MOLP*

We see from the examples in Section 29.5.1 that the plans developed by a path-following interior point algorithm tend to design plans that use many angles. In fact, these plans use so many angles that they are not practical – i.e. the time that it would take to deliver such a plan is well beyond the 15 minutes of a typical treatment. The problem here is that a path-following interior point algorithm terminates with a solution that strictly satisfies as many inequalities as possible, and as already mentioned, this is favorable because we design a plan that strictly satisfies as much of the prescription as possible. However, this is bad because we also design a plan that uses as

Figure 29.6 A contour plot showing how the deposition pattern 'bends' around the critical structure

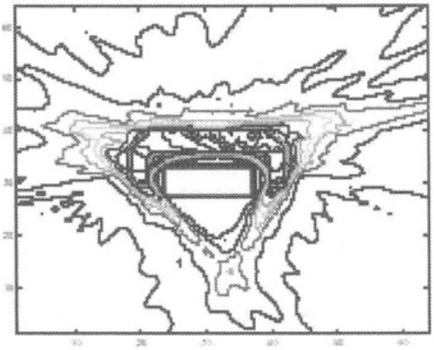
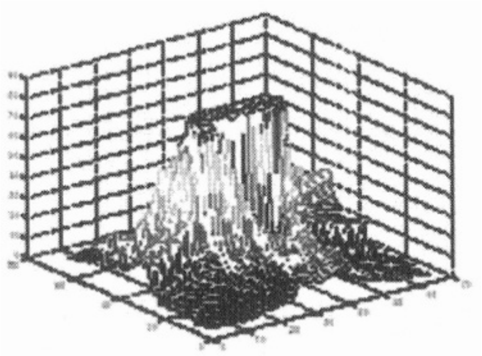


Figure 29.7 The vertical height is the amount of radiation delivered by the plan over the image



many angles and sub-beams as possible. What is needed is a technique to prune the collection of possible angles so that the 'best' angles remain in the pruned collection. This and the following section develop ways to measure the importance of an angle when the priorities of the treatment are uncertain.

The set of pareto optimal solutions of ***MOLP,*** called the *efficient frontier,* induces an optimal partition for the multiple objective program that is similar to the linear programming optimal partition [13].

Figure 29.8 A difficult geometry to plan because the tumor is almost surrounded by low-dose critical structures.

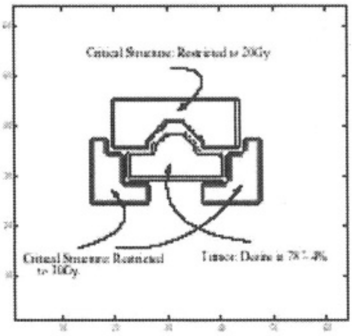

Critical Structure: Restricted to 20Gy
Critical Structure: Restricted to 10Gy
Tumor: Desire is 78 ± 4%

Figure 29.9 The vertical height is the amount of radiation delivered by the plan over the image

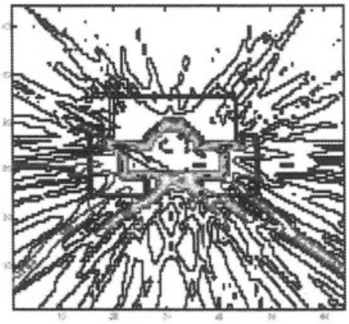
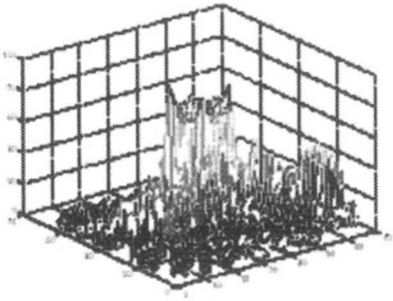


Figure 29.10 The amount of radiation over the image. The amount over the tumor is fairly uniform, but there is a spike between the tumor and one of the critical structures



Definition 1 *Let be the efficient frontier of MOLP. The MOLP optimal partition, denoted* $(\overset{MOLP}{B} \mid \overset{MOLP}{N})$ *is defined by*

$$\overset{MOLP}{N} = \{i : x_i = 0 \quad \mathit{for} \quad \mathit{all} \quad x \in \xi\} \ \mathit{and}$$

$$\overset{MOLP}{B} = \{1,2,3,\cdots,n\} \setminus N.$$

The definition of the MOLP optimal partition retains the quality that an index being in *N* indicates that the component is zero in every pareto optimal solution. Likewise, an index in *B* demonstrates that the component is allowed to be positive on the efficient frontier. A property that is unfortunately lost is that the MOLP optimal partition is not capable of characterizing the efficient frontier, i.e.

$$\xi \neq \left\{(x,\alpha,\beta,\gamma) \in F : (x,\alpha,\beta,\gamma)_i = 0, i \in \overset{MOLP}{N}\right\}$$

However, we do have

$$\xi \subseteq \left\{(x,\alpha,\beta,\gamma) \in F : (x,\alpha,\beta,\gamma)_i = 0, i \in \overset{MOLP}{N}\right\}$$

An algorithm to compute the MOLP optimal partition is found in [13]. This algorithm uses the parameterization in (3) by finding the linear programming optimal partition for every value of θ between 0 and 1. The set $\overset{MOLP}{B}$ is the union of all the linear programming *B* sets, and $\overset{MOLP}{N}$ is $\{1,2,3,\cdots n\} \setminus \overset{MOLP}{B}$. So, $\overset{MOLP}{N}$ indexes the sub-beams that are not used for any $\theta \in (0,1)$. Recall that the value of θ is a measure of how important it is to deliver a uniform, tumoricidal dose, with values near 0 and 1 giving the cancerous regions a high and low importance, respectively. This means that $\overset{MOLP}{N}$ contains the sub-beams that are not used in any weighting of the objectives, and hence, these sub-beams should never be used in a treatment plan. The other side of this is that $\overset{MOLP}{B}$ indexes the sub-beams that are used for some collection of weights, and hence, there is at least one circumstance where each of these sub-beams is used.

Because of the way the algorithm works, we actually acquire more information than just described. Consider the situation depicted in Figure 29.11, where a tumor is surrounded by three critical structures. This experiment was run with 72 equally spaced angles, each with 32 sub-beams. The tumor uniformity level was $\pm 4\%$ and an average analysis was used. Figure 29.12 and 29.13 provide information about which angles are used, and not used, as θ traverses the interval (0,1). We saved the optimal partition for each θ and used this information to calculate how often an angle is used. We say that an angle is *used at Level k* if there are *k* sub-beams from that angle with positive amounts of radiation in the optimal plan. Furthermore, we say that an angle is *on* provided its level of use is at least 1. In Figure 29.12 we calculated each angle's level of use, and then added these together for each of the 399 different optimal partitions. These values are recorded above the circle around the image. The highest peak is at 90° and has a value of 1,424, which means that 1,424 sub-beams were used from this angle as θ traversed the interval (0,1). Figure 29.13 is similar, but instead of accumulating sub-beams from each angle, the percentage of times an angle is on is displayed over the circle. Angle 85° was on in 100% of the optimal partitions, and while 90° had the highest amount of sub-beam usage, it was not on in each optimal partition (it was used in 99.75% of the optimal partitions).

The point of Figures 29.12 and 29.13 is that a dosimetrist can easily decide which angles are, and are not, important. The most definitive information lies in the angles that are never used, as there is no situation where these angles are in an optimal plan. Similarly, any angle whose use is 100%, which is only 85° for this example, is used in every situation. The graphs provide a measure of an angle's usefulness in other situations. For example, if a dosimetrist wants a three-beam plan, then he or she might decide to use angles 85°, 40°, and 205°, all of which have higher peaks in Figure 29.13.

29.5.3 *Using MOLP' to reduce the number of angles*

Recall that we use lexicographic optimization for *MOLP'* and that the order in which the objectives are considered is indicated by the subscript. Lexicographic optimization has its own optimal partition, which is easily calculated for our problem [13]. As an example, consider $MOLP'_{(C,T,G)}$, where we calculate the lexicographic optimal partition as follows.

Step 1 Solve $\min\{u_c^T\beta : (x,\alpha,\beta,\gamma) \in F\}$ and let $(B^1 \mid N^1)$ be the optimal partition.

Figure 29.11 A tumor surrounded by three critical structures

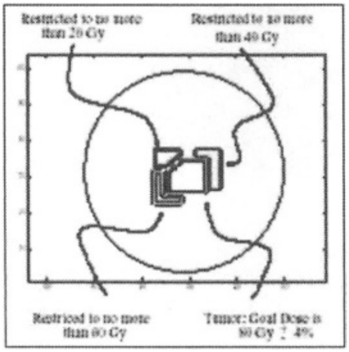

Restricted to no more than 20 Gy
Restricted to no more than 40 Gy
Restricted to no more than 60 Gy
Tumor: Goal Dose is 80 Gy ± 4%

Figure 29.12 Accumulative totals of sub-beam usage along each angle

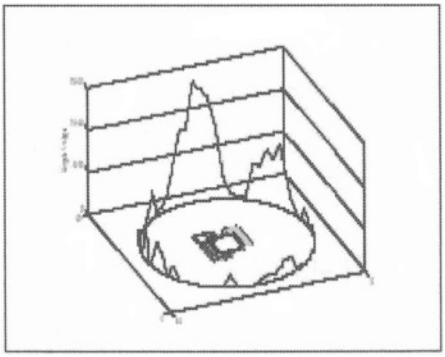
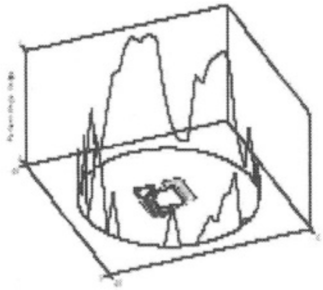


Figure 29.13 Percent totals for each angle



Step 2 Solve $\min\{l^T\alpha : (x,\alpha,\beta,\gamma) \in F, (x,\alpha,\beta,\gamma)_i = 0, i \in N^1\}$ and let $(B^2 \mid N^2)$ be the optimal partition.

Step 3 Solve $\min\{u_G^T\alpha : (x,\alpha,\beta,\gamma)\in F, (x,\alpha,\beta,\gamma)_i = 0, i\in N^1\}$ and let $(B_L \mid N_L)$ be the optimal partition.

The last partition, $(B_L \mid N_L)$, is called the *lexicographic optimalpartition.* If we rearranged the priorities, to say *(G,C,T),* the only difference would be that the objective function in step 1 would become $u_C^T\beta$, in step 2 it would become $u_C^T\gamma$, and in step 3 it would become $l^T\alpha$.

Similar to the *MOLP* optimal partition, the lexicographic optimal partition provides an insight into the usefulness of an angle. As an example, consider the problem in Figure 29.14, where a tumor has gown around a critical structure. There were 72 equally spaced angles, each containing 32 subbeams. The tumor uniformity level was $\pm 4\%$, and an average analysis was used, We calculated the lexicographic optimal partition for each of the 6 possible orderings of the objectives, and Figures 29.15 through 29.20 show each angles level of use in me corresponding optimal plan. Again, the most definitive information comes from the angles that have a zero level of use, for these angles are not to be considered for that priority list. As an example, angle 170° has a zero level of use for the priority list *(T,C,G),* but has a level of use of two for the priority list *(T,G,C).* There were nine angles whose level of use was zero in all six priority lists: 55°, 75°, 120°, 125°, 205°, 225°, 250°, 310°, and 355°. These are the angles that can be excluded from consideration regardless of how the dosimetrist orders the priorities (which means we have removed 12.5% of the angles from consideration).

29.6 CONCLUSION

We conclude this chapter with a plea to the operations research community. Operations research has successfully been used in many disciplines, but one of the few areas that has not witnessed the benefits of the field is clinical medicine. Of course, the statistical training that most operations researchers posses is useful in drug trials and other data intensive medical applications. However, many of the medical procedures that are used in practice have not been mathematically modeled and optimized, which means that any improvement in the treatment is found by trial and error. The area of IMRT design is starting to benefit from the optimization process, and we are now at a point where the operations researchers can make a substantial improvement in a patient's treatment. The author urges those who work in the field of operations research to consider working on a problem that involves some clinical treatment, for all of humankind benefits from this work.

Figure 29.14 A tumor that has grown around a critical structure

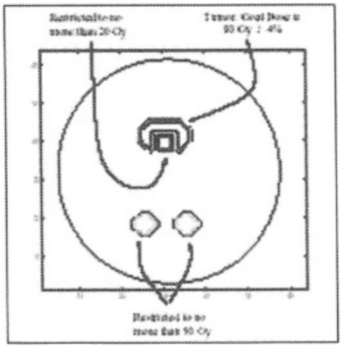

Restricted to no more than 20 Gy
Tumor: Goal Dose ≥ 90 Gy ± 4%
Restricted to no more than 90 Gy

Figure 29.15 Each angle's use for priority list *(T,C,G)*

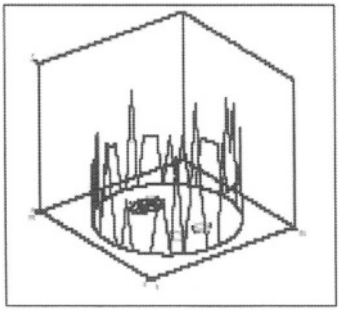


Figure 29.16 Each angle's use for priority list (*T,G,C*)

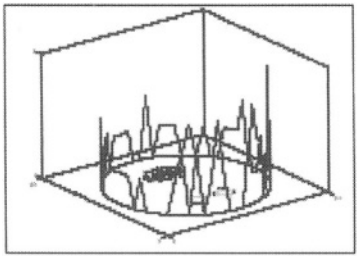


Figure 29.17 Each angle's use for priority list *(G,T,C)*

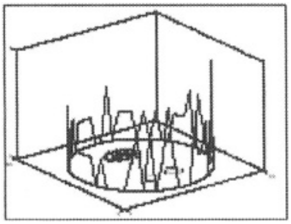


Figure 29.18 Each angle's use for priority list *(G,C,T)*

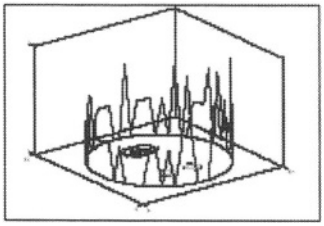


Figure 29.19 Each angle's use for priority list *(C,G,T)*

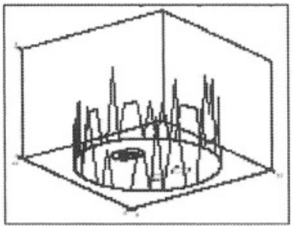
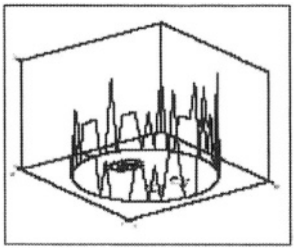


Figure 29.20 Each angle's use for priority list (*C,T,G*)



References

[1] Censor, Y. (1991). Mathematical aspects of radiation therapy treatment planning: Continuous inversion versus full discretization and optimization versus feasibility. In Borgers, C. and F. Natterer, Eds., *Computational Radiology and Imaging: Therapy and Diagnostic.* Springer-Verlag, New York, NY, 101-112.

[2] Cormack, A. and E. Quinto (1990). The mathematics and physics of radiation dose planning using x-rays. *Contemporary Mathematics,* 113, 41-55.

[3] Bartolozzi, F., et al. (2000). Operational research techniques in medical treatment and diagnosis. A review. *European Journal of Operations Research,* 121, 435-466.

[4] Censor, Y., M. Altschuler, and W. Powlis (1988). A computational solution of the inverse problem in radiation-therapy treatment planning. *Applied Mathematics and Computation,* 25, 57-87.

[5] Powlis, W., M. Altschuler, Y. Censor, and E. Buhle (1989). Semi-automatic radiotherapy treatment planning with a mathematical model to satisfy treatment goals. *International Journal of Radiation Oncology, Biology, Physics,* 16, 271-276.

[6] Goitein, M. and A. Niemierko (1988). Biologically based models for scoring treatment plans. Scandinavian Symposium on Future Directions of Computer-Aided Radiotherapy.

[7] Raphael, C. (1992). Mathematical modeling of objectives in radiation therapy treatment planning. *Physics in Medicine and Biology,* 37, 1293-1311.

[8] Withers, H., J. Taylor, and B. Maciejewski (1987). Treatment volume and tissue tolerance. *International Journal of Radiation Oncology, Biology, Physics,* 14, 751-759.

[9] Wolbarst, A. (1984). Optimization of radiation therapy II: The critical-voxel model. *International Journal of Radiation Oncology, Biology, Physics,* 10, 741-745.

[10] Ferris, M. and M. Voellker (2002). Neuro-dynamic programming for radiation treatment planning. Technical Report NA-02/06, Numerical Analysis Group, Computing Laboratory, Oxford University.

[11] Boland, N., H. Hamacher, and F. Lenzen (2002). Minimizing beam-on time in cancer radiation treatment using multileaf collimators. Technical Report KLUEDO: 2002-02-10, Universittsbibliothek Kaiserslautern.

[12] Holder, A. (2003). Designing radiotherapy plans with elastic constraints and interior point methods. *Health Care Management Science,* 6, 5-16.

[13] Holder, A. (2001). Partitioning multiple objective solutions with applications in radiotherapy design. Technical Report 54, Department of Mathematics, Trinity University, San Antonio, TX.

[14] Bahr, G.K., J.G. Kereiakes, H. Horwitz, R. Finney, J. Galvin, and K. Goode (1968). The method of linear programming applied to radiation treatment planning. *Radiology,* 91, 686-693.

[15] Hodes, L. (1974). Semiautomatic optimization of external beam radiation treatment planning. *Radiology,* 110, 191-196.

[16] Langer, M. (1987). Optimization of beam weights under dose-volume restrictions. *International Journal of Radiation Oncology, Biology, Physics,* 13, 1255-1260.

[17] Langer, M., R. Brown, M. Urie, J. Leong, M. Stracher, and J. Shapiro (1990). Large scale optimization of beam weights under dose-volume restrictions. *International Journal of Radiation Oncology, Biology, Physics,* 18, 887-893.

[18] Legras, J., B. Legras, and J. Lambert (1982). Software for linear and non-linear optimization in external radiotherapy. *Computer Programs in Biomedicine,* 15, 233-242.

[19] Lodwick, W., S. McCourt, F. Newman, and S. Humphries (1998). Optimization methods for radiation therapy plans. In Borgers, C. and F. Natterer, Eds., *Computational, Radiology and Imaging: Therapy and Diagnosis.* Springer-Verlag, New York, NY.

[20] Morrill, S., I. Rosen, R. Lane, and J. Belli (1990). The influence of dose constraint point placement on optimized radiation therapy treatment planning. *International Journal of Radiation Oncology, Biology, Physics,* 19, 129-141.

[21] Rosen, I., R. Lane, S. Morrill, and J. Belli (1991). Treatment plan optimization using linear programming. *Medical Physics,* 18, 141-152.

[22] Sonderman, D. and P. Abrahamson (1985). Radiotherapy treatment design using mathematical programming models. *Operations Research,* 33, 705-725,

[23] Morrill, S., R. Lane, G. Jacobson, and I. Rosen (1991).Treatment planning optimization using constrained simulated annealing. *Physics in Medicine and Biology,* 36, 1341-1361.

[24] Chinneck, J. (1995). An elective polynomial-time heuristic for the minimum-cardinality set-covering problem. *Annals of Mathematics and Artificial Intelligence,* 17, 127-144.

[25] Chinneck, J. (1997). Finding a useful subset of constraints for analysis in an infeasible linear program. *INFORMS Journal on Computing,* 9, 164-174.

[26] Greenberg, H. (1993). How to analyze results of linear programs, part 3:Infeasibility diagnosis. *Interfaces,* 23, 120-139.

[27] Greenberg, H. (1996). Consistency, redundancy and implied equalities in linear systems. *Annals of Mathematics and Artificial Intelligence,* 17, 37-83.

[28] Berman, A. and R. Plemmons (1979). *Nonnegative Matrices in the Mathematical Sciences.* Academic Press, New York, NY.

[29] Shepard, D., M. Ferris, G. Olivera, and T. Mackie (1999). Optimizing the delivery of radiation therapy to cancer patients. *SIAM Review,* 41, 721-744.

[30] Roos, C., T. Terlaky, and J.-P. Vial (1997). *Theory and Algorithms for Linear Optimization: An Interior Point Approach.* John Wiley and Sons, New York, NY.

30 OPTIMIZATION TOOLS FOR RADIATION TREATMENT PLANNING IN MATLAB

Michael C. Ferris[1], Jinho Lim[2] and David M. Shepard[3]

[1] Computer Sciences Department
University of Wisconsin
Madison, WI 53706

[2] Department of Industrial Engineering
University of Houston
Houston, TX 77204

[3] Department of Radiation Oncology
University of Maryland School of Medicine
Baltimore, MD 21201

SUMMARY

This chapter describes a suite of optimization tools for radiation treatment planning within the Matlab programming environment. The data included with these tools was computed for real patient cases using a Monte Carlo dose engine. The formulation of a series of optimization models is described that utilizes this data within a modeling system. Furthermore, visualization techniques are provided that assist in validating the quality of each solution. The versatility and utility of the tools are shown using a sequence of optimization techniques designed to generate a practical solution. These tools and the associated data are available for download from www.cs.wisc.edu/~ferris/3dcrt.

KEY WORDS

Optimization, Radiation treatment planning, Matlab

30.1 INTRODUCTION

The optimization of radiation treatment for cancer has become an active research topic in recent years [1-9]. Many types of cancer are treated by applying radiation from external sources, firing beams into a patient from a number of different angles in such a way that the targeted tumor lies at the intersection of these beams. The increasing sophistication of treatment devices – the aperture through which the beams pass can take on a variety of shapes, multiples apertures can be delivered for each beam angle, and wedges can be used to vary the radiation intensity across the beam – allows delivery of complex and sophisticated treatment plans, achieving a specified dose to the target area while sparing surrounding tissue and nearby critical structures. Optimization techniques are proving to be useful in the design of such plans.

This chapter describes a selection of tools that allow optimization approaches to be applied, visualized and iteratively refined. The tools use the Matlab programming environment for overall control and visualization, and the GAMS modeling language for formulation and solution of the underlying optimization models. One of the strengths of this work is that we utilize data that corresponds to real patient cases and thus include a variety of inhomogeneities due to different tissue types. Several useful tools for visualization of the results, along with a sophisticated use of an existing interface between the Matlab programming environment and the GAMS modeling language, are explained via example. We also show how a succession of optimization problem solutions can be used to satisfy the constraints of a realistic plan, for example dose-volume histogram constraints and the location-specific control of hot and cold spots.

The tools described in this chapter can be used for creating radiation therapy treatment plans delivered using either of two treatment techniques: (1) three-dimensional conformal radiotherapy (3DCRT) or (2) intensity modulated radiation therapy (IMRT).

3DCRT is the most common delivery technique used in radiation therapy. With this approach, each beam is shaped to match the view of the tumor from the given direction. In addition, one can choose to include a wedge filter in the beam, which results in a linear variation in intensity across the beam. This is particularly useful for treating tumors near the patient's surface and for compensating for the curved surface of the patient.

IMRT is a more advanced delivery technique that significantly increases the complexity of radiation delivery but provides an improved ability to conform the radiation to the tumor volume. In IMRT, a nonuniform radiation intensity is delivered from each beam angle. The dose delivered to the tumor volume from each beam angle is typically highly nonuniform, but the total dose delivered to the tumor from all angles provides adequate dose uniformity in the tumor with a rapid falloff in dose in the normal tissue.

While the data and techniques used to solve both types of models have much in common, we restrict the discussion in this chapter to 3DCRT for clarity and ease of exposition. Section 30.2 describes the data generation for the treatment planning problem. We believe that there is a fundamental core of optimization models that are useful for treatment planning. Several of these optimization models are discussed in Section 30.3, and the use of the environment in a simple example is given in Section 30.4. Matlab routines are presented to examine the solution quality in Section 30.5. All of these routines are available from www.cs.wisc.edu/~ferris/3dcrt.

We understand that these models should be used in combination and in an iterative fashion to achieve the goals of the planner. Thus, although the core optimization model remains the same, the data that describes a particular instantiation of the model can change in an iterative way as the optimization proceeds. We believe that this is where our environment will be most useful. In Section 30.6, we outline two more complex examples of its use, showing how to incorporate sampling techniques and multiple dose volume histogram (DVH) constraints on a single sensitive structure. While the suite of tools we have developed here are already useful for treatment design and refinement, we believe mat their strength is their easy extensibility to provide the basis for significant treatment improvements over the coming years.

30.2 PROBLEM DATA

Our planning tool is designed within the Matlab programming environment. The problem data consists of two broad components, namely a set of structures (organs) to which a certain level of radiation must be delivered, and a set of beamlets for delivering this radiation. The amount of data is large and patient/case specific. We have designed our environment to allow the use of actual patient data as well as simulated data. We have a growing collection of examples of such problems. We have chosen to store these examples as a “gdx” file [29], a “Gams-Data-eXchange” format that allows the data to be accessed very quickly within a GAMS optimization model, and also (via an API) by other programs.

For patient cases, three-dimensional organ geometries are outlined by a physician on a set of CT or MRI images. The physician outlines the GTV (“Gross Tumor Volume,” the tumor region) and OARs (for “Organs At Risk,” also known as “sensitive structures” or “critical structures”). Since the coordinates of these geometries are continuous, they are not directly usable within our optimization models and have to be converted to a discrete set of voxels. A program “gendata” converts these three-dimensional organ geometries into a set of discrete set of voxel coordinates and stores these (named) sets in the gdx file.

For treatment planning, the planning target volume (PTV) must also be constructed. To construct the PTV, one begins with the GTV that encompasses known macroscopic disease. Next, a margin is added to include regions of suspected microscopic disease. The new volume is called the clinical target volume (CTV). An additional margin is added to account for anatomical and patient setup uncertainties. The final volume is the PTV. While this procedure is standard, it has some drawbacks in that some voxels may appear in the PTV and also in a sensitive structure. The normal tissue is implicitly stored in the gdx file by saving a rectangular grid of voxels encompassing all the organs of interest.

The second component of the data is the beamlets. The specific details of their generation are given in Section 30.2.1. We store all the resulting patient specific beamlet data in the same gdx file as the organ structures. It should be noted that the beamlet data is very large and that it does not need to be present in the Matlab environment. We simply use an external program to generate the gdx file, and use the beamlet data only within our optimization models. Such design allows the optimization process to be independent of Matlab if desired. In the 3DCRT case, the beamlet data is given for every voxel indexed over a set of angles and optionally a set of wedges. Section 30.2.1 describes the generation process in detail, and it can be outlined without continuity loss to the reader.

We also provide a suite of problems based on a water cylinder. These problems can be generated using rotational symmetry from a single beamlet and hence can be stored much more economically. Due to this fact, we ensure that our program “gendata” can generate a corresponding gdx file based on this compact representation. The details of the extra input that is needed in this case is given in the Appendix. In the cylinder case, we also provide a Matlab routine “neworgans” to create simulated organ structures, that may be of use in tuning models. This routine enables users to create simulated organ structures within the cylinder.

Finally, the desired or required dose information for each region is specified by the planner, typically as a sequence of dose volume constraints. For example, requirements are of the form “no more than X% of structure A should receive more than Y Gy”. For each patient case, these prescriptions are available from the web site.

The organ structures can be manipulated directly within Matlab. In Matlab, each structure is stored as an m × 3 matrix consisting of the m "voxels" that are part of that structure. We provide two routines that allow a user to read a structure from a gdx file and to write a new structure to a gdx file. Access to these structures is useful both for visualization and for sampling as we demonstrate in the sequel.

30.2.1 Pencil beams and apertures

Modern linear accelerators use a multileaf collimator, located inside the head of the accelerator, to shape the beam of radiation [11, 12]. To calculate the radiation dosage that can be delivered by a beam applied from a given angle, the rectangular aperture obtained by opening the collimator as widely as possible is divided into rectangular subfields arranged in a regular $M \times N$ rectangular pattern, as shown in Figure 30.1. Each of the subfields is called a *pencil beam* or *beamlet.* M represents the number of leaf pairs in the multileaf collimator, while N represents the number of possible settings we allow for each leaf. We identify each beamlet by the index pair *(i,j),* where $i = 1, 2,..., M$ and $j = 1, 2,..., N$. In our work, the leaves of the multileaf collimator are 1 cm wide, and a pencil beam is assigned a length of 0.5 cm. Thus, for a 10 cm by 10 cm field, we would use $M = 10$ and $N = 20$, giving a total of 200 beamlets.

A separate three-dimensional dose distribution is computed for each pencil beam. The dose distribution matrix for each pencil beam from each angle is calculated using a Monte Carlo technique, which simulates the track of individual radiation particles, for a large number of particles. A unit-intensity, non-wedged beam is assumed for the purposes of these calculations. Each dose distribution consists of the radiation deposited by the beam into each of the small three-dimensional regions ("voxels") into which the treatment area is divided.

As described in the introduction, the pencil beam data sets provided by this tool can be used for either 3DCRT or IMRT. For IMRT optimization, the pencil beam intensities are optimized directly. Thus, an optimized intensity map (fluence map) is produced for each beam angle. In conformal radiotherapy, the shape of each beam is set to match the beam's-eye view (BEV) of the tumor volume, which is essentially the projection of the three-dimensional shape of the tumor onto the plane of the multileaf collimator [11, 13-17]. One technique for determining the BEV is to employ a ray-tracing algorithm from the radiation source to the tumor volume, setting the beam's-eye view to include all of the rays that pass through the tumor volume. We use an alternative approach based on the dose matrices of the pencil beams. We include in the BEV all pencil beams whose field of

Figure 30.1 Division of aperture into pencil beams (shaded area represents one beamlet)

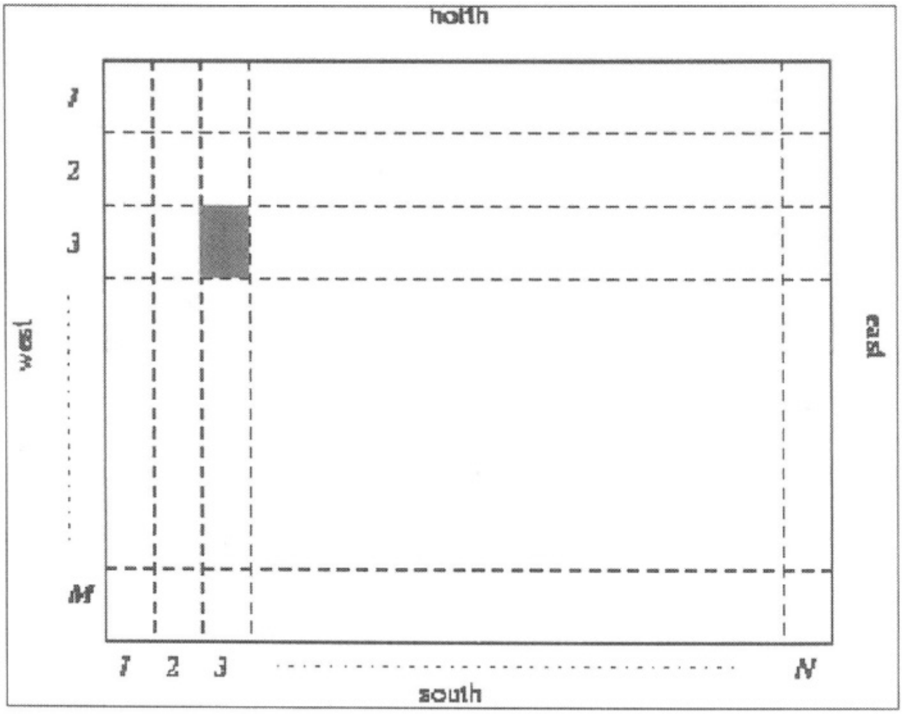
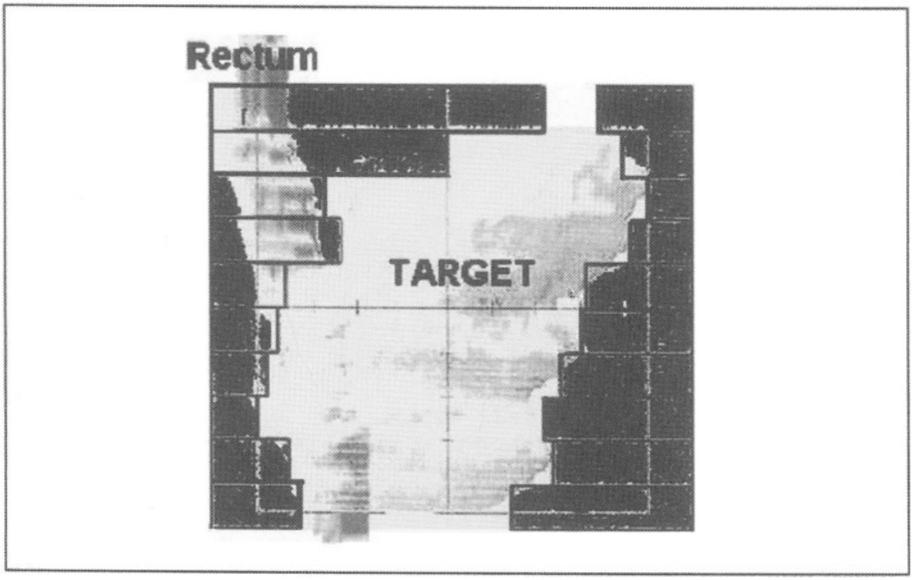


Figure 30.2 An example of beam's-eye view



significant dose intersects with the target region. To be specific, given a threshold value *T,* we include a pencil beam in the BEV if its dose delivered to at least one voxel within the target region is at least *T*% of the dose delivered by that pencil beam to *any* voxel. Figure 30.2 shows an example of a BEV. Once the BEV from a particular angle has been chosen, we can construct the dose matrix for the BEV aperture by simply summing the dose matrices of all the pencil beams that make up the BEV.

The choice of threshold parameter *T* is critical. If the value of *T* used in the determining the BEV is too small, the BEV overestimates the target, producing an aperture that irradiates not only the target but also nearby normal tissue and organs at risk. On the other hand, if the value of *T* is too large, the BEV underestimates the target, and the optimizer might not be able to find a solution that adequately delivers radiation dose within the required range to all parts of the target. The best value of *T* to use depends somewhat on the shape of the tumor. We choose *T* as the minimum value such that the resulting BEVs provide a complete 3D coverage of the target from all beam angles considered in the problem. Based on our experiments, a value of *T* of between 7% and 13% appears to be appropriate.

30.3 OPTIMIZATION MODELS

Optimization models can be used in conjunction with the data outlined above to provide treatment plans that specify how to operate the particular delivery device to generate a dose distribution over the area of interest. The basic optimization models are described more fully in [18], including techniques used to reformulate the problems suitably for optimization solvers.

In conformal radiation therapy, the data that we generate using the procedure outlined above is provided to an optimization model as:

$D_{(i,j,k),A}$ – the dose contribution to voxel *(i,j,k)* from a beam of weight 1 from angle *A*,
T – a collection of voxels on the target,
S – a collection of voxels on the sensitive structure(s),
N – a collection of voxels on the normal tissue

When wedges are allowed in the optimization, the data will be provided as

$D_{(i,j,k),A,F}$ – the dose contribution to voxel *(i,j,k)* from a beam of weight 1 from angle *A,* using wedge orientation *F*

30.3.1 Beam weight optimization

The classical optimization problem in conformal radiation therapy is to choose the weights (or intensity levels) to be delivered from a given a set of angles. While much of the literature reformulates the constraints of the problem into the objective using penalization, we propose here to exploit the power of constrained optimization. Thus, these problems can be cast as a quadratic programming problem, minimizing the Euclidean distance between the dose delivered to each voxel and the prescribed dose [4, 19-21]. Our formulation uses w_A to represent the beam weight delivered from angle *A*, $D_{(i,j,k)}$ for the total dose deposited to voxel *(i,j,k)* and λ to represent the relative weighting factors in the objective function. For simplicity we assume that a target prescription of θ is given, the penalty on overdose in the target is the same as for underdose, and that a representative sensitive structure regards a dose exceeding ϕ as being hot.

$$\begin{array}{lll}
\underset{w}{\min} & \lambda_t \dfrac{\left\|D_T - \theta_{eT}\right\|_2^2}{card(T)} + \lambda_s \dfrac{\left\|(D_S - \phi_{eS})_+\right\|_2^2}{card(S)} + \lambda_n \dfrac{\left\|D_N\right\|_2^2}{card(N)} & (1) \\
s.t. & D_\Omega = \sum_{A \in A} D_{\Omega,A} w_A, \qquad \Omega = T \cup S \cup N, & \\
 & l \le D_T \le u, & \\
 & 0 \le w_A, \qquad \forall A \in A. &
\end{array}$$

Note that $(\bullet)_+ := \max(\bullet, 0)$ can be reformulated using extra constraints and variables as a smooth quadratic as detailed in [18]. Furthermore, we have imposed hard upper and lower bound constraints on the target dose, as well as the objective requirement to be close to the prescription. Other constraints may be useful to deal with prescription constraints at other locations. An implementation of this model is given as *qp.gms* in the Appendix.

Linear programming (LP) has also been extensively used to improve conventional treatment planning techniques [20, 22-25]. The strength of LP is its ability to control *hot* and *cold* spots or integral dose on the organs using constraints, and the presence of many state-of-the-art LP solvers.

An example of an LP formulation is as follows:

$$\min_{w} \quad \lambda_t \frac{\left\|(D_T - \theta_{eT})_+\right\|_1}{card(T)} + \lambda_s \frac{\left\|(D_S - \phi_{eS})_+\right\|_1}{card(S)} + \lambda_n \frac{\left\|D_N\right\|_1}{card(N)} \tag{2}$$

$$\begin{aligned} s.t. \quad & D_\Omega = \sum_{A \in A} D_{\Omega,A} w_A, \qquad \Omega = T \cup S \cup N, \\ & l \le D_T \le u, \\ & 0 \le w_A, \qquad \forall A \in A. \end{aligned}$$

The model here replaces the Euclidean norm objective function with a polyhedral one, for which standard reformulations (see [18]) result in linear programming problems. This model is available as *lp.gms* on the website. While these techniques still suffer from large amounts of data in $D_{(i,j,k),A}$, they are typically solved in acceptable time frames. These models tend to find optimal solutions more quickly than the corresponding quadratic programming formulations.

Another technique to convert the quadratic (or more generally convex) problem to a linear program is via a piecewise-linear approximation of the objective (see [26]). For a quadratic function, a uniform spacing for the breakpoints guarantees small approximation errors from the piecewise linear interpolant [27]. Since the piecewise linear interpolant is convex, standard techniques can be used to reformulate this as a linear program [28]. The paper [27] suggests a particular formulation and the use of an interior point method to solve the resulting problems.

Recently, some of the medical physics literature has been advocating the use of other forms of objective function in place of the ones outlined above. A popular alternative to those given above is that of generalized equivalent uniform dose (EUD). This is defined on a per structure basis as

$$EUD_a(D,\Omega) \equiv \left(\frac{1}{card(\Omega)} \sum_{(i,j,k)\in\Omega} D_{(i,j,k)}^{a} \right)^{1/a}.$$

Note that EUD is a scaled version of the a-norm of the dose to the particular structure, and hence is known to be a convex function for any $a \ge 1$ and concave for $a \le 1$ [29]. Thus the problem

$$
\begin{array}{lll}
\max\limits_{w} & EUD_a(D,T) & \\
s.t. & D_\Omega = \sum\limits_{A \in A} D_{\Omega,A} w_A, & \Omega = T \cup S \cup N, \\
 & EUD_b(D,S) \le \phi, & \\
 & EUD_c(D,N) \le u, & \\
 & 0 \le w_A, & \forall A \in A
\end{array}
$$

is a convex optimization problem provided $a \le 1$ and $b, c \ge 1$. As such, nonlinear programming algorithms will find global solutions to these problems. An example is provided as *eud.gms* on the website.

30.3.2 Beam angle and wedge selection

Optimization also lends itself to solving the more complex problem of selecting which angles to use as well as their intensities. Mixed Integer Programming (MIP) is a straightforward technique for selecting beam angles from among many candidates.

$$
\begin{array}{lll}
\min\limits_{w,\psi} & \lambda_t \dfrac{\left\|(D_T - \theta_{eT})_+\right\|_1}{card(T)} + \lambda_s \dfrac{\left\|(D_S - \phi_{eS})_+\right\|_1}{card(S)} + \lambda_n \dfrac{\left\|D_N\right\|_1}{card(N)} & \\
s.t. & D_\Omega = \sum\limits_{A \in A} D_{\Omega,A} w_A, & \Omega = T \cup S \cup N, \\
 & l \le D_T \le u, & \\
 & 0 \le w_A \le M\psi_A, & \forall A \in A \\
 & K \ge \sum\limits_{A \in A} \psi_A, & \\
 & \psi_A \in \{0,1\}, & \forall A \in A.
\end{array}
\qquad (3)
$$

The variable φ_A is used to determine whether or not to use an angle A for delivery. The choice of M plays a critical role in the speed of the optimization; further advice on its choice is given in [18]. This is implemented as *optangle.gms.*

Finally, we describe an optimization model that simultaneously optimizes beam angles, wedge orientations, and beam intensities. Wedges are placed in front of the collimator to produce a gradient over the dose distribution and can be effective for reducing dose to organs at risk. This can be done by adding an extra dimension F to the variable w_A:

$$\min_{w} \; \lambda_t \frac{\left\|(D_T-\theta_{eT})_+\right\|_1}{card(T)} + \lambda_s \frac{\left\|(D_S-\phi_{eS})_+\right\|_1}{card(S)} + \lambda_n \frac{\left\|D_N\right\|_1}{card(N)} \tag{4}$$

$$\begin{aligned} s.t. \quad & D_\Omega = \sum_{A\in A} D_{\Omega,A,F} w_A, F, \qquad && \Omega = T \cup S \cup N, \\ & w_{A,F} \le M\psi_A, \\ & l \le D_T \le u, \\ & K \ge \sum_{A\in A} \psi_A, \\ & \psi_A \in \{0,1\}, && \forall A \in A. \end{aligned}$$

Note that the data for this problem is considerably larger, increasing by a factor related to the number of wedge orientations allowed. An implementation of (4) is available as *optwedge.gms.* Additional optimization models relating to (4) can also be found in [18].

It should be noted that beam energy and non-coplanar beams can also be treated in a similar fashion to wedges within this framework at the cost of increasing the amount of data.

The models described above have been implemented within the GAMS modeling system. The Appendix gives an example for model (1). Most of the notation used in this GAMS file tries to imitate the mathematical symbols used in (1) with a few exceptions: PTV represents T, OAR is for S, Normal is used for N, and sumDose represents D. For debugging purposes, this model can be executed directly at the command prompt:

```
% gams qp
```

Note that "matsol.gms" is a reserved file for the system.

30.4 OPTIMIZATION PROCESS

We demonstrate the entire treatment planning process using a set of prostate data that is generated by our phantom cylinder configuration. There are two organs in this example, namely "prostate" and "rectum". The tumor volume has 5245 voxels, while the rectum consists of 1936 voxels. We are interested in optimizing beam intensities of 36 beam angles for a treatment plan.

We take two steps to generate data that is needed for the rest of treatment planning. The first step is to collect appropriate input data for the Matlab command "gendata" using the second example given in the Appendix. It creates a Matlab structure array "prob" with data values necessary to utilize the "prostate" example. The second step is to produce the necessary data for the optimization using the inputs (prob) created above:

```
gendata(prob);
```

This generates both a GAMS include file (initdata.gms) and a GDX (GAMS Data Exchange) file [10], typically named *data.gdx,* that are used in the GAMS models. The file *initdata.gms* is a problem specific file that defines sets and parameters that are used in the optimization model. It is described in more detail in the Appendix. It is included at the very beginning of any GAMS files in our toolbox:

```
$include initdata.gms;
```

A Matlab program "rungms" generates a treatment plan based on this data:

```
Dose = rungms('qp');
```

This command uses the interface [30] to execute the GAMS program "qp.gms" of the Appendix. At the end of the run, the Matlab variable "Dose" contains the dose that is delivered as determined by the optimization of the model (1). If necessary, GAMS options can be added followed immediately after the GAMS file name. The last two lines of "qp.gms" facilitate the return of the Dose variable to the Matlab environment.

Since it is typical that a user will wish to update the various organs that are contained in the model and visualize the DVH plots of various structures in the problem, we provide a Matlab command "readstruct" to retrieve the coordinates from the GDX file. For example,

```
PTV = readstruct('data.gdx','prostate');
```

retrieves the three-dimensional coordinates of “prostate” from “data.gdx” into PTV. These coordinates are stored in a Matlab matrix that has a positive number of rows and three columns. Each row holds the three-dimensional coordinates of a given voxel in the structure.

More complex use of “rungms” is as follows.

```
rungms('GAMS_file_name [-options]',organ1,organ2,...,data);
```

Users can specify the GAMS solver options immediately after the GAMS file name. The additional (optional) input arguments are Matlab structures representing organs. Each organ structure must have a name field. The name field must be the string that is the set name used in GAMS. For an example, we define a Matlab structure array for the target, another for the sensitive structure, and the third for the normal tissue as follows:

```
target = struct('name','prostate');
sensitive = struct('name','rectum','data',samplerect);
normal = struct('name','Normal');
```

In the above example, the target (prostate) is extracted from the “data.gdx” file, whereas the sensitive structure (rectum) uses the voxels provided in the samplerect matrix (which has the same format as that described above for PTV). The M-file “rungms” automatically writes out “samplerect” to a GDX file (rectum.gdx) using a Matlab command “writestruct”:

```
writestruct('rectum.gdx',samplerect);
```

The set normal contains the set of those voxels that are neither target nor sensitive and is generated automatically.

The specified GAMS file is executed and returns the three-dimensional matrix of the final dose distribution. This can be used to evaluate the treatment quality using DVH and dose distribution plots as will be explained in the next section.

A final output is always returned, namely the intensities *w:*

```
[Dose,w] = rungms('qp',target);
```

In the “rungms” example immediately above, the intensities are returned in *w*. Note that the “qp” model will only have a target structure in this case, i.e.

the sensitive and normal structures will be empty in the optimization problem.

The amount of voxels in normal can be unnecessarily large. We provide a Matlab command "genrind" that can be useful to reduce the amount of normal voxels. For example,

```
genrind(PTV,5);
```

generates a set of three-dimensional voxel coordinates that surrounds the PTV within a distance of five voxels. The rind of the PTV can be used as a substitute for the normal structure to reduce the number of voxels in the optimization.

Thus the following Matlab commands set up a rind of the PTV as normal tissue, again with no sensitive structures included in the optimization:

```
normal = struct('name','Normal','data',genrind(PTV,5));
Dose = rungms('qp',target,normal);
```

While "initdata.gms" contains default values that allow the model to be run, we also provide a mechanism to update the values of the parameters in the optimization model. For example,

```
data=struct('kBeams',6,'thetaL',0.97,'thetaU',1.05,'phi',struct('rectum',0.4),...
              'lambda',struct('prostate',2,'normal',0.5));
Dose = rungms('qp',target,sensitive,normal,data);
```

generates new values for various parameters in "qp.gms". The data argument (a Matlab structure) must be the last argument to "rungms". To distinguish it from the organ structures, it cannot have a 'name' field. The remaining fields use execution time assignments to update the values of the given fields. Thus, *kBeams* gets reset to the value 6, θ_L to the value 0.97, etc. For parameters defined over one-dimensional sets (e.g., ϕ), we update a subset of its components by specifying new values in a structure. Thus ϕ_{return} will be reset to 0.4 in the above example. The communication mechanism to facilitate this is a file "updatedata.gms".

30.5 VISUALIZATION

Medical experts rely on visualizations of the dose to examine the quality of treatment plans, instead of the objective function that operation researchers typically employ. Two such visualizations are the dose volume histogram (DVH) and a slicewise dose distribution plot. The Matlab routines "dvh" and “doseplot” are provided to allow both sets of users to determine if the quality of solutions are in broad agreement.

30.5.1 Plotting DVH

The quality of a treatment plan is typically specified and evaluated using a DVH. The DVH shows what fraction of the volume receives at least a certain level of dose. To make a DVH plot of the current solution, the final dose distribution and sets of three-dimensional organ coordinates are passed to a Matlab routine “dvh”. “Dose” must be the first input argument for “dvh”, followed by the organs of interest:

```
dvh(Dose,PTV,OAR);
```

This generates a Matlab figure with dose volume histograms for the specified organs. Optionally, the user can specify a line property for the DVH plot as follows:

```
dvh(Dose,PTV,'b-',OAR,'r');
```

where the color blue (‘b-’) with a solid line is specified for the “PTV” and red (‘r’)for the “OAR.”

An example DVH is shown in Figure 30.3. The x-axis is normalized so that the target prescribed dose (θ) is one. The y-axis represents the fraction of the volume. For example, the line of the normal tissue approximately passes through the coordinate (0.2,0.2). This means that 80% of the normal tissue receives 20% or less of the target prescribed dose level. Note that the labels on the structures are created manually using the Matlab figure editor; alternatively the “legend” command could be used.

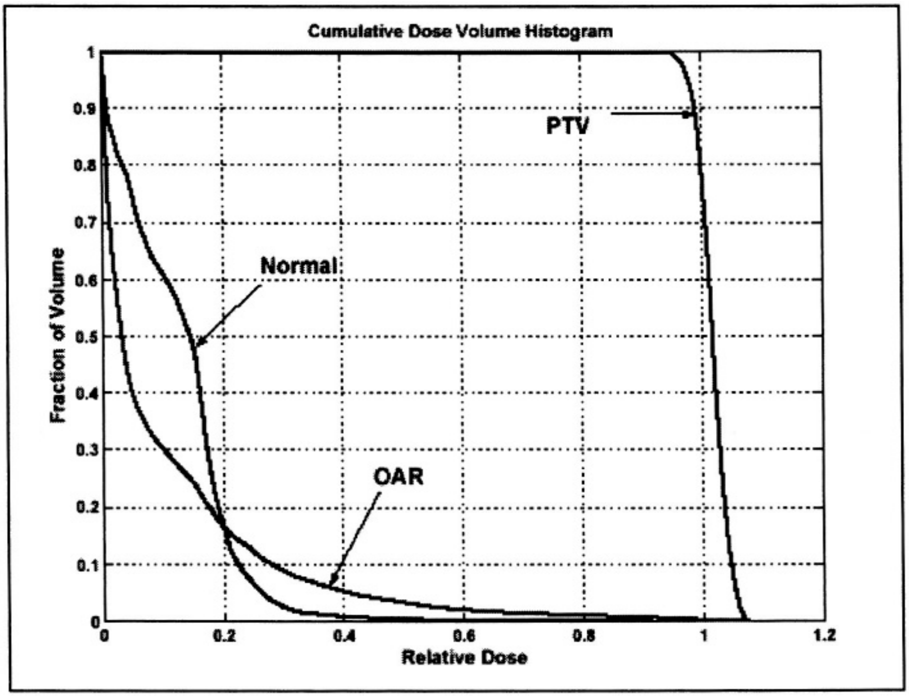
30.5.2 Plotting dose distribution

Although a DVH provides information about the fraction of each organ receiving each dose level, it does not give spatial information with regard to the location of “hot spots”, “cold spots”, or “streaking”. Visualizing the dose distribution becomes important to finalize treatment plans for practical use.

To meet this need, a Matlab routine “doseplot” is provided to examine the dose distribution from different viewpoints: axial, sagittal, and coronal. To visualize the dose distribution of the current solution from the axial viewpoint, the following command suffices:

```
doseplot(Dose);
```

Figure 30.3 Dose volume histogram

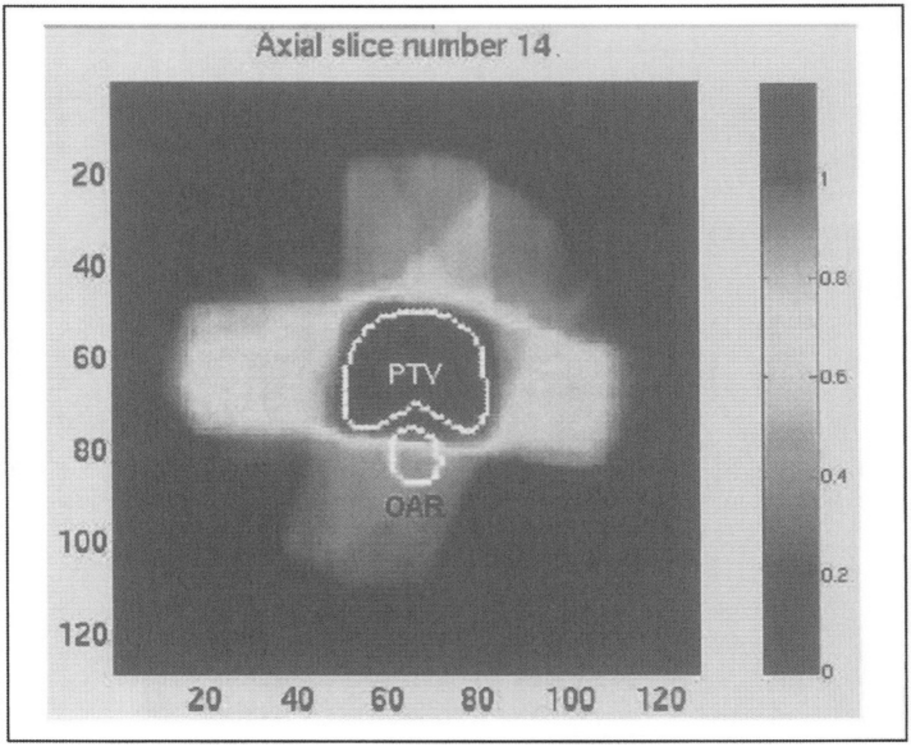


This produces a Matlab figure with dose distribution of the current solution using the default viewpoint ("axial" slice). All slices can be viewed one slice at a time by pressing any button on the keyboard. Optional arguments are also allowed. These include sets of three-dimensional organ coordinates, a string ("axial", "coronal" or "sagittal") representing the viewpoint of the image, and a vector of slice numbers:

```
doseplot(Dose,PTV,OAR,'axial');
doseplot(Dose,PTV,OAR,'axial',15:20);
```

In the above examples, the PTV and OAR structures will be outlined on the same slice of the dose distribution. The slice number must come after the choice of the viewpoint. The PTV and OAR structures are indicated on the resulting "axial" output as shown in Figure 30.4. The second case allows only slices 15 to 20 to be displayed as opposed to all slices. In a color display of this figure, the color blue would represent the cold spots while red would indicate hot spots, as depicted in the Matlab color bar located on the right side of Figure 30.4. For each structure, an optional argument facilitates the change of the outline color of a given organ, e.g.,

Figure 30.4 Dose distribution plot in axial slice



```
doseplot(Dose,PTV,'b',OAR,'coronal');
```

outlines the target in blue on all coronal slices.

30.6 EXAMPLE USAGE OF TOOLBOX

We outline the use of some of the tools described above for a particular example of 3D conformal radiation therapy treatment planning in a prostate case. We use the model format (3) that has been encoded in a GAMS file "optangle.gms".

To start the process, we use the Matlab command "gendata" to generate files *initdata.gms* and *data.gdx* as discussed in Section 30.4. These files are required for running optimization models. Using the first example of the Appendix,

```
gendata(prob);
```

generates “initdata.gms” and uses the existing “data.gdx” file. Alternatively, a new GDX file can be generated as discussed in the second example of the Appendix.

A Matlab command “readstruct” is used to extract the organ geometries from the GDX file into the Matlab workspace:

```
prostate = readstruct('data.gdx','prostate');
rectum = readstruct('data.gdx','rectum');
```

As discussed in Section 30.4, some parameters of the optimization model can be updated as follows:

```
data                                                                   =
struct('kBeams',6,'thetaL',0.97,'thetaU',1.05,'phi',struct('rectum',0.4),...
                'lambda',struct('normal',0.5));
```

30.6.1 Data sampling

A large number of voxels comprise the normal tissue. Although voxels in the normal tissue are important for the final treatment plan, it is known that using sampled voxels of the normal structure in the optimization problem has an insignificant effect on the optimal treatment plan dose distribution [18]. A random sampling of voxels can also be used to speed up the computation [31].

In addition to these data reduction techniques, rather than using the complete sets of organ structures, we follow an interactive approach that tries to determine promising beam angles based on an *a%* sampling of the sensitive structure as outlined in [18]. The sampling scheme is also noted elsewhere [32].

A self-explanatory example of our sampling scheme is as follows:

```
goodAngles = zeros(prob.nAngle,1); rate = 0.1; nsamples = 10;
for s = 1:nsamples
   % sample a subset of prostate
   sample = find(rand(length(prostate),1) < rate);
   target = struct('name','prostate','data',prostate(sample,:));
   % sample a subset of rectum
   sample = find(rand(length(rectum),1) < rate);
    sens = struct('name','rectum','data',rectum(sample,:));
   % sample a subset of normal
   sample = find(rand(length(normal),1) < rate/10);
```

```
  snormal = struct('name','normal','data',normal(sample,:));
  % run a GAMS optimization model
  [Dose,w] = rungms('optangle lo=1 optfile=1 optcr=0.03',...
                       target,sens,snormal,data);
  used = find(w > 0);
  goodAngles(used) = goodAngles(used) + 1;
end
```

We used these sampled problems (that solve very quickly) to determine which subset of the angles are promising. The Matlab vector “goodAngles” contains a count of sample how many solutions used as given angle. We then resolve the full model using just those angles that were used in more than one sample case.

```
fixangles = find(goodAngles<=1)-1; % index adjustment
wup = [fixangles,zeros(length(fixangles),1)];
data.wup = wup;
```

Note that the Matlab structure “data” holds new values for parameters that are already declared in the GAMS model “optangle”. In particular, the parameter “wup” is used to set new upper bounds on the weights in the model. By resetting some of these to be zero, the model ignores the corresponding bad angles.

We continue to sample the normal structure, but include all those voxels that are close to the target.

```
target = struct('name','prostate');
sens = struct('name','rectum');
sample = find(rand(length(normal),1) < rate);
newnormal = [genrind(prostate,3);normal(sample,:)];
snormal = struct('name','normal','data',newnormal);
[Dose,w] = rungms('optangle lo=1 optfile=1
optcr=0.03',target,sens,snormal,data);
dvh(Dose,prostate,rectum);
```

30.6.2 Refining solutions

At this stage, the DVH may show a sensitive structure that is violating a “dvh” constraint. To rectify this situation, we have three recourses. We can update the parameter ϕ in (3) using

```
data.phi.rectum = 0.2;
```

This will penalize dose on the rectum that exceeds 0.2, instead of the original value of 0.4. Secondly, we can update the penalty parameters in relation to the organ structures:

```
data.lambda = struct('prostate',0.9,'rectum',2,'normal',0.2);
```

Thirdly, we can update the structure itself to allow a subset of the problem voxels to violate the constraints:

```
rectumBottom = extract(Dose,rectum,'bottom',0.82);
```

By executing the above line, the Matlab command "extract" sorts the values of the dose delivered to all voxels in the rectum. Then it extracts (the bottom) 82% of the voxels that received the low dose. The sensitive structure is updated using this new set of voxels in the rectum:

```
sensitive = struct('name','rectum','data',rectumBottom);
```

This is a heuristic, based on a DVH that allows 20% of the voxels to be essentially ignored; we actually ignore slightly less than this. Note that the command "extract",

```
extract(dose_matrix,organ_coordinates,option_string,value);
```

offers two different option types for the user to sample a fraction of voxels in the specified organ structure: *absolute option* and *relative option.* The relative option was displayed in the previous example of "extract". The same command line can be replaced by

```
rectumBottom = extract(Dose,rectum,'top',0.18);
```

Two options are available for the absolute option: "above" and "below". The command

```
extract(Dose,rectum,'above',0.9);
```

can be used to extract all voxels that received more than 90% of the target prescribed dose in the rectum.

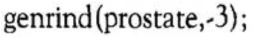
Resolving the newly updated problem,

```
Dose = rungms('optangle',target,sensitive,data);
dvh(Dose,prostate,rectum);
```

gives a new DVH. Furthermore, other changes to the parameters of the formulation may be carried out, or changes to the formulation itself can be implemented by simply updating the GAMS model file, or the Matlab solution process.

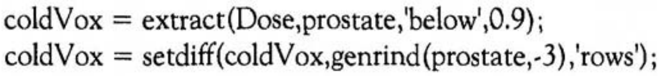
Another example where iterative solution can be useful is in the treatment of location-specific hot or cold spots. The DVH plot does not distinguish between the locations of such spots, but clinically cold spots in the center of a tumor location are more serious than those on the periphery. If one of the models (1), (2) or (3) is used, then this should not be a problem since the model incorporates hard upper and lower bound constraints on the dose in the tumor. However, in practice, these hard constraints can lead to infeasibilities or overly homogeneous solutions, and may be dropped or relaxed within the modeling process.

To isolate voxels on the periphery of a given structure we can use the routine “genrind”. In the following example, we extract the “internal” rind of size 3 from the tumor (in this case, the prostate):



```
genrind(prostate,-3);
```

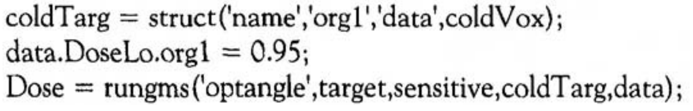
To find the cold spots that are centrally located in the tumor we invoke the routine “extract” to find the voxels that are dosed below a given cutoff value, and throw away those voxels that are close to the periphery as follows:



```
coldVox = extract(Dose,prostate,'below',0.9);
coldVox = setdiff(coldVox,genrind(prostate,-3),'rows');
```

Similar techniques could be used to find hot spots within a sensitive structure.

In order to pass new constraints on these substructures to the optimization routines, we use any number of additional organ structures that our toolbox provides by default. These additional organs are called *org*1,..., *org*100. Thus, to impose a strict lower bound of 0.95 on the dose in the center of the tumor, the following additional lines would suffice:



```
coldTarg = struct('name','org1','data',coldVox);
data.DoseLo.org1 = 0.95;
Dose = rungms('optangle',target,sensitive,coldTarg,data);
```

The additional organs could also be subject to soft constraints - the user just has to define ϕ and λ appropriately. For example,

```
data.lambda.org1 = 10;
Dose = rungms('optangle',target,sensitive,coldTarg,data);
```

30.6.3 Streaking control

When relatively few angles are used for treatment, "streaking" can occur in the normal tissue. Essentially, high dose is delivered along particular rays that due to the decay mechanism results in hot spots of radiation close to beam entry locations. These hot spots must be removed before the treatment plan is finalized. We take two steps to do this. First, the normal structure is divided into two distinct sets, *innormal* and *outnormal.* The set "innormal" is an external rind of the target structure while "outnormal" is defined as the remainder. Second, a strict dose upper bound is applied only on "outnormal". Note that "innormal" is not considered in the optimization model. The following lines of Matlab code show an implementation of this method that also uses sampling to reduce the normal voxels considered.

```
sample = find(rand(length(normal),1) < rate/10);
innormal = genrind(prostate,6);
outnormal = setdiff(normal(sample,:),innormal,'rows');
snormal = struct('name','normal','data',outnormal);
data.DoseUp.normal = 0.8;
[Dose,w] = rungms('optangle lo=1 optfile=1
optcr=0.03',target,sens,snormal,data);
```

The same technique using additional organs *org*1,..., *org*100 could be used to impose different bounds on different pieces of the structure. Thus, to remove hot spots in the normal tissue close to the tumor whilst maintaining streaking control, we would use an additional organ *org*2, for example:

```
org2 = struct('name','org2','data',innormal);
data.DoseUp.org2 = 1.05;
[Dose,w] = rungms('optangle lo=1 optfile=1
optcr=0.03',target,sens,snormal,org2,data);
```

30.7 CONCLUSIONS

We have provided clinical data for radiation therapy treatment planning within the Matlab environment. We have shown how to use a small suite of programs (gendata, rungms, genrind, extract, dvh, doseplot) in conjunction with a library of optimization models to provide an effective and adaptive treatment planning procedure. We have demonstrated the utility of the Matlab environment to facilitate more complex data and optimization control.

References

[1] Bortfeld, T. and W. Schlege (1993). Optimization of beam orientations in radiation-therapy – some theoretical considerations. *Physics in Medicine and Biology,* 38, 291-304.

[2] Bortfeld, T., J. Burkelbach, R. Boesecke, and W. Schlegel (1990). Methods of image reconstruction from projections applied to conformation radiotherapy. *Physics in Medicine and Biology,* 25, 1423-1434.

[3] Bortfeld, T., A.L. Boyer, W. Schlegel, D.L. Kahler, and T.J. Waldron (1994). Realization and verification of three-dimensional conformal radiotherapy with modulated fields. *International Journal of Radiation Oncology: Biology, Physics,* 30, 899-908.

[4] Chen, Y., D. Michalski, C. Houser, and J. Galvin (2002). A deterministic iterative least-squares algorithm for beam weight optimization in conformal radiotherapy. *Physics in Medicine and Biology,* 47, 1647-1658.

[5] Intensity Modulated Radiation Therapy Collaborative Working Group (2001). Intensity-modulated radiotherapy: Current status and issues of interest. *International Journal of Radiation Oncology: Biology, Physics,* 51, 880-914.

[6] Jordan, T.J. and P.C. Williams (1994). The design and performance characteristics of a multileaf collimator. *Physics in Medicine and Biology*, 39, 231-251.

[7] Tervo, T. and P. Kolmonen (2000). A model for the control of a multileaf collimator in radiation therapy treatment planning. *Inverse Problems,* 16, 1875-1895.

[8] Webb, S. (1998). Configuration options for intensity-modulated radiation therapy using multiple static fields shaped by a multileaf collimator. *Physics in Medicine and Biology,* 43, 241-260.

[9] Wu, X. and Y. Zhu (2001). A global optimization method for three-dimensional conformal radiotherapy treatment planning. *Physics in Medicine and Biology,* 46, 109-119.

[10] van der Eijk, P. (2002). *GDX facilities in GAMS.* GAMS Contributed Software, http://www.gams.com/contrib/GDXUtils.pdf.

[11] Goitein, M., M. Abrams, S. Rowell, H. Pollari, and J. Wiles (1983). Multi-dimensional treatment planning: I. beam's eye view, back projection, and projection through ct sections. *International Journal of Radiation Oncology: Biology, Physics,* 9, 789-797.

[12] Webb, S. (1997). *The Physics of Conformal Radiotherapy.* Bristol, UK: Institute of Physics Publishing.

[13] Brewster, L., G.S. Mageras, and R. Mohan (1993). Automatic-generation of beam apertures. *Medical Physics,* 20, 1337-1342.

[14] Chen, G.T.Y., D.R. Spelbring, C.A. Pelizzari, J.M. Balter, L.C. Myrianthopoulous, S. Vijayakumar, and H. Halpern (1992). The use of beam eye view volumetrics in the selection of noncoplanar radiation portals. *International Journal of Radiation Oncology: Biology, Physics,* 23, 153-163.

[15] Cho, B.C.J., W.H. Roa, D. Robinson, and B. Murray (1999). The development of target-eye-view maps for selection of coplanar or non-23 coplanar beams in conformal radiotherapy treatment planning. *Medical Physics,* 26, 2367-2372.

[16] McShan, D.L. B.A. Fraass, and A.S. Lichter (1989). Full integration of the beam's eye view concept into computerized treatment planning. *International Journal of Radiation Oncology: Biology, Physics,* 18, 1485-1494.

[17] Myrianthopoulos, L.C., G.T.Y. Chen, S. Vijayakumar, H. Halpern, D. R. Spelbring, and C.A. Pelizzari (1992). Beams eye view volumetrics – an aid in rapid treatment plan development and evaluation. *International Journal of Radiation Oncology: Biology, Physics,* 23, 67-375.

[18] Lim, J., M.C. Ferris, S.J. Wright, D.M. Shepard, and M.A. Earl (2002). An optimization framework for conformation radiation treatment planning. Optimization Technical Report 02-08, Computer Sciences Department, University of Wisconsin, Madison, WI.

[19] Redpath, A.T., B.L. Vickery, and D.H. Wright (1976). A new technique for radiotherapy planning using quadratic programming. *Physics in Medicine and Biology,* 21, 781-791.

[20] Shepard, D.M., M.C. Ferris, G. Olivera, and T.R. Mackie (1999). Optimizing the delivery of radiation to cancer patients. *SIAM Review,* 41, 721-744.

[21] Starkschall, G. (1984). A constrained least-squares optimization method for external beam radiation therapy treatment planning. *Medical Physics,* 11, 659-665.

[22] Bahr, G.K., J.G. Kereiakes, H. Horwitz, R. Finney, J. Galvin, and K. Goode (1968). The method of linear programming applied to radiation treatment planning. *Radiology,* 91, 686-693.

[23] Langer, M. and J. Leong (1987). Optimization of beam weights under dose volume restriction. *International Journal of Radiation Oncology: Biology, Physics,* 13, 1255-1260.

[24] Morrill, S., R. Lane, J. Wong, and I.I. Rosen (1991). Dose-volume considerations with linear programming. *Medical Physics,* 6, 1201-1210.

[25] Rosen, I.I., R. Lane, S. Morrill, and J. Belli (1990). Treatment plan optimization using linear programming. *Medical Physics,* 18, 141-152.

[26] Romeijn, H.E., R.K. Ahuja, and J.F. Dempsey (2003). A new linear programming approach to radiation therapy treatment planning problems. Technical Report, University of Florida, Gainesville, FL.

[27] Kontogiorgis, S. (2000). Practical piecewise-linear approximation for monotropic optimization. *INFORMS Journal on Computing,* 12, 324-340.

[28] Ho, J.K. (1985). Relationships among linear formulations of separable convex piecewise linear programs. *Mathematical Programming Study,* 24, 126-140.

[29] Choi, B. and J. O'Deasy (2002). The generalized equivalent uniform dose function as a basis for intensity-modulated treatment planning. *Physics in Medicine and Biology,* 47, 3579-3589.

[30] Ferris, M.C. (1998). MATLAB and GAMS interfacing optimization and visualization soft ware. Mathematical Programming Technical Report 98-19, Computer Sciences Department, University of Wisconsin, Madison, WI.

[31] Rowbottom, C.G., V.S. Khoo, and S. Webb (2001). Simultaneous optimization of beam orientations and beam weights in conformal radiotherapy. *Medical Physics,* 28, 1696-1702.

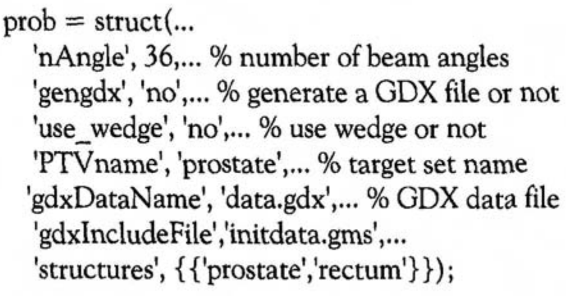
[32] Langer, M., R. Brown, M. Urie, J. Leong, M. Stracher, and J. Shapiro (1990). Large-scale optimization of beam-weights under dose volume restrictions. *International Journal of Radiation Oncology: Biology, Physics,* 18, 887-893.

Appendix

A. Problem Specification

The routine "gendata" reads in user input that specifies the total number of beam angles considered in the problem, a flag indicating if wedges are to be used, the name of the PTV, GAMS set names for the organ structures, the GAMS include file name that will store set and parameter definitions, organ geometries, and the dose matrix. This data is provided as a Matlab structure.

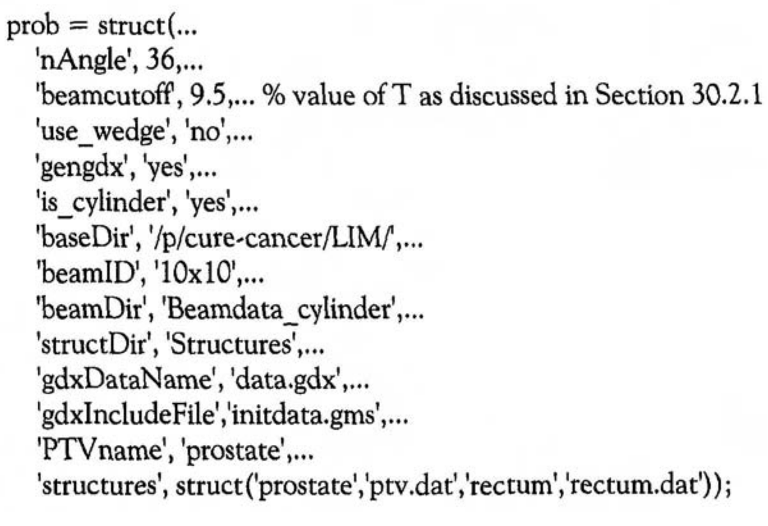
For documentation purposes, we have provided Matlab code that generates the appropriate structure for some existing examples. The first example



```
prob = struct(...
   'nAngle', 36,... % number of beam angles
   'gengdx', 'no',... % generate a GDX file or not
   'use_wedge', 'no',... % use wedge or not
   'PTVname', 'prostate',... % target set name
  'gdxDataName', 'data.gdx',... % GDX data file
   'gdxIncludeFile','initdata.gms',...
   'structures', {{'prostate','rectum'}});
```

retrieves its data from a GDX file. Note that the flag "gengdx" is set to "no" here since the GDX file is assumed to exist already.

The second example shows how to utilize the existing Beamdata and structure information to generate a new GDX file:



```
prob = struct(...
   'nAngle', 36,...
   'beamcutoff', 9.5,... % value of T as discussed in Section 30.2.1
   'use_wedge', 'no',...
   'gengdx', 'yes',...
   'is_cylinder', 'yes',...
   'baseDir', '/p/cure-cancer/LIM/',...
   'beamID', '10x10',...
   'beamDir', 'Beamdata_cylinder',...
   'structDir', 'Structures',...
   'gdxDataName', 'data.gdx',...
   'gdxIncludeFile','initdata.gms',...
   'PTVname', 'prostate',...
   'structures', struct('prostate','ptv.dat','rectum','rectum.dat'));
```

B. File initdata.gms

This file is generated automatically from the problem input described above. This file has six components. First, basic sets and their dimensions are defined for solving the optimization problems. An example is shown below.

```
Sets   I       /0*125/,
       J       /0*125/,
       K       /0* 31/,
       nAngle /0*35/,
       angle(nAngle);
```

Note that the set “angle” is a subset of all angles considered in the optimization. This set is particularly useful for MIP problems ((3) and (4)) to automatically reduce the solution search space. The use of “angle” can be seen in the Appendix.

The next component is to define sets of the three-dimensional coordinates and their dimensions. Sets “prostate” and “rectum” provide coordinates of organs, “PTV”, “OAR”, and “Normal” are auxiliary set definitions for the target, the sensitive, and the normal structures respectively. The name of the parameter to store the dose distribution is also defined:

```
Sets          prostate(I,J,K), rectum(I,J,K),
              PTV(I,J,K), OAR(I,J,K), Normal(I,J,K);
Parameter     Dose(I,J,K,nAngle);
```

The third component is to define the collection of organ names:

```
Set allorgans /'prostate','rectum','Normal', org1*org100/;
```

Note that the set “allorgans” has dummy organs “orgl*org 100”. These extra empty organs play an important role when new organ structures (not defined in the GDX file) are added. As shown in Section 30.6, an immediate application of this feature can be seen on the local dose control over a subset of an organ structure.

In the next component, global variables for the GAMS model are defined for the target and the critical structures. We then define a set “organs” as the collection of critical structures of interest for the particular instance. In our example, the following three lines

```
$setglobal target 'prostate'
$setglobal critical 'rectum'
```

```
Set organs(allorgans) /%critical%/;
```

make the global variable target contain “prostate” and the variable critical contain “rectum”. If there is more than one critical (sensitive) structure, the string “rectum” is replaced by a single quoted string of comma separated organ names, e.g., ‘rectum, bladder’. The set “organs” is defined over the critical structure “rectum”. The “rungms” interface automatically updates this set if a user excludes organs from consideration as outlined in Section 30.4.

All necessary data for the optimization is stored (by “gendata”) in GAMS GDX format. Therefore, the next component is written to retrieve the data in the GAMS file:

```
$GDXIN data.gdx
$LOAD PTV=%target% prostate rectum
$LOAD Dose
$GDXIN
```

“$GDXIN data.gdx” opens the GDX file “data.gdx” for input. The second line is used to load sets from “data.gdx”. Note that a set can be renamed at this stage: the stored set “target” is now named “PTV” in the GAMS file. The last line “$GDXIN” closes the file “data.gdx”.

Finally, a set “Sensitive” of the sensitive structures is defined as a collection of “allorgans” in the GAMS file. Each organ must be defined explicitly as shown in the second line below:

```
Set Sensitive(I,J,k,allorgans);
Sensitive(I,J,K,'rectum') = yes$rectum(I,J,K);
```

Note that gendata forms this file for a given input of the types shown at the start of the Appendix.

C. Program qp.gms

This program solves the 3D conformal radiation treatment problem.

The solution includes: optimal beam weights

```
option limrow=0, limcol=0, solprint=off;
scalar theta 'dose level prescribed for target' /1.0/;
scalar thetaU 'hot spot control parameter on the target' /1.07/;
scalar thetaL 'cold spot control parameter on the target' /0.9/;
scalar ubar 'dose upper bound on the target voxels' /1.15/;

*-- Read in dose matrix and set definition of the model --*

$include initdata.gms

*-- Define parameters --*

Parameter phi(allorgans) 'hot spot control parameters for organs';
phi(allorgans) = 0.3;
Parameter wup(nAngle) 'upper bound of w(nAngle)';
wup(nAngle) = inf;
Parameter DoseUp(allorgans) 'dose upper bound on organ structures';
DoseUp(allorgans) = inf;
Parameter DoseLo(allorgans) 'dose lower bound on organ structures';
DoseLo(allorgans) = -inf;
Parameter lambda(allorgans) 'objective function penalty parameters';
lambda(allorgans) = 1;

*-- Optimization model definition --*

Positive variables w(nAngle), dS(I,J,K,allorgans);
Variable        sumDose(I,J,K), z;
Equations       Def4sens(I,J,K,allorgans),
                Def4sumDose(I,J,K),
                Obj;
Def4sumDose(I,J,K)$(PTV(I,J,K) or OAR(I,J,K) or Normal(I,J,K)) ..
   sumDose(I,J,K) =e= sum(angle,Dose(I,J,K,angle)*w(angle));
Def4sens(OAR,organs)$Sensitive(OAR,organs) ..
   -sumDose(OAR) + dS(OAR,organs) =g= -phi(organs);
Obj ..
   z =e= lambda('%target%')*sum(PTV,sqr(sumDose(PTV)))/card(PTV)
         + sum(organs,lambda(organs)*sum(OAR$Sensitive(OAR,organs),

sqr(dS(OAR,organs)))/max(1,sum(OAR$Sensitive(OAR,organs),1)))
         + lambda('normal')*sum(Normal,sqr(sumDose(Normal)))
                 /max(1,card(Normal));

*--- Bound constraints & Data update ---*

Normal(I,J,K)=yes;
$if exist updatedata.gms $include updatedata.gms
wup(nAngle) = min(wup(nAngle),ubar/smax(PTV,Dose(PTV,nAngle)));
```

```
w.up(nAngle) = wup(nAngle);
angle(nAngle) = yes$(w.up(nAngle) gt w.lo(nAngle));
OAR(I,J,K) = yes$sum(organs$Sensitive(I,J,K,organs),1);
Normal(PTV) = no; Normal(OAR) = no; OAR(PTV) = no;
sumDose.up(PTV) = DoseUp('%target%');
sumDose.lo(PTV) = DoseLo('%target%');
sumDose.up(Normal) = DoseUp('normal');
loop(organs,
   sumDose.up(I,J,K)$(Sensitive(I,J,K,organs)) = DoseUp(organs);
);
model conf / all/;
solve conf using nlp minimizing z;

*--- Write out the solution for Matlab users ---*

sumDose.l(I,J,K) = sum(angle,Dose(I,J,K,angle)*w.l(angle));
$libinclude matout sumDose.l I J K
$libinclude matout w.l nAngle nWedge
```

31 TRANSMISSION MODEL ANALYSIS OF NONTYPEABLE HAEMOPHILUS INFLUENZAE

James S. Koopman[1,2], Ximin Lin[1], Stephen E. Chick[3,4]
and Janet R. Gilsdorf[1,5]

[1] Department of Epidemiology
University of Michigan
Ann Arbor, MI 48109

[2] Center for the Study of Complex Systems
University of Michigan
Ann Arbor, MI 48109

[3] Department of Industrial and Operations Engineering
University of Michigan
Ann Arbor, MI 48109

[4] INSEAD
Fontainbleau, France

[5] Department of Pediatrics
University of Michigan
Ann Arbor, MI 48109

SUMMARY

The effects of immunity stimulated by natural colonization with Nontypeable *Haemophilus influenzae* (NTHi) were assessed using population models of transmission and data from the literature on NTHi colonization prevalence by age, NTHi acute otitis media (AOM) incidence by age, NTHi antibody levels, and colonization duration. The models allowed both contact patterns and immunity to influence colonization and disease patterns by age. To fit the data, the models required colonization to stimulate immunity affecting both transmission (susceptibility and contagiousness) and pathogenicity (AOM given colonization). Model analysis demonstrated that immunity affecting transmission influenced AOM incidence in the first year of life from 4.6 to 39.5 times as much as immunity reducing pathogenicity. This differential decreased with age until age three and then rose again. It was important, however, across all age groups. The conclusion that immunity affecting transmission had larger effects on AOM incidence than immunity affecting pathogenicity was robust to model form and to reasonable variation in the data. Because sensitivity to NTHi strain interactions and age patterns of infection by strain could not be assessed and because data on the distribution of NTHi strains across all ages are deficient, this conclusion must still be viewed as tentative. Nonetheless, these results make it imperative that trials of potential NTHi vaccines be designed to insure accurate assessment of effects on transmission. The models presented here provide the basis for the construction of discrete individual simulation models for use in designing the most informative and powerful vaccine trials.

KEY WORDS

Infection, Models, Vaccines, Transmission, Bacteria

31.1 INTRODUCTION

It is often unclear whether vaccines should be developed to protect against transmission of infectious agents or to protect against illness given transmission. Likewise, it is controversial whether vaccine trials should be designed to detect effects on transmission or just protection against disease. The *de facto* stance has been just to assess effects on disease. We address this issue here for Nontypeable *Haemophilus influenzae* (NTHi) where little is known about the extent to which naturally acquired immunity keeps NTHi from passing from one person's throat to another's or the extent to which such immunity prevents disease once a person's throat is colonized. The analysis we performed used a model of NTHi colonization and transmission that kept NTHi at its endemic levels. Our results show that immunity against transmission has larger effects on AOM incidence than immunity against disease given that the agent is in the throat.

We have organized our presentation first with a statement of the problem and presentation of some background on NTHi and vaccines in Section 31.2. In Section 31.3 we present our methods and the data to which we fit our models. Section 31.4 presents the results with a focus on comparing inferred immune effect on transmission versus disease and the robustness of our conclusions to a wide spectrum of model and data variations. Finally in Section 31.5 we discuss our results with a focus on explaining their significance and encouraging engineers to work with epidemiologists so that more analyses of this type will be performed.

31.2 BACKGROUND

31.2.1 NTHi

Haemophilus influenzae are bacteria that inhabit the human nasopharynx. They are classified by the presence of their capsular polysaccharide as types *a* through *f* or, in the absence of the capsule, as nontypeable. When *Haemophilus influenzae* are found in someone's nose or throat, we say that person is colonized. Most often colonization is asymptomatic but these organisms do cause several types of human disease. Type b strains cause bacteremia, septic arthritis, cellulitis, and meningitis – invasive infections that are successfully prevented by the *Haemophilus influenzae* type b (Hib) vaccine, which consists of the type b capsule. Nontypeable H. influenzae (NTHi) cause respiratory tract infections such as acute otitis media (AOM) and sinusitis in healthy individuals, bronchitis in patients with underlying pulmonary disease such as chronic obstructive pulmonary disease and cystic fibrosis, and pneumonia in children in developing countries. NTHi generate medical care costs of greater than $1 billion per year in the US. Indirect costs in lost work are even greater.

These bacteria are highly transmissible and endemically present at all times in most populations. Their only reservoir is the human nasopharynx. Although antibiotics are usually highly effective in treating H influenzae infections, emerging antibiotic resistance will continue to compromise antibiotic efficacy. Thus, prevention by vaccines offers more effective, more enduring, and less costly control of these infections. Progress is being made in developing vaccines against NTHi [1]. Because a simple vaccine target like the capsule of Hib is not available for NTHi, diverse approaches to vaccine development are being pursued and are likely to stimulate varying degrees and types of immunity against H. influenzae transmission or disease.

31.2.2 Immunity against transmission or disease

Immunity effects upon transmission can keep NTHi from passing from one person to another either by impairing the ability of NTHi to colonize the throat of immune individuals, reducing the ability of immune individuals to spread NTHi when they are colonized, or by reducing the time that NTHi stay in a person's throat, thus shortening the time when the individual can transmit the infection.

Immunity has both direct effects on the person who has acquired the immunity and indirect effects on other individuals who may be spared infection because of someone else's immunity. There are three sources of indirect effects. First, vaccinated individuals who are directly protected against colonization by the vaccine's susceptibility effect are no longer a source of infection to other individuals. This sets off a reverberating effect on transmission dynamics that reduces the risk of everyone down all chains of transmission stopped by vaccination. Second, vaccinated individuals who do get colonized can be made less contagious by immunity. In that case we say that the vaccine has a contagiousness effect. That will reduce the infection risk of everyone with whom they come in contact, vaccinated and unvaccinated alike. Again this effect will reverberate down chains of transmission. Third, vaccination can reduce the duration of colonization, thus reducing the time available to transmit infection.

Pathogenicity is defined as the fraction of colonization episodes that result in AOM. An immune response affecting pathogenicity keeps bacteria that colonize the throat from causing symptoms in the respiratory tract, including the ear. When immunity only affects pathogenicity and not transmission, by definition there are no indirect effects. Because asymptomatic colonization of the nose and throat with NTHi is so common, there has been a tendency to think that vaccines should be developed to prevent disease given colonization rather than to reduce NTHi transmission. The analysis we present here indicates that would be a mistake.

31.2.3 Approaches to separately estimating effects on transmission or disease

The analysis we present uses deterministic compartmental models of NTHi transmission to infer immune effect parameters that explain patterns of NTHi colonization prevalence and AOM incidence. It might seem that inferences about the immunity stimulated by NTHi colonization could be better made on the basis of more direct observations. Such direct observations would involve following children in contact with each other carefully and determining the rate of transmission between them and the duration of colonization as a function of the number of times children have been previously colonized. Colonization can only be detected by swabbing the throat and nasopharynx. To avoid missing colonization episodes, swabs should be taken twice a week or more often. From each isolate one needs refined molecular distinction between NTHi strains from the colonized subjects and their contacts. That is because it is so easy for an infection acquired elsewhere to look like a transmission between study children. Because immunity is acquired slowly over the course of many colonization episodes, the time that children would have to be followed would cover the entire preschool years. Not only would this be very expensive, it would be very difficult: swabbing would have to be more frequent and over longer periods than most parents would tolerate, as this is not a pleasant experience for their children.

31.2.4 A changing spectrum of vaccines

Vaccines have been a mainstay of public health for at least seven decades. In the early years, the science that went into developing and evaluating vaccines proceeded mainly on a trial and error basis. Theories of infectious agent disease causality and immune protection were simple and the evaluation of these theories needed only tests of hypothesis rather than detailed causal model analyses. For the infections to which vaccines were directed, it was possible to empirically devise effective vaccines just by denaturing the agents or by developing non-pathogenic variants of them. The focus of early vaccines was on preventing disease, not infection. There was almost no perceived need to analyze vaccine effects on infection transmission dynamics through populations.

Currently, vaccines for more complex infectious agents are being assessed using more complex theories at both the individual level (how vaccines act within individuals) and the population level (how vaccines affect the circulation of infection in populations). To address this complexity, medical scientists and epidemiologists need to collaborate with experts in the analysis of complex system behavior.

Recent vaccine success against *Haemophilus influenzae* type b (Hib) has motivated greater interest in the effects of vaccines on the circulation of bacterial agents. Hib vaccine development and evaluation focused almost wholly on vaccine effects against disease. Hib disease, although most frequent in infants and toddlers, was occasionally found in immunocompromised or elderly adults. Hib colonization was felt to be nearly universal in childhood but only a small fraction of colonized individuals suffered serious disease. Colonization in older age groups was likely to represent repeat colonization. Given this picture, it was thought that the vaccine would not have much of an effect on infection levels – especially in individuals who were not vaccinated. It was thus a surprise to find that vaccination of a small fraction of children reduced circulation of Hib to near zero in many locations [2]

The Hib vaccine effect on transmission is most likely explained by five interacting characteristics:

1. Hib presents a simple and uniform antigenic face to the outside world, as it is covered completely with a type b polysaccharide coat;

2. this type b polysaccharide coat is the key element in pathogenesis by Hib;

3. the vaccine is composed of this type b polysaccharide and the immune response is directed against it;

4. immunity to this type b polysaccharide reduces or eliminates Hib from vaccinated individuals; and

5. herd effects of vaccination in children could reduce the ability of the infectious agent to circulate in all age groups.

Now attention is being turned to other bacterial agents such as NTHi and other agents such as *Moraxella cattharalis* and *Streptococcus pneumonia* that cause AOM and other respiratory and invasive infections. Vaccines against *Streptococcus pneumonia* are directed against surface polysaccharides just like the Hib vaccine. However, there is an important difference. *Streptococcus pneumonia* has over 90 capsular types, and immunity stimulated by one type only slightly protects against infection or disease with other types. The currently licensed vaccine against *Streptococcus pneumonia* contains the seven capsular types most likely to cause invasive infections in children in the United States. Recent studies have shown that vaccinated individuals are colonized with, and develop infections from, capsular types of S. pneumoniae not present in the seven-valent vaccine.

31.2.5 NTHi vaccines

The potential market size of more than $1 billion per year in the United States for NTHi vaccines has stimulated intense efforts at vaccine development. However, NTHi vaccine development differs from Streptococcus pneumonia vaccine development with regard to the much greater the strain-to-strain diversity of potential vaccine target molecules in NTHi and the lack of knowledge about which type of immunity against which targets might provide vaccine protection.

Unlike Streptococcus pneumonia where one surface polysaccharide alone generates protective immunity and therefore an ideal classification criteria, there is no well-accepted classification system to use in choosing serotypic or pathogenic variants for an NTHi vaccine. Many different surface proteins are involved in NTHi immunity. Our lack of understanding regarding the NTHi antigenic variations is one important reason why we are still ignorant as to whether NTHi strains that circulate in adults are significantly different from those that circulate in children. If NTHi transmission dynamics and diversity patterns are not better understood by the time vaccine trials begin, those trials might not be informative enough to choose the best of the variety of vaccines being developed and to determine how those vaccines should be used to maximize their population benefit.

31.3 MODEL DEVELOPMENT STRATEGY

31.3.1 General approach

We used a standard deterministic compartmental model approach based on differential equations that parameterize the flow of population between compartments. Age groups, daycare attendance, colonization or disease status, and level of immunity defined the compartments. Three alternative compartment structures for the natural history of infection and immunity were used in order to insure that our conclusions were not sensitive to this aspect of compartment structure conformation. We lumped all NTHi strains together and did not define different compartments for different strains of NTHi.

We fixed population structure parameters to correspond to a developed country population. Nine other parameters were fixed by finding values that fit specified patterns of NTHi related AOM incidence and NTHi colonization prevalence. These included three contact pattern parameters, three immune effect parameters, a parameter specifying the duration of colonization episodes in the absence of immunity, a parameter specifying the fraction of colonization episodes that result in AOM in the absence of immunity, and a parameter specifying the speed with which immunity acquired after colonization is lost. Both contact patterns and naturally acquired immunity affecting transmission could contribute to the decreasing prevalence of NTHi colonization with age. The model parameters are listed in Table 31.1.

For simplicity in this first stage of model development, our model defined only a single disease state, AOM. Again for the sake of simplicity, and in the absence of clear evidence to the contrary, it treated every colonization episode as being contagious and as stimulating immunity that moves individuals up one level of immunity. Immunity rose after every colonization episode and was not acquired any differently when colonization led to AOM.

31.3.2 Infection prevalence and otitis incidence data to fit model

Given nine parameters to fit, we selected various sets of nine data points covering a range of values consistent with the literature. Four of the data points represent NTHi colonization prevalence in 1) preschool children attending daycare, 2) preschool children not in daycare, 3) school children, and 4) adults. The remaining five data points represent AOM incidence across daycare and non-daycare children for each of the first five preschool years of life. The literature can only specify these nine values imprecisely and they vary considerably from one population to another. Therefore we specified various data sets that we felt cover the plausible values for these nine data points and capture patterns to which our results might be sensitive. We then fit parameters to each of the selected data sets.

To select sets of colonization prevalence and AOM incidence for our analyses we used the literature summarized in Tables 31.2 and 31.3 along with data on the fraction of AOM where NTHi is found [3-7]. We calculated expected prevalence for children in and out of daycare assuming that 33% of preschool children attended daycare and that the relative risk of NTHi colonization comparing children in daycare to those not in daycare was 2.5. We varied age patterns of AOM incidence and colonization prevalence explicitly to cover a range of patterns we thought might affect our results. The final sets of data points used are presented in Table 31.4.

We fit the nine parameters to each combination of high and low baseline AOM incidence and colonization prevalence.

Table 31.1 Variables and parameters in the models examined

Indices

i = Categories of individuals defined by age and daycare attendance specified in Figure 31.1 ($i = 1, ..., 20$)

m = Level of immunity as in Figure 31.3 ($m = 1, ..., 4$)

k = Mixing site (general, school, daycare as in Figure 31.2) ($k = 1, 2, 3$)

Variables from Figure 31.3

C_{iam} = Recently infected population in group i at immunity level m

C_{ibm} = Infected population in group i at immunity level m that will either cure their infection or go on to disease

D_{im} = Population in group i with AOM at immunity level m

C_{icm} = Population in group i recovered from AOM but still infective at immunity level m

S_{im} = Susceptible population in group i at immunity level m

Immunity Parameters *

θ = One minus the fraction by which susceptibility is reduced after an infection

γ = One minus the fraction by which the fraction of infected individuals going on to disease is reduced by infection

χ = One minus the fraction by which contagiousness of further infection is reduced after an infection

δ = One minus the fraction by which duration of further infection is reduced after an infection

$\mu = \chi = \delta =$ One minus a single parameter reducing both duration and contagiousness by the same fraction

ω = Rate at which each level of immunity is lost

Contact Parameters

ξ_G = Effective contact rate of all age groups in the general mixing site

ξ_D = Effective contact rate of children in daycare at the daycare setting

ξ_S = Effective contact rate of school children at school

ξ_{ik} = Effective contact rate of individuals in group i at mixing site k

Table 31.1 (cont.) Variables and parameters in the models examined

Other Parameters

σ = Rate of progressing out of each C (colonization) compartment at immunity level 1

π = Fraction of infections that go on to disease at immunity level 1

Calculated Quantities

I_{im} = Size of infective population in group i with immunity level m:
$I_{im} = C_{iam} + C_{ibm} + D_{im} + C_{icm}$

N_{im} = Size of total population in group i with immunity level m:
$N_{im} = I_{im} + S_{im}$

R_{im} = Rate of infection experienced by the susceptible fraction of population i when they are at immunity level m

*Each model has only one of χ, δ, or μ

For sensitivity analysis, we fitted models to two alternatives with sharper and flatter age trends for each baseline data set. For a sharper age trend by colonization prevalence, we decreased prevalence in school children to half of baseline and that in adults to one-fourth of baseline levels. For the sharper age trend of AOM, instead of the flat 30% that we assumed for the baseline proportion of AOM caused by NTHi, we set the first year at 40% and decreased each subsequent year by 5%. For a flatter trend in colonization prevalence we set the prevalence in school children equal to that in non-daycare preschool children and the prevalence in adults twice its baseline level. For a flatter trend in AOM by age, we set the proportion of AOM caused by NTHi in the first year at 20% and increased each subsequent year by 5%.

31.3.3 Modelpopulation structure

Each compartment in the model has several dimensions. First is the division of individuals into compartments defined by demographic characteristics. Compartments and flows between them are presented in Figure 31.1.

Table 31. 2 Reported NTHi prevalence by age from studies in developed countries

	Reported NTHi prevalence by age			
Reference	**Preschool Children**	**Daycare Children**	**School children**	**Adults**
UK [8]	28% (up to 6 yrs)	----	----	----
US [9]		81% (0-5 yrs)		
US [10]	9%-11% (6-24 mos.*)		----	----
Sweden [11]	13.2% (≤6 yrs)	----	6.1% (7-15 yrs)	2.7% (≥16 yrs)
Spain [12]	----	51.1% (up to 6 yrs)	----	----
Italy [13]	----	13% (0-5 yrs)	9.4% (6-7 yrs)	----
Netherlands [14]	11% ** (3-36 mos.)	37% (3-36 mos.)	----	----

* Faden et al. [20] found that monthly NTHi prevalence peaks at 6 months of age and remains stable through 24 months of age.

** Non-daycare children

Table 31.3 Reported AOM incidences from studies in developed countries

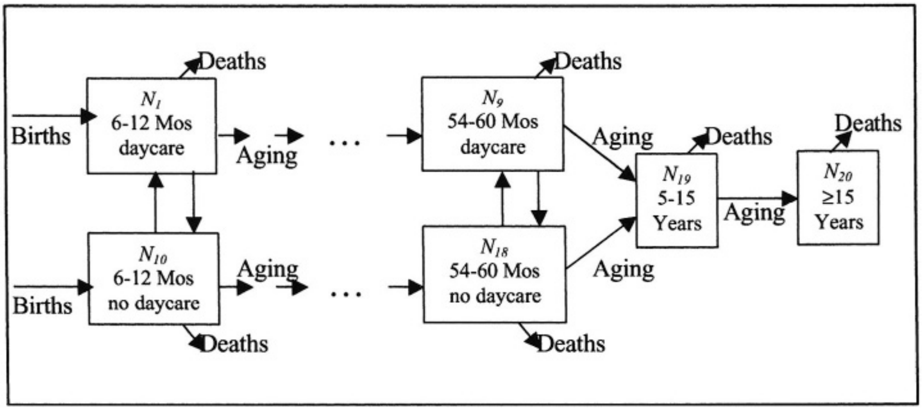
	Reported AOM incidences by age (per person-year)				
Reference	**0-1 year**	**1-2 years**	**2-3 years**	**3-4 years**	**4-5 years**
Sweden [15]	0.29	0.44	0.29	0.23	0.22
Boston [16]	1.2	1.1	0.7	0.8	0.7
Finland [17]	0.73	1.18	----	----	----

Table 31.4 Endemic NTHi colonization prevalence and AOM incidence due to NTHi to which model parameter values were fit*

	Low Values	High Values
Colonization prevalence values fitted		
Colonization prevalence ages 0-5 when in daycare	23%	51%
Colonization prevalence ages 0-5 when not in daycare	9.5%	21%
Colonization prevalence ages 6-15	7%	15%
Colonization prevalence in adults	4%	9%
AOM Incidence values fitted		
Annual NTHi AOM incidence age* <1	8.7/100	21.8/100
Annual NTHi AOM incidence age 1-2	13.2/100	33.0/100
Annual NTHi AOM incidence age 2-3	8.7/100	21.8/100
Annual NTHi AOM incidence age 3-4	6.9/100	17.2/100
Annual NTHi AOM incidence age 4-5	6.6/100	16.5/100

* All ages in the table are in years.

Figure 31.1 Structure of the 20 population groups by age and daycare attendance in all models



Because the population under age five years is the key population we seek to protect, and because colonization and disease vary strongly by age in this group, we divide the population into nine six-month age groups under age 5 starting with the second six months of life and ending at age 5. For the sake of simplicity, we model maternal immunity as completely protective for the first six months of life and therefore disregard this age group. We divide the preschool children into those that do and do not attend daycare. We define constant rates of entering and leaving daycare that result in patterns of daycare attendance corresponding to national observations[18] At age five all children enter school and leave daycare if they were in daycare. Beyond age five we divide the population into a school age group and adults. The birth, death, and daycare flows from Figure 31.1 are shown in Table 31.5. This demographic conformation has 20 categories of individuals in terms of age and daycare attendance. We use the subscript *"i"* to represent these groups in subsequent descriptions.

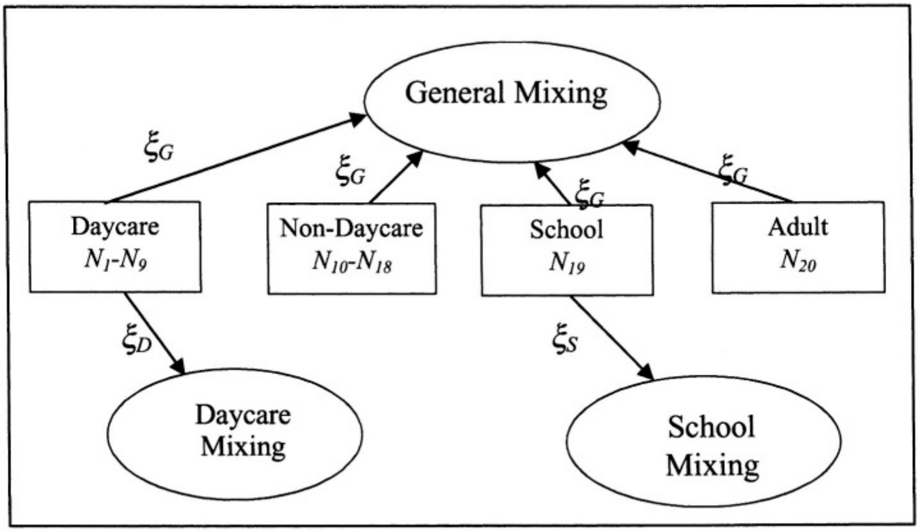
Table 31.5 Population conformation parameters

Parameter	Value
1 year death rate	0.00181
2-3 years death rate	0.00036
4-5 years death rate	0.00036
6-15 years death rate	0.00021
16 years and over death rate	0.01086
Annual birth rate into 7-12 month age group	0.00938
Rate at which children enter daycare	0.1740
Rate at which children leave daycare	0.0358
Daycare attendance at 6 months	0.0785

31.3.4 Model contact structure

To define the contact structure through which NTHi are transmitted, we break the population into four groups. These are preschool children not in daycare, preschool children in daycare, school age children, and adults. By specifying three mixing sites (daycare, school, and a general site) where the four groups mix according to structured mixing formulations [19] as indicated in Figure 31.2, and by assuming that all age groups have the same effective contact rate at a general site, we reduce the mixing parameters to three as seen in Figure 31.2. We use effective contact rate parameters, ξ, which are equivalent to the total number of contacts an individual makes per

Figure 31.2 Structure of mixing by different age groups at different sites



unit time multiplied by the transmission probability per contact given that both susceptibles and infectives have never been previously infected.

The base model has all four population groups making contact at the general site in proportional relations of 1 : 1 : 1: 1 for daycare : non-daycare : school : and adults. As a perhaps more realistic alternative for sensitivity analysis, we also used relationships 1 : 1.5 : 2 : 3.

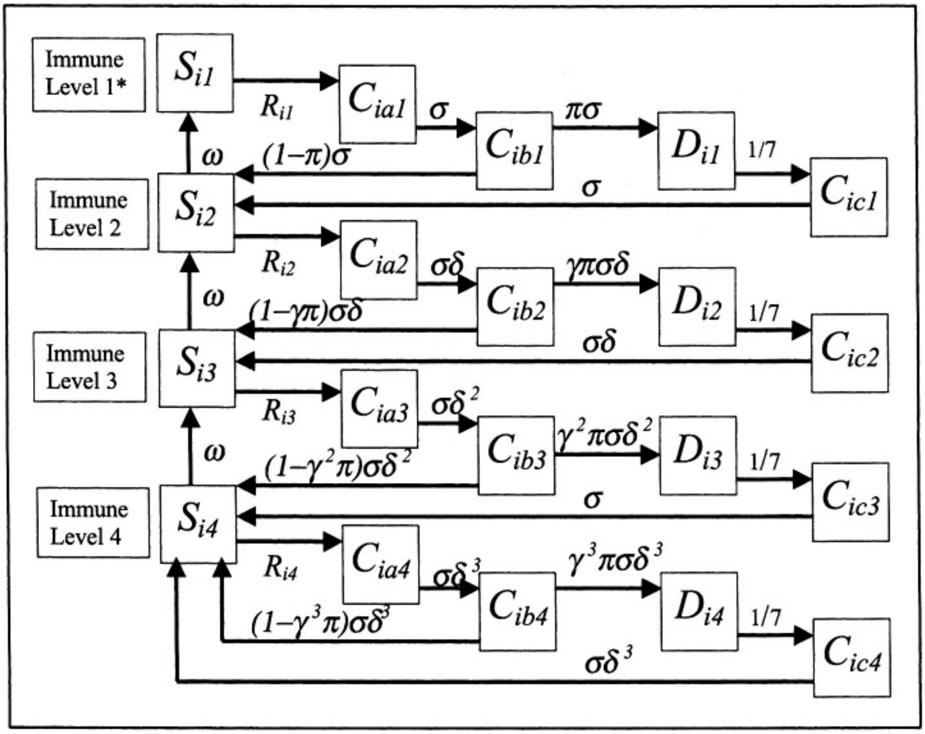
31.3.5 Natural history of infection and immunity structure

We model all NTHi strain variants jointly by modeling the acquisition of immunity to the collection of strains as a step-wise process. In our model each infection ratchets one up to the next higher level of immunity but one's immunity is continually waning so individuals are always moving down to lower immunity levels. In the real world this waning of immunity has two sources. The first is the biological loss of immunity due to immune system dynamics in the host. The second is the loss of immunity that results from changes in the circulating strains of NTHi that enable them to escape the immune responses that have been stimulated.

An alternative approach would be to model a set of cross-reacting strains whose frequencies fluctuate according to the immune pressure that affects their circulation. This would provide greater realism than our approach as well as insights about strain dynamics that might prove crucial for the design and evaluation of vaccines. However, modeling such a group of individual strains is complicated, expensive, and less readily focused on the issue we address.

The natural history of infection and immunity is modeled as flows between compartments as shown in Figure 31.3. This structure is consistent with the observations that AOM occurs within the context of nasopharyngeal infection and that ear infection is more readily eliminated by treatment than is nasopharyngeal infection. Each of the *C* compartments represents individuals with nasopharyngeal colonization. The *D* compartments represent individuals with AOM. All *C* and *D* compartments at any level are equally contagious. Only individuals without current colonization or disease can be newly colonized. The first of the four levels of these variables has no acquired immunity. Three levels of acquired immunity are modeled. An infection moves an individual only one immunity level higher. All *C* compartments have the same average duration. In some analyses, this depends upon the immunity level. We fixed the average stay in the D compartment at seven days for all immunity levels.

Figure 31.3 Structure of natural history of infection and immunity[+]

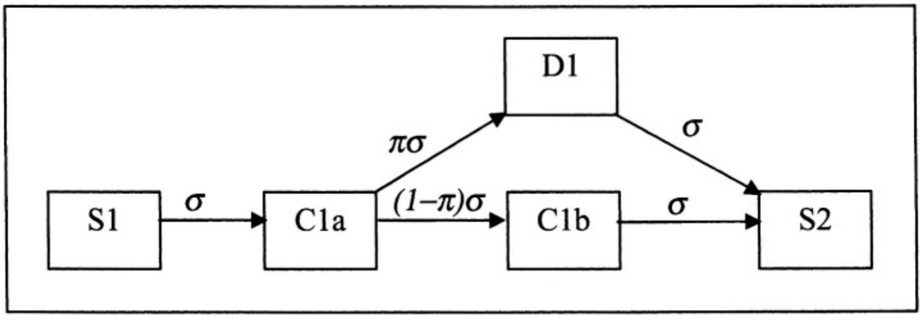


[+] All notation is defined in Table 31.1.

* All individuals who have never before been infected fall in Level 1.

In sensitivity analysis we also examined a model with eight levels of immunity following the pattern in Figure 31.3. In a further sensitivity analysis, we examined a model that eliminated one compartment for each immunity level but otherwise preserved the same parameter structure. The flow diagram of that arrangement from immunity level one to level two is shown in Figure 31.4. The susceptibility, contagiousness, and duration parameters act in the same way as for Figure 31.3, as described in the next section.

Figure 31.4 Alternative structure of natural history of infection and immunity considered in sensitivity analysis



31.3.6 Immunity effects of natural infection

Acquired immunity can decrease the fraction of colonization episodes that progress to disease (captured by the pathogenicity effect parameter γ in Table 31.1), decrease susceptibility to colonization (captured by the susceptibility effect parameter θ in Table 31.1), decrease the contagiousness of colonized individuals (captured by the contagiousness effect parameter χ in Table 31.1), and/or decrease the duration of colonization (captured by the duration effect parameter δ in Table 31.1). In all models we used both susceptibility and pathogenicity effects, as models without either of these could not fit the data. Immunity effects on either contagiousness or duration reduce the total number of transmissions from a colonized individual. Using just a contagiousness parameter fit the data better than using just a duration parameter. This also gave patterns of immunity levels by age and durations of infection that were more consistent with observations from the literature. Consequently we chose this conformation as our base model. In sensitivity analyses we used duration instead of contagiousness or a single parameter defining equal contagiousness and duration effects.

Each infection up to and including the third has the same multiplicative immune effects. All the immunity parameters have values from zero to one and we use 1 minus the values of these parameters as the measure of the amount of immunity stimulated by an infection (Table 31.1). So, the lower the value of these parameters, the greater is the immunity stimulated by colonization.

31.3.7 Rate-of-infection formulation

The rate of infection varies by age, daycare attendance, and level of immunity. The rate calculations for these categories are as follows:

$$R_{im} = \theta^{m-1} \sum_{k=1}^{3} \xi_{ik} \sum_{i=1}^{4} \sum_{m=1}^{4} \frac{\chi^{m-1} I_{im} K_{ik}}{\sum_{i=1}^{4} \sum_{m=1}^{4} N_{im} K_{ik}}$$

$$\text{where} \begin{bmatrix} K_{ik} = 1 \text{ if group } i \text{ attends site } k \\ K_{ik} = 0 \text{ if group } i \text{ does not attend site } k \end{bmatrix}$$

See Table 31.1 for definitions relevant to this equation. In this equation the susceptibility parameter θ acts on individuals equally regardless of where they are mixing. There are three mixing sites for which the force of infection an individual experiences is calculated. This force is determined by the effective contact rate ξ of the individual at each site, the fraction of individuals I that are infected at each site, and the degree of contagiousness χ of those individuals.

31.3.8 Fitting parameters to data on AOM incidence and NTHi prevalence

We use the boundary values function of Berkeley Madonna [20] to find the values of the nine parameters that in the endemic state of the model reproduce observed infection and disease frequencies. To check for identifiability problems, wide search intervals that put the estimated values at each end of those intervals are used and results are checked to see that the fitted parameter values do not change.

After fitting, we check to see that the patterns generated by our models are consistent with available data on antibody levels by age group and with data on the duration of NTHi colonization. Antibody levels peak during school age years [21] so we only accept models that similarly generate peak levels of immunity in this group. Duration of colonization measurements are less firmly established but we reject models that generate average durations of less than a week or more than six months.

31.4 RESULTS

31.4.1 Fitted parameter values

Parameter values that fit all four data combinations were found. As seen in Table 31.6, the root mean square errors were small in each case. The identifiability of the parameter values found was confirmed by observing that the same values were reached from diverse starting points. For all four data conformations the average durations of colonization and average immunity levels by age group are consistent with observations. The most important observation in Table 31.6 is that across all data conformations, of the three immunity effects in the model, the strongest effects are on susceptibility to colonization. In three of the four data conformations examined, the second strongest effects are on pathogenicity and in the fourth case contagiousness came in second place.

31.4.2 Relative effects of immunity against transmission or disease given infection

In Table 31.7 we compare the effects of a 10% decrease in the effect of infection on immunity to transmission (effect on susceptibility and contagiousness parameters) versus a 10% decrease in the effect of infection on the risk of disease given colonization (an effect on the pathogenicity parameter). Here we see that no matter what data conformation is being fit, disease incidence is much more sensitive to immunity affecting transmission than to immunity affecting disease given colonization. Even if transmission effects were slightly less than pathogenicity effects, they would deserve consideration for the design of vaccines and the measurement of vaccine effects. We see in Table 31.7, however, that transmission effects are truly dominant. This is especially true at the very youngest ages.

To understand why the effects of immunity on transmission are so important in the young ages, note that immunity effects on pathogenicity start out low in the youngest group and then progressively increase with age (Table 31.7). In contrast, immunity effects on transmission start out high, decrease, and then increase. For the models fitted to low colonization prevalence and high AOM incidence, this pattern is not evident because the rebound began very early in the second year.

The increasing effect with age of reducing colonization-induced immunity affecting pathogenicity is due to the fact that immunity effects are cumulative across sequential colonization episodes. Changing the pathogenicity parameter thus has a greater effect the higher the level of immunity. Given that older children have experienced more infections and

Table 31.6 Parameter values that fit the observations in Table 31.4

	Colonization prevalence and AOM incidence data fit*			
	H col H AOM	H col L AOM	L col H AOM	L col L AOM
Goodness of fit (root mean square error)	0.07	0.05	0.05	0.02
Duration of each level of immunity (years), $1/\omega$	3.7	4.7	3.4	9.8
Duration of each stage of colonization at the lowest level of immunity (years), $1/\sigma$	0.104	0.107	0.0613	0.0549
Probability of AOM given colonization at the lowest level of immunity, π	0.343	0.127	0.374	0.136
% decrease in AOM probability per immunity level (pathogenicity effect), $1-\gamma$	0.334	0.301	0.294	0.279
% decrease in susceptibility per immunity level, $1-\theta$	0.597	0.594	0.732	0.481
% decrease in contagiousness per immunity level, $1-\chi$	0.582	0.237	0.116	0.24
Effective contact rate per year at general site, ξ_G	173	80.1	50.3	94.4
Effective contact rate per year at daycare site, ξ_D	655	218	359	113
Effective contact rate per year at school site, ξ_S	301	68	217	61

* The four right-hand columns are intended to indicate the four combinations of the two colonization levels and two AOM incidence levels in Table 31.4.

are at a higher level of immunity, this means that any change in the pathogenicity parameter has a greater effect on older children.

The same phenomenon accounts for the late rise in AOM incidence by age when colonization-induced immunity affecting transmission is reduced. The difference with immunity effects, however, is that they have indirect effects. These indirect effects are strongest in the youngest ages because the

Table 31.7 Increase in AOM incidence when the effect of a single infection on immunity affecting transmission or pathogenicity is decreased by 10%

Data Conformation Fitted			Age				
Colonization Prevalence	**AOM Incidence**	**Immunity Type Decreased**	**0-1 year**	**1-2 years**	**2-3 years**	**3-4 years**	**4-5 years**
High	High	Pathogenicity	1.6%	3.9%	7.9%	10.9%	12.5%
		Transmission	12.0%	9.5%	11.8%	17.8%	23.4%
High	Low	Pathogenicity	1.6%	3.8%	7.6%	10.2%	13.2%
		Transmission	23.4%	14.6%	15.3%	23.6%	32.8%
Low	High	Pathogenicity	1.4%	2.9%	5.1%	6.8%	8.1%
		Transmission	15.9%	19.2%	32.6%	48.7%	62.7%
Low	Low	Pathogenicity	1.8%	3.7%	6.7%	9.0%	10.4%
		Transmission	59.7%	34.1%	33.5%	53.2%	70.3%

youngest are most susceptible. Indirect effects decrease with the acquisition of immunity and therefore decrease with age.

31.4.3 Sensitivity of transmission dominance to model form, assumptions, and parameters

To strengthen the conclusion from Table 31.7 that immunity affecting transmission has far greater effects on the risk of AOM in young children than does immunity against AOM given colonization, we conducted extensive sensitivity analyses. These are presented in Table 31.8. We generated tables like Table 31.7 for each of the sensitivity analyses. To present the results of these compactly, we divided each relative increase from transmission parameters by the relative increase from the pathogenicity parameter and then we averaged these ratios across the four different data sets. This reduces the eight numbers that could appear in each column to a single summary number.

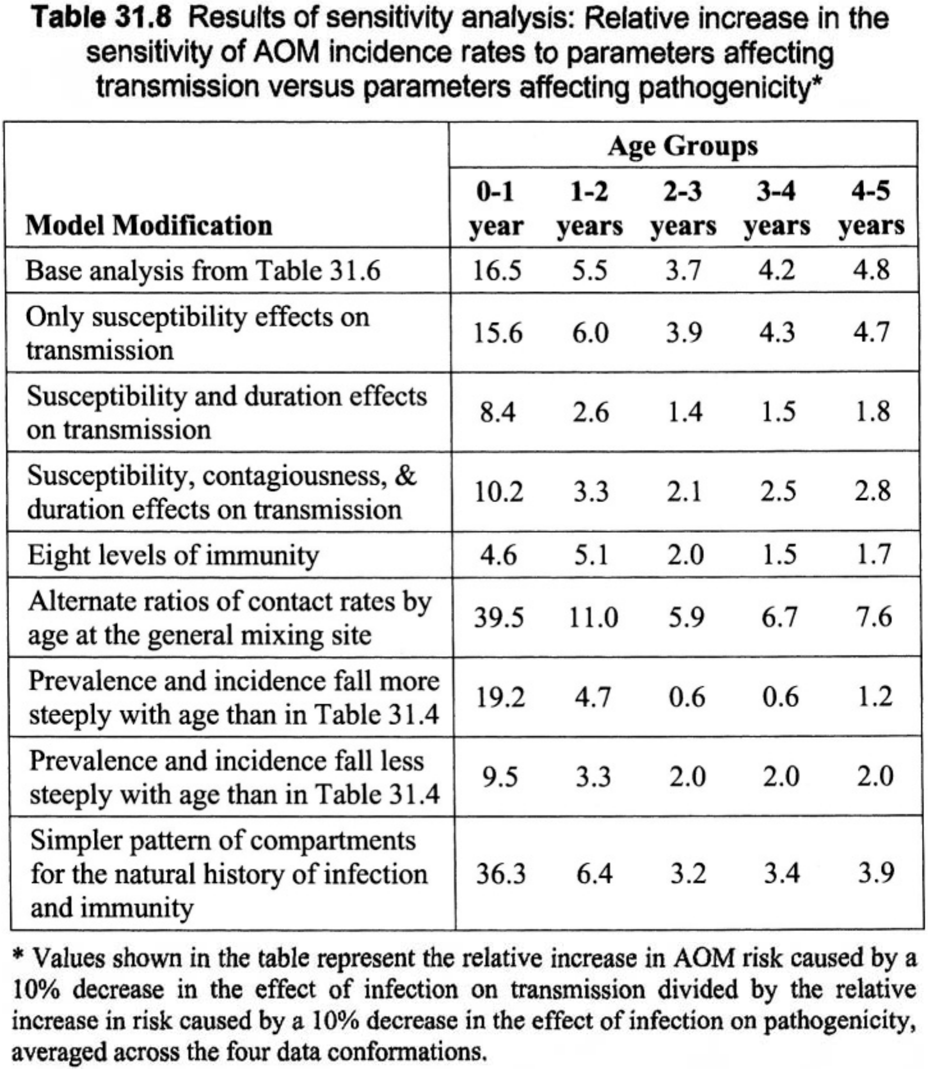


Table 31.8 Results of sensitivity analysis: Relative increase in the sensitivity of AOM incidence rates to parameters affecting transmission versus parameters affecting pathogenicity*

Model Modification	Age Groups				
	0-1 year	1-2 years	2-3 years	3-4 years	4-5 years
Base analysis from Table 31.6	16.5	5.5	3.7	4.2	4.8
Only susceptibility effects on transmission	15.6	6.0	3.9	4.3	4.7
Susceptibility and duration effects on transmission	8.4	2.6	1.4	1.5	1.8
Susceptibility, contagiousness, & duration effects on transmission	10.2	3.3	2.1	2.5	2.8
Eight levels of immunity	4.6	5.1	2.0	1.5	1.7
Alternate ratios of contact rates by age at the general mixing site	39.5	11.0	5.9	6.7	7.6
Prevalence and incidence fall more steeply with age than in Table 31.4	19.2	4.7	0.6	0.6	1.2
Prevalence and incidence fall less steeply with age than in Table 31.4	9.5	3.3	2.0	2.0	2.0
Simpler pattern of compartments for the natural history of infection and immunity	36.3	6.4	3.2	3.4	3.9

* Values shown in the table represent the relative increase in AOM risk caused by a 10% decrease in the effect of infection on transmission divided by the relative increase in risk caused by a 10% decrease in the effect of infection on pathogenicity, averaged across the four data conformations.

First we examined the effects of different parameterizations of immunity effects on transmission. Models that lacked immunity affecting susceptibility to colonization or immunity affecting pathogenicity would not fit the data from Table 31.4 well in any of the four combinations. Thus for sensitivity analyses we always kept a susceptibility effect and changed contagiousness and duration effects. The baseline has susceptibility and contagiousness effects but no duration effect (first row in Table 31.8). The row after the baseline in Table 31.8 has no duration or contagiousness effects. Then a model with no contagiousness effect but with a duration effect is examined. Finally we examined a model where contagiousness and duration effects were entered but with a single parameter (fourth row of Table 31.8).

One might speculate that a reason our analysis showed immunity effects on transmission to be greater than effects on pathogenicity is that we fit two parameters affecting transmission and only one affecting pathogenicity. We see no logic regarding either transmission dynamics or the way we fit our parameters to support such a conclusion. Indeed when transmission is only affected by a single transmission parameter, namely only a susceptibility parameter, the effects are not much different from those in the base model. Likewise, the model with a duration effect parameter instead of a contagiousness effect parameter still shows a dominance of immunity affecting transmission. The effect is less than with the base model, but still impressive. The same can be said for the model where a single parameter was used to equally affect immunity to both contagiousness and duration of colonization.

Our choice of four levels of immunity for our model was initially due to the fact that we did not see improvement of fit to higher numbers of levels. But using only four levels of immunity to a single agent to capture the effects in the real world of a diversity of cross-reacting agents is simplification whose effects on the issue we are addressing is unclear. When we examined a model with eight levels of immunity, our conclusions were unchanged (fifth row of Table 31.8). Even though the relative increases in AOM incidence were less than with four immunity levels, they were still impressive.

In addition we examined a model with perhaps a more realistic assumption about general mixing as described in Section 31.2 under “Model Contact Patterns”. With the new relationship the immune effects on transmission are even more dominant (sixth row of Table 31.8).

The one analysis where in some age groups changing immunity to pathogenicity affected AOM incidence more than immunity to transmission was when we used steeper age curves of NTHi prevalence and AOM incidence (seventh row of Table 31.8). In Table 31.8 we see that this is true for ages 3 and 4 as the ratio of transmission effects to pathogenicity effects is less than one. In no way, however, should this pattern be interpreted as representing a greater effect of immunity to pathogenicity. When the immunity to transmission is decreased in this situation, its effects on increasing circulation of NTHi are so great that much higher stages of immunity are reached in the third and fourth years of life. Referring back to our explanation of the U-shaped curves of transmission effects, fitting our model to steeper age drops in colonization and in AOM incidence generates a steeper drop in the indirect effects of immunity on transmission and a faster rise in the direct pathogenicity effects on AOM.

Fitting our model to flatter age drops in colonization and AOM (eighth row of Table 31.8) has the opposite effect. The indirect effects of immunity affecting transmission are less at the start but fall more slowly while the effects of immunity on pathogenicity rise more slowly. Overall whether the age relationships are steeper or flatter, the dominance of immunity affecting transmission remains impressive.

Finally, we examined a different compartmental flow structure for the natural history of infection and immunity. This is the flow structure illustrated in Figure 31.4. Use of the alternative structure had little effect on our conclusions (last row of Table 31.8).

31.4.4 Age groups sustaining transmission

Another question to address is which groups can sustain circulation of NTHi. We isolated preschool children, school children, and adults in our models and found that in the absence of contact with other age groups, each group can sustain the circulation of NTHi by themselves. This finding was robust to the NTHi incidence and AOM prevalence conformations to which model parameters were fit. The finding was similarly unaffected by whether immunity affected contagiousness or duration of colonization. This is certainly in marked contrast to the observation that Hib immunization of only young children has been effective in dramatically reducing Hib infection and disease levels in all age groups.

31.5 DISCUSSION

31.5.1 Major conclusions

Descriptions of NTHi infection patterns and NTHi transmission dynamics are sketchy and imprecise. But our analysis of the NTHi transmission system shows that infection and disease observations are sufficient to conclude that acquired immunity to transmission will be an essential element of a successful vaccine against this agent. Across a broad range of data conformations that are consistent with developed country situations, and across a broad range of model conformations, we observed that AOM incidence was more sensitive to immunity affecting transmission than to immunity affecting the risk of AOM given NTHi colonization.

Our estimates of relationships between pathogenicity and transmission effects probably underestimate the dominance of transmission effects because of the way we handle NTHi diversity. We tried to capture the effects of diversity with four immune levels. We treated all age groups as having identical immunity effects within each immunity level and identical rates of waning immunity. Most likely there is considerable diversity in NTHi transmissibility, pathogenicity, and the strength of immunity stimulated. A strain with greater invasive and pathogenicity potential is likely to stimulate a stronger immune response than a strain with less invasive and pathogenic potential. That means that immunity will be acquired first to the most pathogenic strains and that NTHi colonizing adults will have less pathogenic potential than NTHi colonizing children. Thus if we were to model just the most pathogenic strains, we would have a steeper age curve of colonization and our estimates of immunity against transmission would be greater.

Our model-based conclusions are consistent with studies showing that individuals with nasopharyngeal anti-P6 sIgA antibody obtained during NTHi colonization were protected from re-colonization [22-24]. Also the observation that NTHi have evolved two different anti-IgA proteases [25] strongly argues for immunity against colonization being a dominant force driving the evolution of NTHi.

In our models each age group in isolation can sustain circulation if NTHi. This indicates that indirect effects of vaccinating preschool children will not have the dramatic effects on NTHi prevalence in other age groups that they had on Hib. However, we did demonstrate strong indirect effects on the youngest age groups that most urgently need protection against infection. This is encouraging, as it is always more difficult to stimulate adequate immunity with a vaccination program in the youngest age groups.

31.5.2 Solidifying and extending results by melding data and models more productively

While strong, our conclusion about the dominance of immunity affecting transmission over immunity affecting pathogenicity of NTHi is tentative. In order for pathogenicity parameters to have greater effects than transmission parameters, the prevalence relationships by age would have to be flatter and the fraction of AOM due to NTHi would have to fall more steeply by age than seems likely. Better data on NTHi colonization prevalence by age across the entire age range as well as better data on the age distribution of AOM caused by NTHi is needed to solidify this conclusion. Studies that gather both types of data from the same population are needed. Our conclusions are solid enough, however, to mandate that in any trial of NTHi vaccines, assessing effects on transmission should be a major objective.

One way to solidify the theory we have presented regarding immunity effects is to use the theory to make predictions about the outcomes of novel observations. When such predictions made by two theories conflict, a basis for choosing between theories exists. One such area that deserves exploration relates to the extent of fluctuations in endemic NTHi colonization prevalence. The stronger the effects of immunity on circulation, the greater the fluctuations should be in endemic prevalence. If there are diverse strains with little cross-reactive immunity, fluctuations will have to be specified by type to evaluate such a prediction.

Besides solidifying the general conclusion about transmission effects, we should pursue better definition of how effects on susceptibility, contagiousness and colonization duration contribute to these effects. This is necessary to predict the population effects of vaccination programs. But, as we argued in the background section, direct measurement of these effects with very few model assumptions is infeasible. Making immunity and vaccine effects reliably identifiable from collectible data should be a goal of future modeling efforts. The approach taken in this paper needs to be refined so that parameter estimation procedures can be developed to provide valid statistical inference for these parameters using data collected in carefully planned studies that maximize power. Model explorations in this direction should not confine themselves to consideration of only pathogen culture results. Immunity information is likely to add powerfully to estimation potential. IgA measurement from throat washes might be particularly useful. Also, molecular studies that can indicate when transmission could or could not have occurred between individuals would be helpful. However, such studies will not be able to resolve immunity effects unless they are designed and analyzed in the context of transmission system models.

The use of data on the dynamic fluctuations of NTHi prevalence seems likely to provide more powerful parameter estimates than achieved by the procedures in this paper, which assume observed prevalence values represent endemic prevalence levels. Molecular distinction between strains would add value to such studies by distinguishing infections in families or daycare centers that come from within or from outside of those families or daycare centers. Even with models like those in this paper that do not distinguish different strains, the strain data could be very useful for constraining the range of parameter estimates. Models with different strains might be even more helpful in extracting knowledge from such data.

31.5.3 Models and strains of NTHi

Our understanding of NTHi immune effects would be improved by ascertaining the molecular determinants of cross-reactive immunity and of pathogenicity. Use of transmission models to analyze epidemiological data may be essential to achieve these ends for two reasons. First, a model that conforms better to reality than statistical models that assume away transmission may be essential to perceive patterns of cross-reactive immunity and pathogenicity [26]. Second, even with fine molecular distinctions, model assumptions may be necessary to get reasonable power because of the frequency and duration of swabbing needed to directly measure these effects.

A first step in this direction should be studies and models that assess strain distribution by age. There is a great tendency to confine epidemiological data on NTHi colonization to the age group suffering most from infection. But if the effects of immunity are to be understood on the transmission dynamics of NTHi, data from across the age range will be essential. It seems remarkable so little data exist on differences in NTHi molecular patterns by age. We have such data for *Streptococcus pneumoniae* mainly because of the well-established ties between immunity and infection risks to dominant surface antigens. In NTHi where surface antigens and genome patterns are more diverse, the picture can be confusing. But a simple distinction between molecular patterns by age would go a long way to helping fit models that include multiple strains.

31.5.4 The need for further transmission system analysis

We see three reasons why the analyses we present here need to be extended. First, further assessment of the sensitivity of our conclusions to model form and parameter values is needed to solidify our conclusions. Second, the population consequences of vaccines stimulating different patterns of immunity should be explored using our models to help direct vaccine research into the most useful directions. Finally, and most importantly, the design of vaccine trials needs to be explored using discrete individual stochastic models based on our deterministic compartmental models to insure that the effects of vaccines on transmission are fully and efficiently assessed.

Further sensitivity analyses We approximate reality roughly by handling the phenomenon of repeated NTHi colonization using four levels of immune response. We argued in the previous section that models with more realistic strain diversity would only amplify the effects, leading to the conclusion that immunity against transmission should be the primary objective of vaccine development. But the NTHi transmission system is complex, and surprises that invalidate predictions in complex systems are not uncommon. Thus our logic that models with multiple strains will even more strongly support our conclusions needs to be assessed by building and analyzing such models.

Exploration of potential vaccine effects How sensitivity to immunity parameters translates into vaccination effects cannot be deduced accurately without an appropriate model analysis. Many issues such as the number of repeat vaccinations required, how vaccine effects vary by age, how effectively young children can be reached by a vaccination program, and how vaccine effects differ from natural infection effects will affect the overall effects of vaccine induced immune effects on transmission or pathogenicity. In general, however, it seems that adding realism in any of these dimensions will only increase the importance of indirect effects from vaccine effects on transmission.

Perhaps the most important issue regarding vaccine effects that deserves exploration is how vaccination programs could affect the mix of agents with different pathogenic potential. How important is it to insure that vaccines preferentially target the most pathogenic organisms? What would be the consequences of decreasing the circulation of less pathogenic strains?

Exploration of vaccine trial designs It is evident from our analysis that vaccine trial designs that fail to capture effects on transmission would be a mistake. To capture transmission effects, trials must be conducted in transmission units like families or daycare centers. The classical design that studies only individuals who are randomly assigned vaccine or placebo can only detect susceptibility effects. Many questions deserve exploration regarding the design of trials that will detect transmission effects. What size of transmission unit maximizes the efficiency of trial designs? What age range of children to be vaccinated maximizes trial efficiency? How do different randomization schemes affect the predictive value of parameters that can be estimated in a trial? What frequency of swabs in a trial maximizes the efficiency of the trial? Is it worthwhile to study effects of infant and toddler vaccination on older sibling and parent colonization rates?

The models we have presented could be used as a base for transition to the individual event history models needed to answer these questions. The Model Transition Sensitivity Analysis (MTSA) strategy is indicated for this [27]. Biomedware, Inc. has been supported by the National Institutes of Health to develop software that will effect transition at the click of a mouse from the deterministic compartmental models we used here to stochastic individual models needed to assess trial design.

31.5.5 Promoting involvement of systems scientists in transmission analyses

No single body of scientists is ready to undertake the type of tasks we have listed above. Acquired skills in visualizing the consequences of changes in system conformation and parameters is needed to identify all of the sensitivity analyses needed to solidify model analysis conclusions. Likewise skill in designing computer experiments of the issues outlined above is needed if the issues that need resolution are to be resolved efficiently. Perhaps most importantly, the design of analytic methods to estimate system parameters from data is a crucial issue to which diverse system scientists need to contribute. Collaborations between systems scientists and epidemiologists seem essential. To promote such collaboration, epidemiology needs to develop a specialty area of systems science, and engineering needs to develop a specialty area in infection transmission system analysis. Broader conference and journal forums need to extend the forum of this book if such collaborations are to flourish.

References

[1] Poolman, J.T., et al. (2000). Developing a nontypeable Haemophilus influenzae (NTHi) vaccine. *Vaccine,* 19, S108-S115.

[2] Barbour, M.M.-W., R.T. Coles, C. Crook, and D.W.M. Moxon. (1995). The impact of conjugate vaccine on carriage of Haemophilus influenzae type b. *Journal of Infectious Diseases,* 171, 93-98.

[3] Eskola, J., et al. (2001). Efficacy of a pneumococcal conjugate vaccine against acute otitis media. *New England Journal of Medicine,* 344, 403-409.

[4] Heikkinen, T., M. Thint, and T. Chonmaitree (1999). Prevalence of various respiratory viruses in the middle ear during acute otitis media. *New England Journal of Medicine,* 340, 260-264.

[5] Bluestone, C.D. (1982). Otitis media in children: to treat or not to treat? *NewEngland Journal of Medicine,* 306, 1399-1404.

[6] Teele, D.W., J.O. Klein, and B.A. Rosner (1980). Epidemiology of otitis media in children. *Annals of Otology, Rhinology, & Laryngology - Supplement,* 89, 5-6.

[7] Del Beccaro, M.A., et al. (1992). Bacteriology of acute otitis media: a new perspective. *Journal of Pediatrics,* 120, 81-84.

[8] Howard, A.J., K.T. Dunkin, and G.W. Millar (1988). Nasopharyngeal carriage and antibiotic resistance of Haemophilus influenzae in healthy children. *Epidemiology & Infection,* 100, 193-203.

[9] St. Sauver, J.L., C. Marrs, B. Foxman, P. Somsel, R. Madera, and J.R. Gilsdorf. (2000). Relationship of otitis media risk factors to carriage of multiple strains of *H. influenzae* and *S. pneumoniae. Emerging Infectious Diseases,* 6, 622-630.

[10] Faden, H., L. Duffy, A. Williams, D.A. Krystofik, and J. Wolf (1996). Epidemiology of nasopharyngeal colonization with nontypeable Haemophilus influenzae in the first two years of life. *Acta Oto-Laryngologica - Supplement,* 523, 128-129.

[11] Gunnarsson, R.K., S.E. Holm, and M. Soderstrom (2000). The prevalence of potentially pathogenic bacteria in nasopharyngeal samples from individuals with a long-standing cough-clinical value of a nasopharyngeal sample. *Family Practice,* 17, 150-155.

[12] Fontanals, D., et al. (2000). Prevalence of Haemophilius influenzae carriers in the Catalan preschool population. Working Group on Invasive Disease Caused by Haemophilus influenzae. *European Journal of Clinical Microbiology & Infectious Diseases,* 19, 301-304.

[13] Principi, N., P. Marchisio, G.C. Schito, and S. Mannelli (1999). Risk factors for carriage of respiratory pathogens in the nasopharynx of healthy children. Ascanius Project Collaborative Group. *Pediatric Infectious Disease Journal,* 18,517-523.

[14] Peerbooms, P.E., M.N. Stokman, D.A. van Benthem, B.H. van Weert, M.L. Bruisten, S.M. van Belkum, and R.A. Coutinho (2002). Nasopharyngeal carriage of potential bacterial pathogens related to day care attendance, with special reference to the molecular epidemiology of Haemophilus influenzae. *Journal of Clinical Microbiology,* 40, 2832-2836.

[15] Lundgren, K. and L. Ingvarsson (1983). Epidemiology of acute otitis media in children. *Scandinavian Journal of Infectious Diseases - Supplementum,* 39, 19-25.

[16] Teele, D.W., J.O. Klein, and B. Rosner (1989). Epidemiology of otitis media during the first seven years of life in children in greater Boston: a prospective, cohort study. *Journal ofInfectious Diseases,* 160, 83-94.

[17] Alho, O.P., M. Koivu, M. Sorri, and P. Rantakallio (1991). The occurrence of acute otitis media in infants. A life-table analysis. *International Journal of Pediatric Otorhinolaryngology,* 21, 7-14.

[18] U.S. Bureau of the Census (1998). *Childcare Arrangements for Preschoolers by Family Characteristics: Fall 1995 Percentages.* http://www.census.gov/population/socdemo/child/ppl-138/tab01b.txt.

[19] Jacquez, J.A., C.P. Simon, and J.S. Koopman (1989). *Structured Mixing: Heterogeneous Mixing by the Definition ofActivity Groups.* Springer-Verlag Lecture Notes in Biomathematics.

[20] Macey, R. and G. Oster (2003). Berkeley Madonna – Modeling and Analysis of Dynamic Systems. http://www.berkeleymadonna.com/.

[21] Yamanaka, N. and H. Faden (1993). Antibody response to outer membrane protein of nontypeable Haemophilus influenzae in otitis-prone children. *Journal ofPediatrics,* 122, 212-218.

[22] Harabuchi, Y., et al. (1994). Nasopharyngeal colonization with nontypeable Haemophilus influenzae and recurrent otitis media. Tonawanda/Williamsville Pediatrics. *Journal of Infectious Diseases,* 170, 862-866.

[23] Yamanaka, N. and H. Faden (1993). Local antibody response to P6 of nontypeable Haemophilus influenzae in otitis-prone and normal children. *Acta Oto-Laryngologica,* 113, 524-529.

[24] Williams, R. and R. Gibbons (1972). Inhibition of bacterial adherence by secretory immunoglobulin A: a mechanism of antigen disposal. *Science,* 177, 697-699.

[25] Rao, V.K., G.P. Krasan, D.R. Hendrixson, S. Dawid, and J.W. St. Geme, 3rd (1999). Molecular determinants of the pathogenesis of disease due to non-typeable Haemophilus influenzae. In *FEMS Microbiology Reviews.*

[26] Koopman, J.S., I.M. Longini, J.A. Jacquez, C.P Simon, D. Ostrow, W.R. Martin, and D.M. Woodcock (1991). Assessing risk factors for transmission. *American Journal of Epidemiology,* 133, 1199-1209.

[27] Koopman, J.S., S.E. Chick, C.P. Riolo, C.P. Simon, and G. Jacquez (2002). Stochastic effects of disseminating versus local infection transmission. *MathematicalBiosciences,* 180, 49-71.

32 THE IMPACT OF NOVEL TREATMENTS ON Aß BURDEN IN ALZHEIMER'S DISEASE: INSIGHTS FROM A MATHEMATICAL MODEL

David L. Craft[1], Lawrence M. Wein[2] and Dennis J. Selkoe[3]

[1] Operations Research Center
Massachusetts Institute of Technology
Cambridge, MA 02139

[2] Graduate School of Business
Stanford University
Stanford, CA 94306

[3] Center for Neurologic Diseases
Harvard Medical School
Brigham and Women's Hospital
Boston, MA 02115

SUMMARY

Motivated by recent therapeutic initiatives for Alzheimer's disease, we developed a mathematical model of the accumulation of amyloid, β-protein (Aβ) in the brain. The model incorporates the production and clearance of Aβ monomers, and the elongation and fragmentation of polymers by monomer aggregation and break-off, respectively. Our analysis suggests that Aβ dynamics are dictated by a single unitless measure referred to as the polymerization ratio, which is the product of the production and elongation rates divided by the product of the clearance and fragmentation rates. Cerebral Aβ burden (i.e., the total number of Aβ molecules, whether they exist as monomers or polymers) attains a finite steady-state level if this ratio is less than one, and undergoes sustained growth if this ratio is greater than one. The highly nonlinear relationship between the polymerization ratio and the steady-state Aβ burden implies that a modest reduction in the polymerization ratio achieves a significant decrease in the Aβ burden. Our model also predicts that after initiation or discontinuation of treatment, it may take months to reach a new steady-state Aβ burden. Taken together, our findings suggest that the research community should focus on developing agents that provide a modest reduction of the polymerization ratio while avoiding long-term toxicity. Finally, our model can be used to indirectly estimate several crucial parameters that are difficult to measure directly: the production rate, the fragmentation rate and the strength of treatment.

KEY WORDS

Alzheimer's disease, Smoluchowski equation

32.1 INTRODUCTION

More than four million people in the U.S. suffer from Alzheimer's disease (AD), and the prevalence among people over 85 years old is about 50%. The U.S. spends roughly 100 billion dollars annually to treat AD. The cerebral build-up of amyloid β-protein (Aβ) and its aggregation into oligomers and polymers are key components of the pathogenesis of Alzheimer's disease (AD) [1]. Recent studies suggest that Aβ accumulation plays an important role in AD neurodegeneration [2, 3]. Aβ is produced when two enzymes, β-secretase and γ-secretase, sequentially cut the amyloid precursor protein (APP), which is expressed in the brain [1]. Consequently, a variety of Aβ treatment strategies are being aggressively pursued, including the enhancement of Aβ clearance via immunization with Aβ [4-6], and the reduction of Aβ production by inhibiting γ-secretase [7, 8]. The kinetics of Aβ production, aggregation (polymerization), disaggregation and clearance, and the effect of these novel treatments on the time-dependent variation of Aβ levels are not well understood. Although there exist several mathematical models that focus on either fibrillogenesis [9-16] or plaque formation [17-19], these studies neither consider the impact of treatment nor allow for a continuously renewable source of Aβ from APP molecules. Recently, a mathematical analysis was undertaken of an agent that blocks amyloid formation by capping polymer ends [20]. Here we continue in this vein by formulating and analyzing a parsimonious mathematical model that tracks the dynamics of Aβ production, polymer elongation, fragmentation and clearance during the course of treatment. Simple formulas and numerical results are presented that elucidate the impact of treatment on Aβ burden.

32.1.1 Mathematical Model

In AD brain tissue, the level of extracellular (and perhaps intracellular) Aβ (particularly $A\beta_{42}$, the 42 amino-acid form of Aβ) increases over time, which gives rise to polymerization. Some of these polymers cluster into light microscopically-visible particles consisting of $A\beta_{42}$, which can accumulate to create diffuse (amorphous) plaques, the apparent initial neuropathological lesion of AD. Further, some polymers containing $A\beta_{42}$ and/or $A\beta_{40}$ can eventually fold into long filamentous assemblies called amyloid fibrils. Clumped masses of these are referred to as amyloid (senile) plaques. Our mathematical model idealizes this process in several important ways. First, we only consider extracellular Aβ, which is likely to be in a dynamic equilibrium with intracellular Aβ. We do not distinguish between $A\beta_{40}$ and $A\beta_{42}$, or between soluble and insoluble Aβ, because there is insufficient information in the current literature about the precise dynamic changes in these species over time in the preclinical and clinical phases of AD. Finally, because our primary goal is to understand the impact of treatment on total cerebral Aβ burden (as opposed to the number and size of plaques), we restrict our attention to Aβ polymerization, and do not attempt to capture the downstream processes of fibrillization or plaque formation *in vivo*. However, later in the chapter we discuss how this model might be interpreted in terms of – and generalized to incorporate – plaque formation.

Our model is an infinite system of nonlinear differential equations that tracks the temporal evolution of the concentration of extracellular Aβ *i*-mers, for *i* = 1, 2,.... It is related to a large class of related models in polymer chemistry [21] and actin dynamics [22], and to the Smoluchowski equation [23-24], which has been used for nearly a century to study aggregation processes such as galaxy formation, crystallization, and cloud formation [25-26]. While the polymerization and depolymerization processes in some of these works are more general than the AD-specific processes we assume here, the novelty of our model is to simultaneously incorporate fragmentation, a source (Aβ monomer production), and a sink (Aβ monomer loss).

For i = 1, 2,..., we let $c_i(t)$ denote the concentration of Aβ *i*-mers at time *t;* i.e., *i* = 1 denotes monomers, *i* = 2 denotes dimers, etc. The differential equations dictate the time rate of change of $c_i(t)$, denoted by $\dot{c}_i(t)$, and are given by

$$\underbrace{\dot{c}_1(t)}_{\text{monomers}} = \underbrace{p}_{\text{production}} + \underbrace{f\sum_{i=2}^{\infty} c_i(t)}_{\text{all fragmentations}} - \underbrace{ec_1(t)\Big(2c_1(t) + \sum_{i=2}^{\infty} c_i(t)\Big)}_{\text{all elongations}} - \underbrace{lc_1(t)}_{\substack{\text{loss (e.g.,by}\\ \text{degradation)}}} , \tag{1}$$

$$\underbrace{\dot{c}_2(t)}_{\text{dimer}} = \underbrace{ec_1(t)c_1(t)}_{\substack{\text{elongation}\\ \text{of monomers}}} + \underbrace{fc_3(t)}_{\substack{\text{fragmentation}\\ \text{of 3-mers}}} - \underbrace{ec_1(t)c_2(t)}_{\substack{\text{elongation}\\ \text{of dimers}}} - \underbrace{\frac{f}{2}c_2(t)}_{\substack{\text{fragmentation}\\ \text{of dimers}}} , \tag{2}$$

and for $i \geq 3$,

$$\underbrace{\dot{c}_i(t)}_{\text{i-mers}} = \underbrace{ec_1(\mathrm{t})c_{i-1}(t)}_{\substack{\text{elongation}\\ \text{of } (i-1)\text{-mers}}} + \underbrace{fc_{i+1}(t)}_{\substack{\text{fragmentation}\\ \text{of } (i+1)\text{-mers}}} - \underbrace{ec_1(t)c_i(t)}_{\substack{\text{elongation}\\ \text{of } i\text{-mers}}} - \underbrace{fc_i(t)}_{\substack{\text{fragmentation}\\ \text{of } i\text{-mers}}} . \tag{3}$$

The model captures four processes - production, elongation, fragmentation, and loss (i.e., clearance). For ease of reference the corresponding mnemonic parameters are displayed in Table 32.1. In equation (1), Aβ monomers are produced via cleavage of the APP at a constant rate *p,* and are lost at time *t* at rate $lc_1(t)$; that is, Aβ monomers live for l^{1} time units on average before being cleared by cell internalization (e.g., microglial ingestion), degradation by proteases or removal from the brain via the circulation.

Table 32.1 Some theoretical parameter values for the model, which yield a polymerization ratio of $r = pe/(fl) = 0.84$.

Parameter	Description	Value
p	Production rate of Aβ monomers	1.02×10^{-11} M sec^{-1}
l	Loss rate constant for monomers	1 hr^{-1}
e	Elongation rate constant (only monomer addition)	90 M^{-1} sec^{-1}
f	Fragmentation rate constant (only monomer break-off)	3.93×10^{-6} sec^{-1}

Equations (1)-(3) assume that elongation occurs by monomer addition with elongation constant *e,* so that $ec_1(t)c_{i-1}(t)$ is the rate at which (*i*-1)-mers elongate to *i*-mers for $i \geq 1$. Similarly, to limit the number of model parameters required, we assume that fragmentation occurs only by monomer break-offs, with fragmentation rate constant *f,* so the $fc_i(t)$ is the rate at which an *i*-mer fragments into an (*i*-1)-mer and a monomer. Hence, for $i \geq 2$, the concentrations of *i*-mers in equations (2)-(3) increase due to elongation of (*i*-1)-mers and fragmentation of (*i*+1)-mers, and decrease due to elongation and fragmentation of *i*-mers. The factor 2 in equation (1) arises because the elongation to a dimer requires two monomers, and the term $f/2$ in equation (2) stems from the fact that a dimer can only fragment in one location, while larger *i*-mers possess two potential fragmentation sites. Fragmentation of dimers occurs at rate $f/2$ and creates two monomers, and hence the factors 2 and 1/2 cancel each other out in the fc_2 term in equation (1).

An alternate modeling approach is to allow direct clearance of monomers off of *i*-mers (e.g., to represent microglial ingestion of polymers), rather than requiring a two-step procedure of monomer fragmentation followed by monomer clearance. This alternative would result in the omission of the "all fragmentations" term in equation (1) and the re-interpretation of *f* as an ingestion rate, but would not change the qualitative nature of our results: it would mainly change the precise conditions for a steady-state solution and the relative values of steady-state monomer and dimer concentrations, and slightly hasten the approach to a post-treatment steady-state. Finally, while

Aβ treatment parameters are not explicitly incorporated into equations (1)-(3), treatment can be included in the model, as explained later.

32.2 RESULTS

32.2.1 Steady-state solution

Our analysis begins by seeking conditions for the existence of a steady-state (i.e., equilibrium) solution, c_i, to equations (1)-(3); our analytical approach is standard and the derivations are omitted. To describe our results in a concise manner, we define the polmerization ratio

$$r = \frac{pe}{fl}, \tag{4}$$

which is a unitless quantity that incorporates the four model parameters. Setting the left side of equations (1)-(3) (i.e., the rates of change of i-mer concentration) to zero and solving for c_i reveals that there are two regimes: a steady-state (or subcritical) regime where $r < 1$ and a supercritical regime where $r > 1$. In the former case, the Aβ levels settle into a steady state where

$$c_1 = \frac{p}{l}, \text{ and } c_i = 2\frac{p}{l}r^{i-1} \text{ for } i \geq 2. \tag{5}$$

Hence, i-mer concentrations decay geometrically for $i \geq 2$, and monomers are more prevalent than dimers if $r < 0.5$, and less prevalent than dimers if $r > 0.5$. In contrast, if $r > 1$ then the Aβ burden grows indefinitely. Because the Aβ burden appears to remain relatively stable over time in symptomatic AD patients [27-29], we focus primarily on the steady-state regime, and defer until later a discussion of the supercritical regime.

Equation (5) implies that the steady-state value for the total Aβ concentration (i.e., total number of Aβ molecules in the system, whether they exist as monomers or i-mers), which is denoted by $c = \sum_{i=1}^{\infty} ic_i$ and referred to as the Aβ burden, is given by

$$c = \frac{p}{l}\left(\frac{2}{(1-r)^2} - 1\right) \tag{6}$$

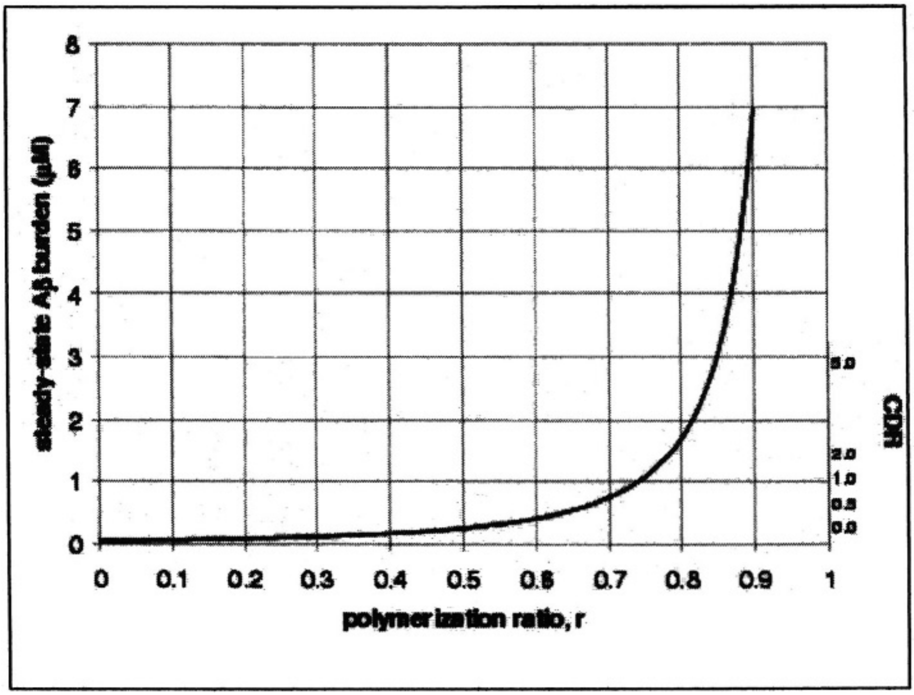
Notice that the total Aβ burden c approaches infinity as the polymerization ratio r approaches 1. We now use equations (5)-(6) to estimate the values of the model parameters. These parameter values, in turn, allow us to calibrate the key relationship (6) (i.e., Aβ burden) to the clinical setting.

32.2.2 Parameter estimation

The value of the Aβ monomer loss rate *l* in Table 32.1 corresponds to a half-life of 41.6 minutes, which is close to the crude experimental value of 38 minutes reported by one group in the brains of APP transgenic mice [8]. Protofibrils [30], fibrils [11, 14] and plaques [31] appear to grow primarily via Aβ monomer addition, and we use an elongation rate *e* in Table 32.1 taken from a synthetic fibril analysis [14], which is within a factor of two of other estimates for fibrils [11] and protofibrils [30].

The fragmentation rate *f* and the monomer production rate *p* are difficult to estimate empirically. Fortunately, we can use equations (5)-(6), together with Aβ estimates from human AD brains, to estimate these two quantities as follows. A recent study [3] estimated that 1.4% of Aβ in the human brain is soluble, and found that this soluble compartment consists primarily of monomers, dimers and trimers. A related study [32] measured the weighted average (of $A\beta_{40}$ and $A\beta_{42}$) soluble fraction of Aβ in human brains to be 1.2%. By taking the average of these two estimates (we ignore a third estimate that is only 0.05% [33]) and assuming that the soluble dimers and trimers are exactly offset by any insoluble monomers, we equate the fraction of Aβ that consists of monomers, c_1/c, to 0.013. We recognize that this analysis will underestimate the percentage of soluble Aβ in the brain if there are large amounts of soluble dimers and trimers in AD brain. By equations (5)-(6), $c_1/c = (2/(1-r)^2 - 1)^{-1}$, and setting this expression equal to 0.013 yields the polymerization ratio r = 0.84. We also equate the Aβ burden on the right side of equation (6) to the average of the total Aβ levels ($A\beta_{40}$ plus $A\beta_{42}$) found in five different cortical regions of wet brain tissue of patients with a clinical dementia rating (CDR) score of 5.0 (severe dementia) in Table 1 of [2] (this estimate of 2819 pmol/g is similar in magnitude to those in [34]), by assuming that the density of wet cortical tissue (which is about 87% water) is equal to that of water. Substituting our estimates for *l* and *r* into the right side of equation (6) and equating this expression to 2819 pmol/g yields the value of the production rate *p* in Table 32.1. Finally, because $r = pe/(fl)$, the polymerization ratio estimate r = 0.84, together with our earlier estimates of *e, l* and *p*, yield the value for the fragmentation rate *f* in Table 32.1. Note that this value of *f* is about 70 times smaller than the value of the loss rate *l,* which suggests that within the context of our model, an Aβ vaccine [4, 5] that directly ingests monomers off of oligomers can be accurately represented as a fragmentation enhancer.

Figure 32.1 Steady-state Aβ burden and clinical dementia rating (CDR) score as a function of the polymerization ratio r, with p and 1 fixed at the values in Table 1. A CDR score of 0.0 (0.5,1.0, 2.0 and 5.0, respectively) corresponds to no (questionable, mild, moderate, and severe, respectively) dementia. The relationship between steady-state Aβ ?burden (left ordinate) and CDR score (right ordinate) is based on Table 1 of [2] as explained in the text.



32.2.3 *Aβ burden*

With these parameter values in hand, we now return to equation (6) for the steady-state total Aβ burden *c*. This nonlinear relationship between the polymerization ratio and the steady-state Aβ burden is shown in Figure 32.1, and implies that a given treatment will generally achieve a greater (relative and absolute) reduction in steady-state Aβ burden for patients with higher pretreatment Aβ burdens. To position this relationship within the clinical context, we convert the Aβ burden in Figure 32.1 to the CDR scale using the data in Table 1 of [2]. This conversion allows us to explore the potential clinical impact (assuming that the neurodegeneration associated with Aβ can be reversed) of a reduction in the polymerization ratio via treatment. Recent studies reveal that Aβ peptide vaccination reduces not only Aβ burden but also cognitive impairment in APP transgenic mice [35, 36], implying that the neurodegeneration is at least partially reversible in mice. We note that the CDR scales on the right ordinates of Figures 32.1, 32.2 and 32.3 can be ignored if the reader believes that CDR will not fall in synchrony with Aβ burden in humans; the latter ordinate is still valid.

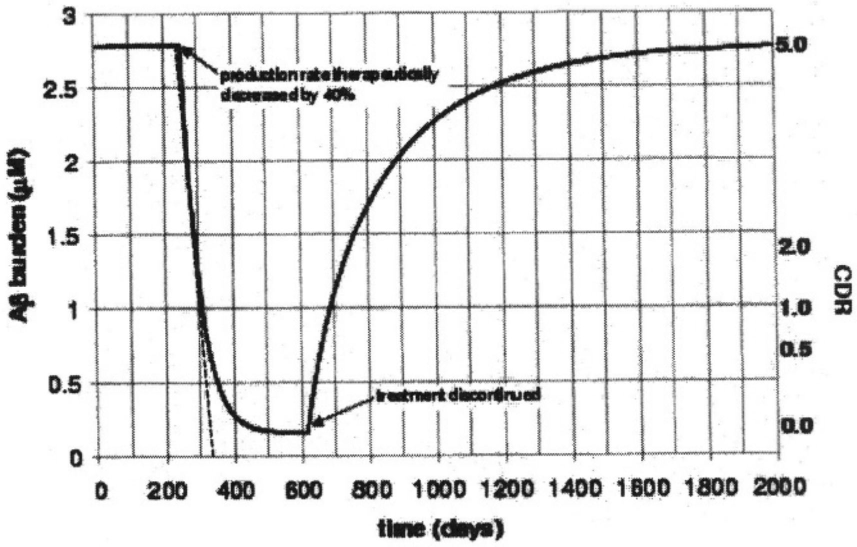
32.2.4 γ-secretase inhibitors

Because of their potentially adverse effect on Notch signaling, which is a crucial mechanism controlling cell differentiation, progression, and death [37-40], γ-secretase inhibitors might best be used to reduce only modestly the Aβ monomer production rate, perhaps by 20-60%. To assess the effect of such treatment, we numerically solve equations (1)-(3) assuming that a symptomatic patient is in a pretreatment steady state on days 0-250, and is administered a γ-secretase inhibitor that decreases the Aβ production rate p by 40% on days 250-620 (i.e., for 1 year). As shown in Figure 32.2, the post-treatment steady state represents an 18-fold drop in the Aβ burden and a theoretical decrease in the concomitant CDR score from 5.0 (severe dementia) to 0.0 (no dementia). The perturbation dynamics in Figure 32.2 are rather sluggish: the relaxation time, which we define to be the time from the start of treatment until the Aβ burden drops to 90% of the way towards the post-treatment steady state (i.e., the time until the Aβ burden equals the steady-state post-treatment Aβ burden plus 0.1 times the difference of the steady-state pretreatment and post-treatment burdens) is 113 days. Interestingly, it takes much longer – about 600 days after the discontinuation of treatment – for the Aβ burden to revert to 90% of its pretreatment steady state (assuming a polymerization ratio r of 0.84 (Table 32.1)).

32.2.5 Comparing different treatments

Equation (6) allows us to compare various treatment approaches. Within the context of our model, treatments can affect all four parameters: γ-secretase inhibitors [7, 8], or other agents that target the upstream mechanism by which Aβ is produced from APP (e.g., β-secretase inhibitors), reduce the production rate p; agents that promote the fragmentation of monomers from polymers and/or enhance the monomer clearance rate [4, 5] increase the fragmentation rate f and loss rate l, respectively; and agents that inhibit the deposition of Aβ monomers reduce the elongation rate e. By equations (4) and (6), if the production rate p is reduced by $x\%$, then the resulting Aβ burden c is $x\%$ lower than if the elongation rate e is reduced by $x\%$. Similarly, if the fragmentation rate f is increased by $y\%$, then the resulting Aβ burden is $y\%$ higher than if the loss rate is increased by $y\%$. Hence, the parameters p and l have somewhat more leverage than e and f respectively, in reducing the Aβ burden. Moreover, an $x\%$ reduction in the production

Figure 32.2 Aβ burden and CDR score versus time. A γ-secretase inhibitor, which reduces the production rate by 40%, is administered from day 250 to day 620 (i.e., for 1 year). The slope of the dashed line is the predicted rate of change in Aβ burden immediately after treatment, $p - l\bar{c}_1$, where $\bar{c}_1$ is given in equation (8).



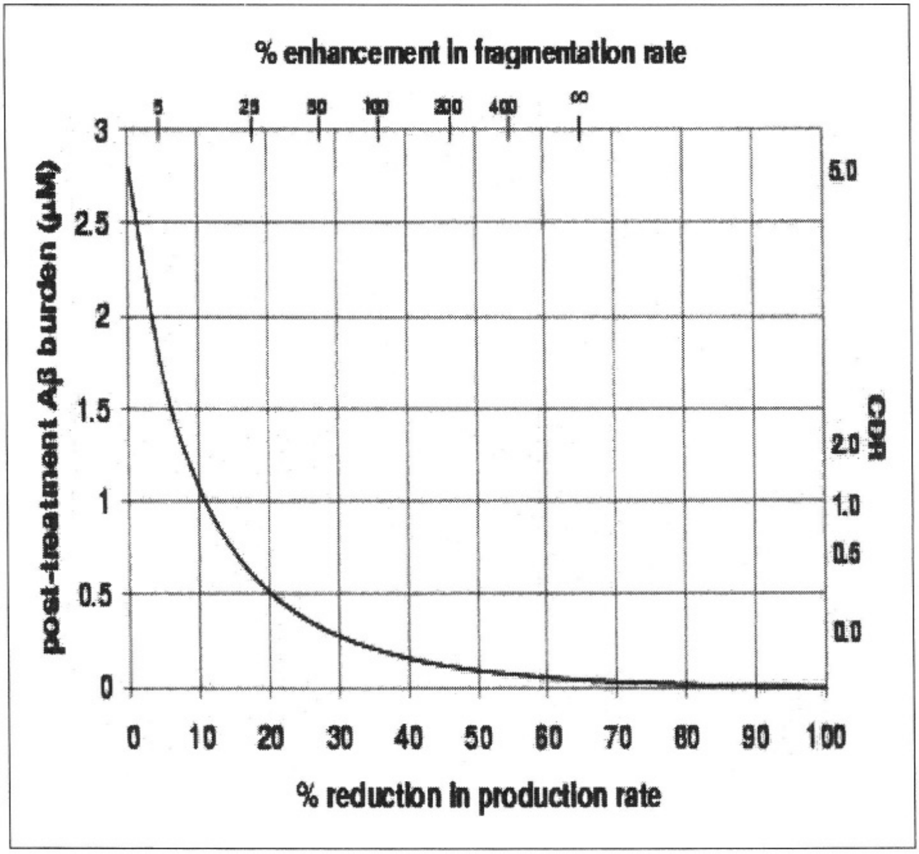
rate (elongation rate, respectively) has the same effect as a $y\%$ increase in the loss rate (fragmentation rate, respectively), where

$$y = \frac{100x}{100 - x}. \tag{7}$$

For example, Figure 32.2, which was computed using a γ-secretase inhibitor that reduces the production rate p by 40%, would also result from a treatment that increased the loss rate by 66.7%.

Unfortunately, there is no simple formula to compare the Aβ burden reduction achieved by changes in p vs. f or by changes in e vs. l. However, in Figure 32.3 we plot the post-treatment steady-state Aβ burden (and corresponding CDR score) as a function of both the percentage reduction in production rate p achieved by a γ-secretase inhibitor and the percentage increase in fragmentation rate f achieved by, for example, an Aβ vaccine (recall that l is 70 times larger than f in Table 32.1). This graph strongly suggests that only a modest (e.g., 10%) change in these parameter values is

Figure 32.3 Post-treatment Aβ burden and CDR score versus the percentage inhibition of production rate (via a γ-secretase inhibitor) and the percentage increase in fragmentation rate (e.g., via an Aβ vaccine).



required to achieve a several-fold reduction in Aβ burden and potentially a clinically significant effect on dementia. Because Aβ monomers need to be cleared after they break off, this figure also shows that an infinite fragmentation rate still leads to a finite steady-state Aβ burden, given the limits of the monomer loss rate *l*.

Additional computational results (not shown here) reveal that a fragmentation rate enhancer has a smaller (i.e., faster) relaxation time than a production rate inhibitor, probably because the relaxation time depends in large part on the fragmentation rate. For example, an agent that increases the fragmentation rate by 118% achieves the same post-treatment steady-state Aβ burden as the 40% γ-secretase inhibitor in Figure 32.2, but the relaxation time is only 40 days compared to 113 days for the γ-secretase inhibitor.

32.2.6 Post-treatment kinetics

To analyze the post-treatment Aβ kinetics, we make the simplifying assumption that the total number of oligomers, $\sum_{i=2}^{\infty} c_i$ remains constant immediately after treatment; we denote this quantity by S_2, which equals $2\frac{p}{l}\left(\frac{r}{1-r}\right)$ by equation (6). The monomer level after treatment quickly reaches a new level and thereafter changes much more gradually. This new level, $\bar{c}_1$, is approximated by setting the left side of equation (1) to zero and solving for c_1 which gives

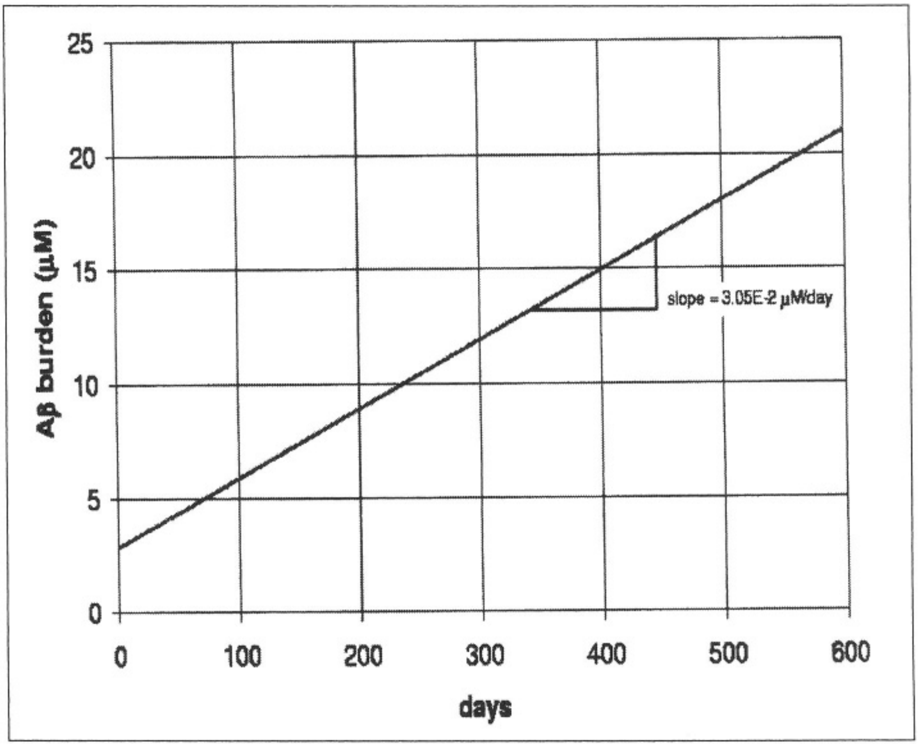
$$\bar{c}_1 = \frac{-(l+eS_2)+\sqrt{(l+eS_2)^2+8e(p+fS_2)}}{4e}. \tag{8}$$

In equation (8), S_2 is the pre-treatment number of oligomers and the four polymerization parameters represent the post-treatment values: for example, if we use a γ-secretase inhibitor that reduces the production rate by 40%, then we set p in equation (8) to 0.6 times the pretreatment production rate. Finally, summing equations (1)-(3) gives $\frac{dc}{dt} = p - lc_1$, and so immediately after treatment we predict that the total Aβ burden changes linearly at rate $p - l\bar{c}_1$, where $\bar{c}_1$ is given in equation (8). The dashed line in Figure 32.2 shows that this approximation is accurate in the initial period after treatment (in this case, for about 15 days), and can be used to estimate drug efficacy parameters (e.g., % inhibition) from post-treatment Aβ data.

32.2.7 Supercritical regime

Thus far, we have assumed that $r < 1$, so that a steady state is achieved. If $r > 1$ then the Aβ burden grows indefinitely, eventually (perhaps after many years) increasing linearly at rate $p(r-1)/r$ (Figure 32.4). Moreover, for all $r \geq 1$, the distribution of polymers no longer decays geometrically, but tends over time to a uniform distribution, where each i-mer (starting with smaller values of i) successively achieves the concentration $c_1 = f/e$, $c_i = 2f/e$ for $i = 2$.

Figure 32.4 A simulation of equations (1)-(3) when the polymerization ration $r = 1.03$. The slope of the curve approaches the asymptotic linear growth rate, $p(r-1)/r$.

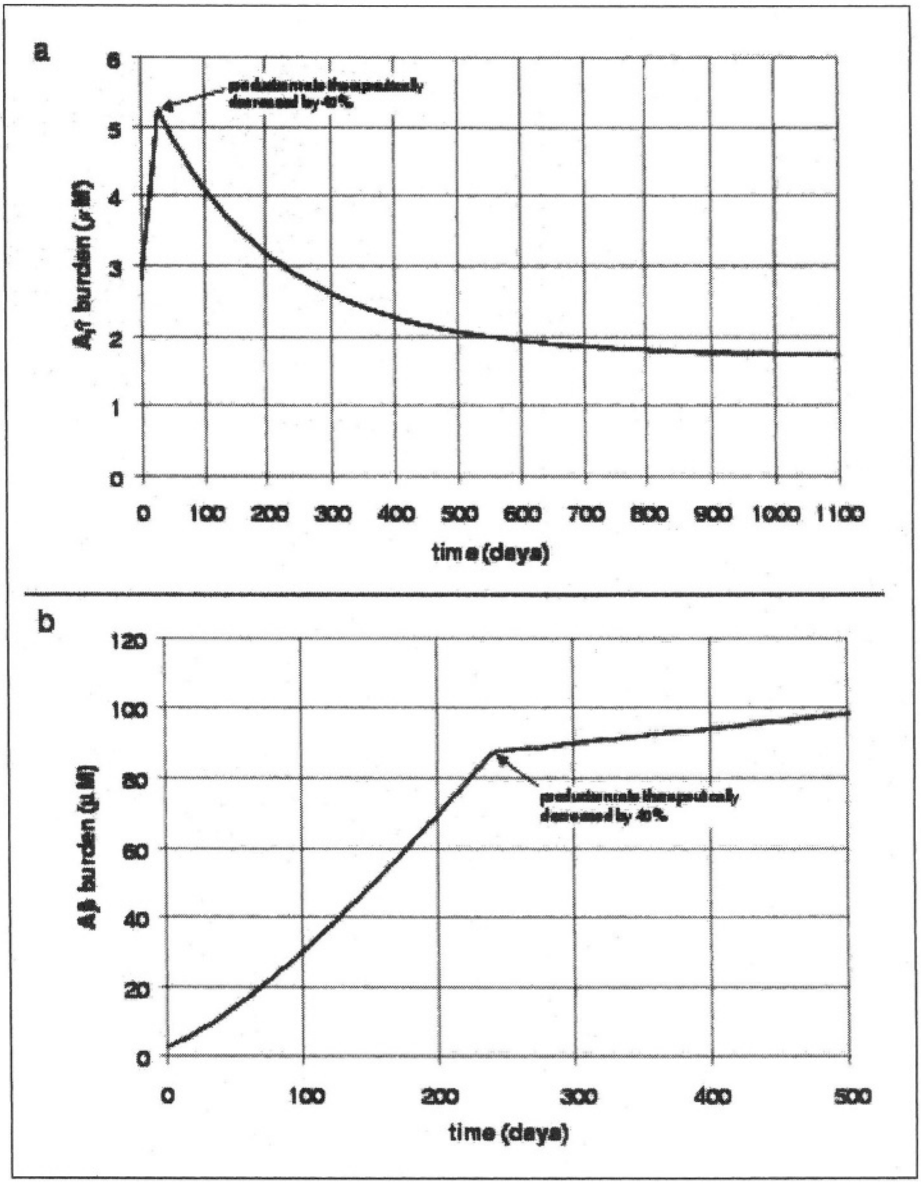


Hence, there are three possibilities for the effect of treatment: (i) the Aβ burden is in a pretreatment steady state and drops to a lower post-treatment steady state, as in Figure 32.2; (ii) treatment causes the Aβ burden to shift from the supercritical regime to the steady-state regime (Figure 32.5a); and (iii) treatment does not allow the Aβ burden to exit the supercritical regime, although it does lower the growth rate (Figure 32.5b). Comparison of Figure 32.5a to Figure 32.2 shows that it may take much longer to achieve a post-treatment steady state in case (ii) than in case (i).

32.2.8 Sensitivity analysis

Because we do not possess precise estimates for our parameter values, we performed a sensitivity analysis that varies the pretreatment polymerization ratio *r*, which is the primary driver of the Aβ dynamics, according to our analysis. We use the same parameter estimation technique as before: fix *l* and *e* to their values in Table 32.1, find *r* from the fraction of Aβ monomer

Figure 32.5 Aβ burden versus time. An Aβ production inhibitor is begun on day 30 in (a) and on day 245 in (b), and is assumed to continue indefinitely. In (a), the polymerization ratio *r* is greater than 1 before treatment, and is less than 1 after treatment. In (b), the polymerization ratio *r* is greater than 1, both before and after the start of treatment.

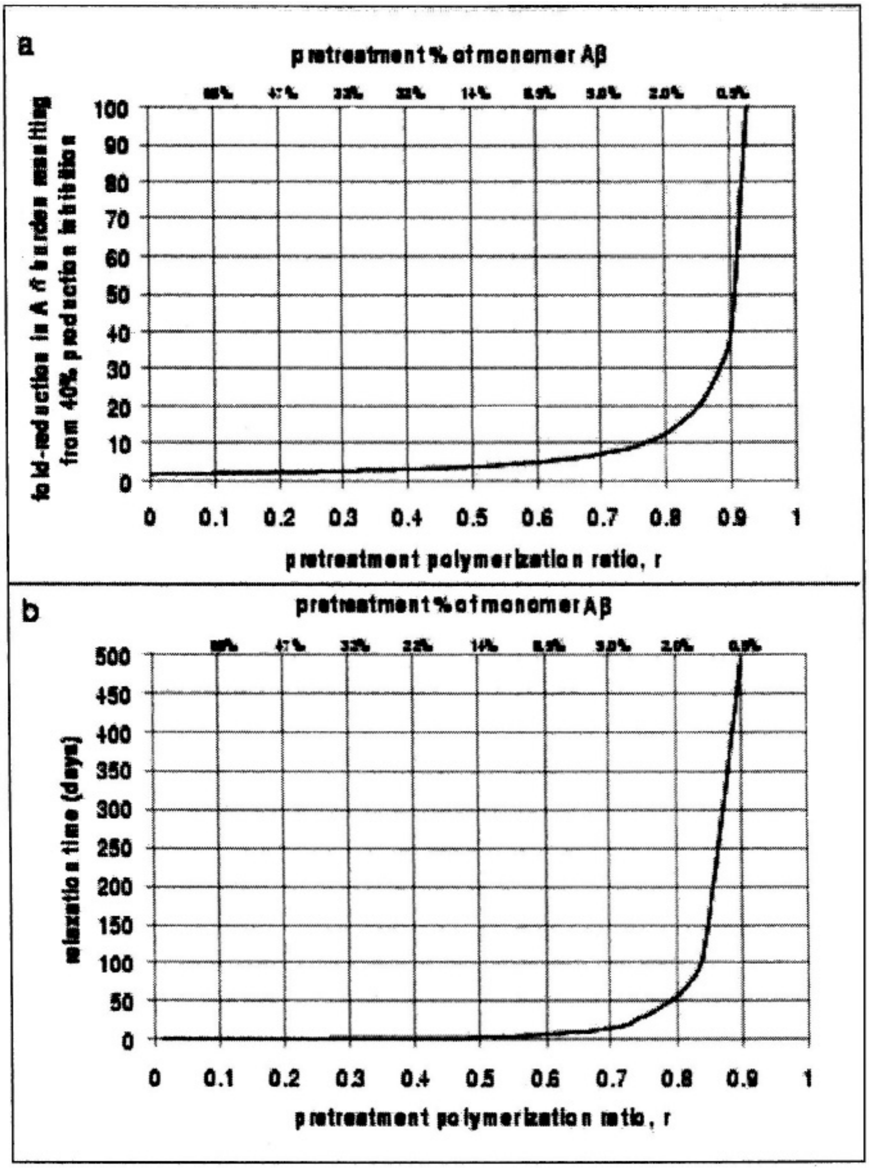


c_1/c, use the average Aβ burden of 2819 pmol/g^2 to derive p, and use r and p to calculate f. Rather than equating c_1/c to 1.3%, we now allow it to vary in order to generate a wide range of values for r, which in turn causes p and f to change. In Figures 32.6a and 32.6b, we investigate the effect caused by a 40% reduction in the production rate (via a γ-secretase inhibitor) as a function of the pretreatment polymerization ratio r. Figure 32.6a plots the pretreatment steady-state Aβ burden divided by the post-treatment steady-state Aβ burden (i.e., the fold-reduction in Aβ burden), and Figure 32.6b shows the relaxation time, which was defined earlier. Recalling our assumption that a three-fold reduction in Aβ burden leads to a clinically significant improvement in CDR score (Figure 32.1), Figure 32.6a shows that the 40% production inhibition appears capable of a significant clinical improvement for essentially all practical values of r (i.e., corresponding to percentages of monomeric Aβ, c_1/c, less than 30%), and a 1-to-2 log drop in Aβ burden as r increases from 0.76 to 0.93. Figure 32.6b shows that our base case of r = 0.84 is on the steep portion of the curve of relaxation time versus polymerization ratio. Hence, if the fraction of monomeric Aβ is actually 5% rather than 1% of total Aβ, then r drops from 0.84 to 0.7 and the relaxation time decreases from 113 days to only two weeks.

32.3 DISCUSSION

The most basic result of our analysis is that there are two possible regimes, depending upon the value of the polymerization ratio r, which equals $pe/(fl)$. If this unitless quantity is less than 1, then the Aβ burden enters a steady-state regime and takes on a finite value. If the polymerization ratio is greater than one, then the Aβ burden grows indefinitely, eventually increasing linearly at rate $p(r-1)/r$. Researchers in other areas of polymerization [22] and in other disciplines have found that the dynamics of some complex systems are dictated by a single measure that leads to a subcritical regime if the measure is less than 1 and a supercritical regime if the measure is greater than 1; two examples are the basic reproductive rate R_0 of an infectious pathogen in epidemiology [41] and the traffic intensity ? in queueing theory [42]. Moreover, as with AD, in these disciplines the ultimate measures of performance (size of the epidemic and queue length, respectively) have a highly nonlinear relationship with the unitless measure. Applied researchers in these disciplines have adopted these unitless measures as a central part of their mental model of these complex systems, and routinely estimate them and monitor achieved reductions in them. We propose that AD researchers follow suit by adopting the polymerization ratio $pe/(fl)$ *as* the key "Aβ driver" in this disease.

Figure 32.6 (a) The pretreatment steady-state Aβ burden divided by the post-treatment steady-state Aβ burden, as a function of the pretreatment polymerization ratio *r.* (b) The relaxation time (see the text for a definition) to a post-treatment steady-state Aβ burden, as a function of the pretreatment polymerization ratio *r.* In both figures, it is assumed that a γ-secretase inhibitor reduces the production rate by 40%. In the parameter estimation procedure, *r* is varied by changing the fraction of Aβ that is monomer (i.e., c_1/c).



Several studies have shown that the Aβ burden does not appear to be closely correlated with the duration or clinical severity of AD [27-29], which suggests that the polymerization ratio is less than 1 for symptomatic AD patients. It is possible that the total Aβ level is in a quasi-steady state, where some of the four parameters change slowly over time (on the order of years), causing a slow increase in the Aβ burden. Although previous mathematical models of plaque formation [18, 19] have also proposed a dynamic equilibrium of aggregation and disaggregation, the model in [19] requires that the disaggregation rate be modulated by the amount of plaque via a feedback mechanism that minimizes the changes that occur in the brain. While some feedback mechanism – for example, a cytotoxic T-lymphocyte response [43] – may exist, our model shows that when Aβ production is incorporated, a steady-state Aβ burden can be achieved by a constant fragmentation rate (i.e., in the absence of any feedback).

Nonetheless, our estimate for *r* is of the order of 10^{-1} (rather than 10^{-2} or smaller), and so it is conceivable that certain aggressive forms of the disease in humans or mice (strongly amyloidogenic presenilin mutations) may achieve a polymerization ratio greater than 1. More importantly, while studies [27-29] assess Aβ burden by counting plaques, a more recent study measures Aβ burden biochemically, and these more refined data suggest that Aβ burden is correlated with clinical severity of AD, which is not inconsistent with linear growth of Aβ burden (i.e., the supercritical regime in our model). Although further research may be required to further elucidate whether Aβ burden is in a steady state or is increasing continually in symptomatic AD patients, our identification of two regimes (steady-state and supercritical) suggests that there are three types of possible treatment outcomes: a reduction in Aβ burden from a pretreatment steady state to a post-treatment steady state (Figure 32.1), a change in regime from pretreatment growth to post-treatment steady state (Figure 32.5a), and a reduction in growth rate from a pretreatment supercritical regime to a post-treatment supercritical regime (Figure 32.5b). These results suggest that the failure of a drug to reduce the Aβ burden in a subset of mice or humans may not necessarily be due to drug inactivity, but rather could signal a pretreatment supercritical regime. Also, even if the pretreatment polymerization ratio is less than 1, a patient may revert to his pretreatment Aβ burden within several months after treatment is discontinued (see Figure 32.6b); consequently, it is important to attempt to measure the Aβ burden throughout the course of treatment in a clinical trial. In contrast, if the pretreatment polymerization ratio is greater than 1, then after discontinuation of treatment a patient's Aβ burden will always be smaller than if he had not received treatment.

Equation (6) provides a simple formula for the steady-state Aβ burden in terms of the polymerization ratio r (provided $r < 1$) and the production rate-to-loss rate ratio, p/l. As pictured in Figure 32.1, the Aβ burden is a highly nonlinear (increasing, convex) function of the polymerization ratio, and approaches infinity as r approaches 1. Hence, the impact of a modest reduction in r appears to be sufficient to produce a many-fold reduction in the Aβ burden; for the parameter values in Table 32.1, a 40% reduction in the production rate via a γ-secretase inhibitor achieves an 18-fold reduction in Aβ burden. Such a reduction is consistent with recently reported data in mice [5, 8], and appears to be sufficient to convert the disease process to a non-pathological state [2, 3, 32], assuming neuronal dysfunction is concommitantly decreased. Figure 32.1 also shows that, while the magnitude of the Aβ burden reduction increases with the strength of treatment, there are decreasing returns to scale as the treatment gets stronger. Our analysis predicts that the percentage reduction in Aβ burden achieved by a given treatment is a unimodal function (i.e., first increasing, then decreasing) of the polymerization ratio r. For $r < 1$, the percentage reduction in Aβ burden increases with r, and hence one might expect that a larger percentage reduction would be achieved in forms of AD that are particularly aggressive (e.g., the Swedish familial AD mutation [44], or the trisomy 21 mutation that occurs in Down syndrome [45]). However, as explained in the previous paragraph, the Aβ burden continues to grow in the face of treatment if the post-treatment polymerization ratio is greater than 1.

Our model is flexible enough to incorporate a variety of treatment approaches that affect any of the four model parameters. Equation (5) allows us to perform an apples-to-apples comparison of production inhibitors, elongation inhibitors, clearance enhancers, and fragmentation enhancers. For a given percentage change in parameters caused by a treatment, this analysis shows that the production inhibitors (clearance enhancers, respectively) are somewhat more effective in reducing the Aβ burden than elongation inhibitors (fragmentation enhancers, respectively). Equation (7) and Figure 32.3 provide an explicit comparison between production inhibitors and fragmentation enhancers. Computational results show that a fragmentation enhancer attains its post-treatment steady-state faster than a production inhibitor. These comparisons, coupled with future data on drug toxicity, may be useful in determining the optimal mix and level of treatments that minimize the Aβ burden subject to toxicity constraints.

Figure 32.2 shows that it takes several months for treatment to reduce the Aβ burden to a new post-treatment steady state, and the return to a pretreatment steady-state Aβ burden following the discontinuation of treatment is even more gradual. However, Figure 32.6b shows that the speed at which the Aβ level changes is highly sensitive to the pretreatment polymerization ratio *r,* suggesting that aggressive forms of AD (having high *r)* are likely to take much longer to get under control. Our lack of a precise estimate for the pretreatment polymerization ratio prevents us from accurately predicting the rapidity of response to treatment, but in some cases this response may be considerably quicker than suggested in Figure 32.2.

This general state of affairs is similar to HIV infection, as revealed several years ago in two seminal papers [46, 47]. The stability of plasma HIV levels over many years lulled the research community into believing that HIV was a relatively static disease process, but a perturbation of this steady state by a powerful protease inhibitor revealed a highly dynamic equilibrium, where a high virus production rate was offset by an equally high virus clearance rate. Hence, antiviral therapy led to a precipitous drop in plasma HIV levels, and discontinuation of treatment allowed for a rapid return of HIV levels to the pretreatment steady state. To borrow an analogy often used in the HIV field, the water level (i.e., the Aβ level) in the bath may change very slowly, but the tap (production of Aβ monomer from APP) and drain (Aβ loss) may be operating at deceptively high rates. It is worthwhile for AD researchers to keep in mind the lessons learned by the HIV clinical research community: (i) if toxicities of various types of agents are non-overlapping, then drug combinations (e.g., a γ-secretase inhibitor plus an Aβ vaccine) are likely to outperform monotherapy; (ii) drug resistance, which is the Achilles heel of HIV treatment [48], may be a problem for AD treatment as well; and (iii) given the dynamic equilibrium of Aβ – which implies that AD patients may require chronic treatment – and the interaction between Aβ production and Notch signaling [37-40], efforts should focus on developing drugs that avoid long-term toxicity, which is beginning to plague the HIV community [49].

Several of the model parameters, particularly the production rate and the fragmentation rate, are difficult to measure *in vivo.* One benefit of our analysis is that equations (5) and (6) allow an indirect estimation of the production rate *p* and the fragmentation rate *f* given published data on the Aβ burden [2] and the fraction of Aβ that is monomer [3, 32] (or, more generally, any quantity, such as mean polymer length, that can be derived from the steady-state distribution of *i*-mers). Equation (5) can also be used to measure the strength (i.e., percentage inhibition or enhancement) of a treatment *in vivo,* given data on the pretreatment and post-treatment steady-state Aβ burdens. Moreover, the transient analysis in equation (8) allows for indirect estimation of a third model parameter from serial measurements of post-treatment Aβ burden.

Our study has two important limitations. First, our model does not currently differentiate between the various forms of Aβ ($A\beta_{40}$ vs. $A\beta_{42}$, soluble vs. insoluble). However, while we have been careful to restrict the model to Aβ polymerization *per se* and ignore protofibril-to-fibril conversion and plaque growth, it is clear that our model can, in fact, be viewed as a representation of plaque growth if deposition onto plaques and fragmentation of plaques are primarily due to monomer addition and break-off, respectively. While monomer deposition onto plaques is believed to be the main form of aggregation *in vitro* [31, 50], it is quite likely that polymers coalesce *in vivo* [19]. Similarly, fragmentation of oligomers [51] and plaques may be more complex than the monomer release that is assumed here. Hence, a more realistic model of plaque growth might use a variation of Smoluchowski's equation that allows *i*-mers and *j*-mers to coalesce (e.g., at a rate proportional to $i^k j^k c_i(t) c_j(t)$, where $k = 2/3$ or 1, depending on the porosity of plaques) and allows fragmentation of monomers from the plaque surface (at rate $fi^{2/3} c_i(t)$) or the entire plaque (at rate $fic_i(t)$). While such a model would drastically change the distribution of Aβ *i*-mers relative to the current model (skewing the distribution towards huge polymers that represent plaques), it is unlikely to significantly alter the basic nature of our results pertaining to the Aβ burden. However, to the extent that the model parameters – including the effect of treatment – vary for $A\beta_{40}$ vs. $A\beta_{42}$ or soluble vs. insoluble Aβ, then these model generalizations would be worth pursuing in the future.

The second main weakness of our study is that our parameter values are not necessarily indicative of human AD brains: the loss rate l was derived from the brains of mice [8], and the elongation rate e was measured from synthetic Aβ in a low pH environment [11, 14]. Also, we crudely equated the fraction of Aβ that is monomer with the fraction of Aβ that is soluble. Our sensitivity analysis (Figures 32.6a and 32.6b) suggests that, regardless of how inaccurate our parameter estimates are, a modest (several-fold) change in a parameter value caused by treatment will still lead to a clinically significant (i.e., at least three-fold) reduction in the steady-state Aβ burden, but the time to achieve this post-treatment steady state (i.e., relaxation time) is quite sensitive to our parameter values.

Despite the model's simplicity, our analysis elucidates the basic nature of the kinetics of the Aβ burden in the face of etiologic treatments, and provides a framework – and a novel metric, the polymerization ratio – within which to interpret the laboratory and clinical results that are likely to be generated during the next few years. As more progress is made on the identity of the toxic moiety [3, 33] the elucidation of the inflammatory and neurotoxicity cascade, and the parameter values and mechanisms related to Aβ polymerization and plaque growth, our hope is that models such as this one will be refined and generalized to provide some guidance about the optimal way to employ the emerging anti-Aβ drug arsenal.

The biggest obstacle preventing the validation and subsequent use of our mathematical model is the infeasibility of unintrusively measuring the Aβ level in the human brain. Indeed, there is an urgent need to develop alternative biomarkers that are easy to monitor. Two obvious choices are the Aβ level in the cerebrospinal fluid (CSF) and plasma. We recently generalized the work presented in this chapter by considering a three-compartment model consisting of the Aβ in the brain, CSF and plasma [52]. In much the same way that Jim Jackson guessed, and then confirmed, the product-form steady-state solution to single-class queueing networks [53], we show that the steady-state brain Aβ levels in the compartmental model are similar to those derived here, except that the actual Aβ production and loss rates in the brain are replaced by effective rate that are derived by analyzing the intercompartmental flows. Our results suggest that production inhibitors reduce Aβ levels in all three compartments. More interestingly, however, treatments that ingest monomers off of polymers, or that increase fragmentation or block elongation, produce little change – or even transient increases – in CSF and plasma Aβ levels. Hence, considerable care must be taken when interpreting these biomarkers. We are currently working with scientists at Elan Pharmaceuticals, makers of the Aβ vaccine [4-5] and several secretase inhibitors, to estimate the key model parameters from their data.

Acknowledgments

This research was partially supported by the Singapore-MIT Alliance (LMW) and by the Foundation for Neurologic Diseases (DJS).

References

[1] Selkoe, D.J. (1999). Translating cell biology into therapeutic advances in Alzheimer's disease. *Nature,* 399 (Supp), A23-A31.

[2] Näislund, J., V. Haroutunian, R. Mohs, K.L. Davis, P. Davies, P. Greengard, and J.D. Buxbaum (2000). Correlation between elevated levels of amyloid β-peptide in the brain and cognitive decline. *Journal oftheAmerican MedicalAssociation,* 283, 1571-1577.

[3] McLean, C.A., R.A. Cherny, F.W. Fraser, S.J. Fuller, M.J. Smith, K. Beyreuther, A.I. Bush, and C.L. Masters (1999). Soluble pool of Aβ amyloid as a determinant of severity of neurodegeneration in Alzheimer's disease. *Annals of Neurology,* 46, 860-866.

[4] Schenk, D., R. Barbour, W. Dunn, G. Gordon, H. Grajeda, T. Guido, K. Hu, J. Huang, K. Johnson-Wood, K. Khan, D. Kholodenko, M. Lee, Z. Liao, I. Lieberburg, R. Motter, L. Mutter, F. Soriano, G. Shopp, N. Vasquez, C. Vandevert, S. Walker, M. Wogulis, T. Yednock, D. Games, and P. Seubert (1999). Immunization with amyloid-β attenuates Alzheimer-disease-like pathology in the PDAPP mouse. *Nature,* 400,173-177.

[5] Bard, F., C. Cannon, R. Barbour, R.-L. Burke, D. Games, H. Grajeda, T. Guido, K. Hu, J. Huang, K. Johnson-Wood, K. Khan, D. Kholodenko, M. Lee, I. Lieberburg, R. Motter, M. Nguyen, F. Soriano, N. Vasquez, K. Weiss, B. Welch, P. Seubert, D. Schenk, and T. Yednock (2000). Peripherally administered antibodies against amyloid β-peptide enter the central nervous system and reduce pathology in a mouse model of Alzheimer disease. *Nature Medicine 6* 916-919.

[6] Weiner, H.L., C.A. Lemere, R. Maron, E.T. Spooner, T.J. Grenfell, C. Mori, S. Issazadeh, W.W. Hancock and D.J. Selkoe (2000). Nasal administration of amyloid-betapeptide decreases cerebral amyloid burden in a mouse model of Alzheimer's disease. *Annals of Neurology,* 48, 567-579.

[7] Wolfe, M.S., W. Xia, C.L. Moore, D.D. Leatherwood, B.L. Ostaszewski, T. Rahmati, I.O. Donkor, and D.J. Selkoe (1999). Peptidomimetic probes and molecular modeling suggest Alzheimer's γ-secretase is an intramembrane-cleaving aspartyl protease. *Biochemistry,* 38, 4720-4727.

[8] Felsenstein, K.M. (2000). The next generation of AD therapeutics: the future is now. *Abstracts from the '7th annual conference on Alzheimer's disease and related disorders,* Abstract 613.

[9] Naiki, H., K. Higuchi, K. Nakakuki and T. Takeda (1991). Kinetic analysis of amyloid fibril polymerization *in vitro. Laboratory Investigation,* 65, 104-110.

[10] Naiki, H. and K. Nakakuki (1996). First-order kinetic model of Alzheimer's β-amyloid fibril extension *in vitro. Labaratory Investigation,* 74, 374-383.

[11] Lomakin, A., D.S. Chung, G.B. Benedek, D.A. Kirschner, and D.B. Teplow (1996). On the nucleation and growth of amyloid β-protein fibrils: detection of nuclei and quantitation of rate constants. *Proceedings of the National Academy of Science USA,* 93, 1125-1129.

[12] Harper, J.D. and P.T. Lansbury (1997). Models of amyloid seeding in Alzheimer's disease and scrapie: mechanistic truths and physiological consequences of the time-dependent solubility of amyloid proteins. *Annual Review of Biochemistry,* 66, 385-407.

[13] Walsh, D.M., A. Lomakin, G.B. Benedek, M.M. Condron, and D. Teplow (1997). Amyloid β-protein fibrillogenesis: detection of a protofibrillar intermediate. *Journal of Biochemistry,* 272, 22364-22372.

[14] Lomakin, A., D.B. Teplow, D.A. Kirschner, and G.B. Benedek (1997). Kinetic theory of fibrillogenesis of amyloid β-protein. *Proceedings of the National Academy of Science USA,* 94, 7942-7947.

[15] Naiki, H., K. Hasegawa, I. Yamaguchi, H. Nakamura, F. Gejyo, and K. Nakakuki (1998). Apolipoprotein E and antioxidants have different mechanisms of inhibiting Alzheimer's β-amyloid fibril formulation *in vitro. Biochemistry,* 37, 17882-17889.

[16] Inouye, H. and D.A. Kirschner (2000). Aβ fibrillogensis: kinetic parameters for fibril formation from Congo red binding. *Journal of Structural Biology,* 130, 123-129.

[17] Hyman, B.T., H.L. West, G.W. Rebeck, S.V. Buldyrev, R.N. Mantegna, M. Ukleja, S. Havlin, and H.E. Stanley (1995).

Quantitative analysis of senile plaques in Alzheimer disease: observation of log-normal size distribution and molecular epidemiology of differences associated with apolipoprotein E genotype and trisomy 21 (Down syndrome). *Proceedings of the National Academy of Science USA,* 92, 3586-3590.

[18] Cruz, L., B. Urbanc, S.V. Buldyrev, R. Christie, T. Gomez-Isla, S. Havlin, M. McNamara, H.E. Stanley, and B.T. Hyman (1997). Aggregation and disaggregation of senile plaques in Alzheimer disease. *Proceedings of the National Academy of Science USA,* 94, 7612-7616.

[19] Urbanc, B., L. Cruz, S.V. Buldyrev, S. Havlin, H.E. Stanley, and B.T. Hyman (1999). Dynamics of plaque formation in Alzheimer's disease. *Biophysical Journal,* 76, 1330-1334.

[20] Masel, J. and V.A.A. Jansen (2000). Designing drugs to stop the formation of prions and other amyloids. *Biophysical Chemistry,* 88, 47-59.

[21] Flory, P.J. (1953). *Principles of Polymer Chemistry.* Cornell University Press, Ithaca, NY.

[22] Oosawa, F. and S. Asakura (1972). *Thermodynamics of the Polymerization of Protein.* Academic Press, London.

[23] von Smoluchowski, M. (1916). Drei vorträge über diffusion, brownsche bewegung und koagulation von kolloidteilchen. *Zeitschrift für Physik,* 17, 557-585.

[24] von Smoluchowski, M. (1917). Versuch einer mathematischen theorie der koagulationskinetic kolloider lösungen. *Zeitschrift für Physik,* 92, 129-168.

[25] Family, F. and D.P. Landau (1984). *Kinetics of Aggregation and Gelation.* North-Holland, Amsterdam.

[26] Sonntag, H. and K. Strenge (1987). *Coagulation Kinetics and Structure Formation.* Plenum, New York.

[27] Hyman, B.T., K. Marzloff, and P.V. Arriagada (1993). The lack of accumulation of senile plaques or amyloid burden in Alzheimer's disease suggests a dynamic balance between amyloid deposition and resolution. *Journal of Neuropathology and Experimental Neurology,* 52, 594-600.

[28] Arriagada, P.V., J.H. Growdon, E.T. Hedley-Whyte, and B.T. Hyman (1992). Neurofibrillary tangles but not senile plaques parallel duration and severity of Alzheimer disease. *Neurology,* 42, 631-639.

[29] Berg, L., D.W. McKeel, J.P. Miller, J. Baty, and J.C. Morris (1993). Neuropathological indexes of Alzheimer's disease in demented and nondemented persons aged 80 years and older. *Archives of Neurology,* 50, 349-358.

[30] Harper, J.D., S.S. Wong, C.M. Lieber and P.T. Lansbury, Jr. (1999). Assembly of Aβ Amyloid protofibrils: an *in vitro* model for a possible early event in Alzheimer's disease. *Biochemistry,* 38, 8972-8980.

[31] Tseng, B.P., W.P. Esler, C.B. Clish, E.R. Stimson, J.R. Ghilardi, H.V. Vinters, P.W. Mantyh, J.P. Lee, and J.E. Maggio (1999). Deposition of monomeric, not oligomeric, Aβ mediates growth of Alzheimer's disease amyloid plaques in human brain preparations. *Biochemistry,*38,10424-10431.

[32] Wang, J., D.W. Dickson, J.Q. Trojanowski, and V.M.-Y. Lee (1999). The levels of soluble versus insoluble brain Aβ distinguish Alzheimer's disease from normal and pathologic aging. *Experimental Neurology,* 158, 328-337.

[33] Lue, L.-F., Y.-M. Kuo, A.E. Roher, L. Brachova, Y. Shen, L. Sue, T. Beach, J.H. Kurth, R.E. Rydel, and J. Rogers (1999). Soluble amyloid β peptide concentration as a predictor of synaptic change in Alzheimer's disease. *American Journal of Pathology,* 155, 853-862.

[34] Gravina, S. A., L. Ho, C.B. Eckman, K.E. Long, L. Otvos, Jr., L.H. Younkin, N. Suzuki, and S.G. Younkin (1995). Amyloid β (Aβ) in Alzheimer's Disease Brain: biochemical and immunocytochemical analysis with antibodies specific for forms ending at Aβ40 or Aβ42(43). *Journal of Biochemistry,* 270, 7013-7016.

[35] Janus, C., J. Pearson, J. McLaurin, P.M. Mathews, Y. Jiang, S.D. Schmidt, M. Azhar Chishti, P. Home, D. Heslin, J. French, H.T.J. Mount, R.A. Nixon, M. Mercken, C. Bergeron, P.E. Fraser, P. St. George-Hyslop, and D. Westaway (2000). Aβ peptide immunization reduces behavioral impairment and plaques in a model of Alzheimer's disease. *Nature,* 408, 979-982.

[36] Morgan, D., D.M. Diamond, P.E. Gottschall, K.E. Ugen, C. Dickey, J. Hardy, K. Duff, P. Jantzen, G. DiCarlo, D. Wilcock, K. Connor, J. Hatcher, C. Hope, M. Gordon and G.W. Arendash (2000). Aβ peptide vaccination prevents memory loss in an animal model of Alzheimer's disease. *Nature,* 408, 982-985.

[37] Wolfe, M.S., W. Xia, B.L. Ostaszewski, T.S. Diehl, W.T. Kimberley and D.J. Selkoe (1999). Two transmembrane aspartates in presenilin-1 required for presenilin endoproteolysis and γ-secretase activity. *Nature,* 398, 513-511.

[38] De Strooper, B., W. Annaert, P. Cupers, P. Saftig, K. Craessaerts, J.S. Mumm, E.H. Schroeter, V. Schrijvers, M.S. Wolfe, W.J. Ray, A. Goate, and R. Kopan (1999). A presinilin-1-dependent gamma-secretase-like protease mediates release of Notch intracellular domain. *Nature,* 398, 518-522.

[39] Struhl, G. and I. Greenwald (1999). Presinilin is required for activity and nuclear access of Notch in Drosophila. *Nature,* 398, 522-525.

[40] Ye, Y., N. Lukinova, and M.E. Fortini (1999). Neurogenic phenotypes and altered Notch processing in Drosophila Presinilin mutants. *Nature,* 398, 525-529.

[41] Diekmann, O., J.A.P. Heesterbeek, and J.A.J. Metz (1990). On the definition and the computation of the basic reproductive ratio R_0 in models for infectious diseases in heterogeneous populations. *Journal of Mathematical Biology,* 28, 365-382.

[42] Kelly, F.P. (1979). *Stochastic Networks and Reversibility.* John Wiley and Sons, New York.

[43] Monsonego, A., R. Maron, A. Bar-Or, J.I. Krieger, D. Selkoe, and H.I. Weiner (2000). The role of T cell reactivity to A-Beta amyloid peptide in the pathogenic processes associated with Alzheimer's disease. *Abstracts from the 7th annual conference on Alzheimer's disease and related disorders,* Abstract 105.

[44] Citron, M., T. Oltersdorf, C. Haass, L. McConlogue, A.Y. Hung, P. Seubert, C. Vigo-Pelfrey, I. Lieberburg, and D.J. Selkoe (1992). Mutation of the β-amyloid precursor protein in familial Alzheimer's disease increases β-protein production. *Nature,* 360, 672-674.

[45] Lemere, C.A., J.K. Blustzjan, H. Yamaguchi, T. Wisniewski, T.C. Saido, and D.J. Selkoe (1996). Sequence of deposition of

heterogeneous amyloid β-peptides and Apo E in Down syndrome: Implications for initial events in anyloid plaque formation. *Neurobiology Disease,* 3,16-32.

[46] Wei, X., S.K. Ghosh, M.E. Taylor, V.A. Johnson, E.A. Emini, P. Deutsch, J.D. Lifson, S. Bonhoeffer, M.A. Nowak, B.H. Hahn, M.S. Saag, and G.M. Shaw (1995). Viral dynamics in human immunodeficiency virus type 1 infection. *Nature,* 373, 117-123.

[47] Ho, D.D., A.U. Neumann, A.S. Perelson, W. Chen, J.M. Leonard, and M. Markowitz (1995). Rapid turnover of plasma virions and CD4 lymphocytes in HIV-1 infection. *Nature,* 373, 123-126.

[48] Condra, J.H., W.A. Schleif, O.M. Blahy, L.J. Gabrelski, D.J. Graham, J.C. Quintero, A. Rhodes, H.L. Robbins, and M. Shivaprakash (1995). *In vivo* emergence of HIV-1 variants resistant to multiple protease inhibitors. *Nature,* 374, 569-571.

[49] Carr, A., K. Samaras, D.J. Chisholm, and D.A. Cooper (1998). Pathogenesis of HIV-1 protease inhibitor-associated peripheral lipodystrophy, hyperlipidemia, and insulin resistance. *Lancet,* 131, 1881-1883.

[50] Esler, W.P., E.R. Stimson, J.R. Ghilardi, H.V. Vinters, J.P. Lee, P.W. Mantyh, and J.E. Maggio (1996). *In vitro* growth of Alzheimer's disease β-amyloid plaques displays first-order kinetics. *Biochemistry,* 35, 749-757.

[51] Cherny, R.A., P.A. Guerette, C. McLean, J.T. Legg, F.W. Fraser, I. Volitakis, C.L. Masters, and A.I. Bush (2000). Oligomeric Aβ in PBS-soluble extracts of human Alzheimer brain. In *Abstracts from the 7th annual conference on Alzheimer's disease and related disorders,* Abstract 62.

[52] Craft, S.L., L.M. Wein, and D.S. Selkoe (In press). A mathematical model of the impact of novel treatments on the Aβ burden in the Alzheimer's brain, CSF and plasma. *Bulletin of Mathematical Biology.*

[53] Jackson, J.R. (1957). Networks of waiting lines. *Operations Research,* 5, 518-521.

INDEX

Early Titles in the

INTERNATIONAL SERIES IN OPERATIONS RESEARCH & MANAGEMENT SCIENCE

Frederick S. Hillier, Series Editor, *Stanford University*

Saigal/ *A MODERN APPROACH TO LINEAR PROGRAMMING*
Nagurney/ *PROJECTED DYNAMICAL SYSTEMS & VARIATIONAL INEQUALITIES WITH APPLICATIONS*
Padberg & Rijal/ *LOCATION, SCHEDULING, DESIGN AND INTEGER PROGRAMMING*
Vanderbei/ *LINEAR PROGRAMMING*
Jaiswal/ *MILITARY OPERATIONS RESEARCH*
Gal & Greenberg/*ADVANCES IN SENSITIVITY ANALYSIS & PARAMETRIC PROGRAMMING*
Prabhu/ *FOUNDATIONS OF QUEUEING THEORY*
Fang, Rajasekera & Tsao/ *ENTROPY OPTIMIZATION & MATHEMATICAL PROGRAMMING*
Yu/ *OR IN THE AIRLINE INDUSTRY*
Ho & Tang/ *PRODUCT VARIETY MANAGEMENT*
El-Taha & Stidham/ *SAMPLE-PATH ANALYSIS OF QUEUEING SYSTEMS*
Miettinen/ *NONLINEAR MULTIOBJECTIVE OPTIMIZATION*
Chao & Huntington/ *DESIGNING COMPETITIVE ELECTRICITY MARKETS*
Weglarz/ *PROJECT SCHEDULING: RECENT TRENDS & RESULTS*
Sahin & Polatoglu/ *QUALITY, WARRANTY AND PREVENTIVE MAINTENANCE*
Tavares/ *ADVANCES MODELS FOR PROJECT MANAGEMENT*
Tayur, Ganeshan & Magazine/ *QUANTITATIVE MODELS FOR SUPPLY CHAIN MANAGEMENT*
Weyant, J./ *ENERGY AND ENVIRONMENTAL POLICY MODELING*
Shanthikumar, J.G. & Sumita, U./*APPLIED PROBABILITY AND STOCHASTIC PROCESSES*
Liu, B. & Esogbue, A.O./ *DECISION CRITERIA AND OPTIMAL INVENTORY PROCESSES*
Gal, T. Stewart, T.J., Hanne, T. / *MULTICRITERIA DECISION MAKING: Advances in MCDM Models, Algorithms, Theory, and Applications*
FOX, B.L. / *STRATEGIES FOR QUASI-MONTE CARLO*
Hall, R.W. / *HANDBOOK OF TRANSPORTATION SCIENCE*
Grassman, W.K. / *COMPUTATIONAL PROBABILITY*
Pomerol, J-C. & Barba-Romero, S. /*MULTICRITERION DECISION IN MANAGEMENT*
Axsäter, S./*INVENTORY CONTROL*
Wolkowicz, H., Saigal, R., & Vandenberghe, L. / *HANDBOOK OF SEMI-DEFINITE PROGRAMMING: Theory, Algorithms, and Applications*
Hobbs, B.F. & Meier, P. / *ENERGY DECISIONS AND THE ENVIRONMENT: A Guide to the Use of Multicriteria Methods*
Dar-El, E. / *HUMAN LEARNING: From Learning Curves to Learning Organizations*
Armstrong, J.S. / *PRINCIPLES OF FORECASTING: A Handbook for Researchers and Practitioners*
Balsamo, S., Personé, V., & Onvural, R./ *ANALYSIS OF QUEUEING NETWORKS WITH BLOCKING*
Bouyssou, D. et al. / *EVALUATION AND DECISION MODELS: A Critical Perspective*
Hanne, T. / *INTELLIGENT STRATEGIES FOR META MULTIPLE CRITERIA DECISION MAKING*
Saaty, T. & Vargas, L. / *MODELS, METHODS, CONCEPTS and APPLICATIONS OF THE ANALYTIC HIERARCHY PROCESS*
Chatterjee, K. & Samuelson, W. / *GAME THEORY AND BUSINESS APPLICATIONS*
Hobbs, B. et al. / *THE NEXT GENERATION OF ELECTRIC POWER UNIT COMMITMENT MODELS*
Vanderbei, R.J. / *LINEAR PROGRAMMING: Foundations and Extensions, 2nd Ed.*
Kimms, A. / *MATHEMATICAL PROGRAMMING AND FINANCIAL OBJECTIVES FOR SCHEDULING PROJECTS*
Baptiste, P., Le Pape, C. & Nuijten, W. / *CONSTRAINT-BASED SCHEDULING*
Feinberg, E. & Shwartz, A. / *HANDBOOK OF MARKOV DECISION PROCESSES: Methods and Applications*
Ramík, J. & Vlach, M. / *GENERALIZED CONCAVITY IN FUZZY OPTIMIZATION AND DECISION ANALYSIS*
Song, J. & Yao, D. / *SUPPLY CHAIN STRUCTURES: Coordination, Information and Optimization*
Kozan, E. & Ohuchi, A. / *OPERATIONS RESEARCH/MANAGEMENT SCIENCE AT WORK*
Bouyssou et al. / *AIDING DECISIONS WITH MULTIPLE CRITERIA: Essays in Honor of Bernard Roy*

Early Titles in the

INTERNATIONAL SERIES IN OPERATIONS RESEARCH & MANAGEMENT SCIENCE

(Continued)