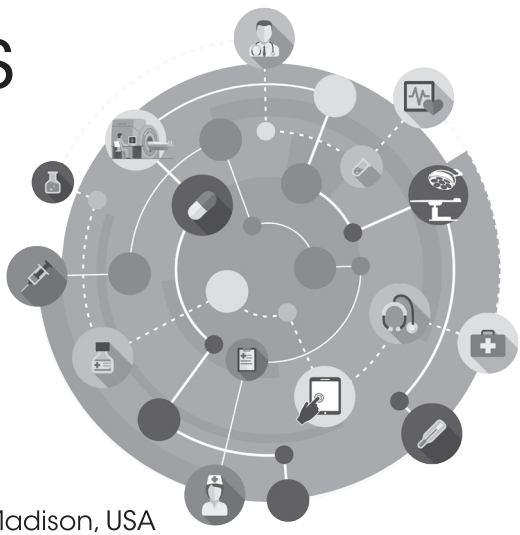
Stochastic Modeling and Analytics in Healthcare Delivery Systems

This page intentionally left blank

Stochastic Modeling and Analytics in Healthcare Delivery Systems



Editors

Jingshan Li
University of Wisconsin-Madison, USA

Nan Kong
Purdue University, USA

Xiaolei Xie
Tsinghua University, China

World Scientific

NEW JERSEY • LONDON • SINGAPORE • BEIJING • SHANGHAI • HONG KONG • TAIPEI • CHENNAI • TOKYO

Published by

World Scientific Publishing Co. Pte. Ltd.

5 Toh Tuck Link, Singapore 596224

USA office: 27 Warren Street, Suite 401-402, Hackensack, NJ 07601

UK office: 57 Shelton Street, Covent Garden, London WC2H 9HE

Library of Congress Cataloging-in-Publication Data

Names: Li, Jingshan, Dr., editor. | Kong, Nan, editor. | Xie, Xiaolei, editor.

Title: Stochastic modeling and analytics in healthcare delivery systems / [edited by] Jingshan Li, Nan Kong, Xiaolei Xie.

Description: New Jersey : World Scientific, 2017. | Includes bibliographical references and index.

Identifiers: LCCN 2017020925 | ISBN 9789813220843 (hardcover : alk. paper)

Subjects: | MESH: Delivery of Health Care | Health Services | Stochastic Processes | Models, Statistical

Classification: LCC RA394 | NLM W 84.1 | DDC 362.101/12--dc23

LC record available at https://lccn.loc.gov/2017020925

British Library Cataloguing-in-Publication Data

A catalogue record for this book is available from the British Library.

Typeset by Stallion Press

Email: enquire@stallionpress.com

Printed in Singapore

To our families

This page intentionally left blank

Preface

There has been growing interest toward research and practices in healthcare systems worldwide to improve safety, quality, and efficiency; reduce cost; and achieve better patient outcome. In an effort to increase awareness and highlight the work in these areas, this book offers a collection of papers that throw light on healthcare system management and optimization. It focuses on research and best practices in stochastic modeling and analytics in the area of healthcare engineering and technology assessment. Scientists, researchers, and practitioners are invited to present their current research outcomes in healthcare system stochastic modeling, simulation, optimization, and management.

First, patient flow, work flow, and operation management within hospitals and clinics are studied in the first five chapters.

Chapter 1 (by Joonyup Eun, Sangbok Lee, and Yuehwern Yih) presents two types of patient appointment scheduling problems — outpatient appointment scheduling in clinics and surgery scheduling — and focuses on methodologies used to solve the problems. It also provides a detailed literature review related to each specific scheduling problem: a simulation-based approach for the outpatient appointment scheduling and a stochastic optimization approach for surgery scheduling.

Chapter 2 (by Lina Aboueljinane and Evren Sahin) introduces a discrete event simulation model to study the current performance of the emergency medical service (EMS) system SAMU, which stands for the French acronym of Urgent Medical Aid Services, as well as to

investigate the effects of potential process changes that can lead to enhanced operational efficiency, in terms of the target 20 min coverage performance of primary rescues.

Chapter 3 (by Wanying Chen, Alain Guinet, and Tao Wang) studies an emergency department of a large-sized Italian hospital in normal and overcrowding (due to a major event) situations. The IDEF0 method is used to develop the conceptual models, and SIMIO is selected to simulate the conceptual models in detail. The factorial design is, then, used to analyze the impact of resource dimensioning. Finally, improvement rules are proposed.

Chapter 4 (by Na Geng) introduces a new magnetic resonance imaging (MRI) examination reservation process. A contract-based approach aims to reduce the waiting time of stroke patients for MRI examination without degrading the utilization of MRI scanner. A stochastic programming model is proposed to simultaneously determine contract decisions, and an average cost Markov decision process (MDP) approach is used to identify the structural properties of the optimal control policy.

Chapter 5 (by Zexian Zeng, Xiaolei Xie, Xiang Zhong, Barbara A. Liegel, Sue Sanford-Ring, and Jingshan Li) uses a computer simulation model to study the discharge process in medical units at the University of Wisconsin Hospital. Two main constraints of the discharge process are identified: waiting times for the physician's order and before final discharge. Reduction in physician prescription processing time and better coordination of events among discharge teams are the two potential areas for improvement.

Second, beyond hospitals and clinics, connections with other healthcare facilities and the whole healthcare network are considered in Chapters 6, 7, and 8.

Chapter 6 (by Xuxue Sun, Zhouyang Lou, Mingyang Li, Nan Kong, and Pratik J. Parikh) throws light on patient transition problems in two projects. The first project uses a binary classifier, based on the conditional logistic regression model, to predict 30-day hospital readmission incidence and utilizes decision trees to identify influential risk factors. The second project proposes a Bayesian latent heterogeneity modeling and quantification approach to characterize

time-to-transition of elderly individuals from the community to the long-term care systems. Real case studies in both projects are carried out to demonstrate the usefulness of the methods.

Chapter 7 (by Jianpei Wen and Jie Song) describes a multi-agent simulation model to quantitatively analyze the impact of different factors on a patient's choice of a healthcare facility in hierarchical healthcare systems in China, which consist of general hospitals (GHs) and community healthcare centers (CHCs). GHs and CHCs were established in urban China to improve the accessibility of healthcare services. The results show that improving the quality of treatment at CHCs and reducing CHC-related costs can attract patients toward CHCs, reducing congestion at GHs and balancing the loads between GHs and CHCs.

Chapter 8 (by Rachel M. Townsley, Maria E. Mayorga, A. Sidney Barritt IV, and Eric Orman) investigates trends in liver transplantation and evaluates the effects of these trends on the transplant-recipient population and resulting predictors of survival, that is, the survival probability and the D-MELD score. Population dynamics models are used to predict the donor and the recipient population. Then, donors and recipients are matched with a survival analysis model to predict survival outcomes posttransplant.

Third, more broad analyses related to health data management, social network, and public health are investigated in the remaining chapters.

Chapter 9 (by Chen Kan and Hui Yang) presents a new visualization and data analytics tool for stochastic modeling and analysis of cardiac electrical signals. The tool advances cardiac tele-healthcare service with exceptional features, such as personalization, responsiveness, and superior quality, and develops the next-generation cardiac mHealth system, namely, the mobile and E-network smart health (MESH).

Chapter 10 (by Yu Teng, Nan Kong, and Torsten Reimer) showcases a recent study on an agent-based social influence simulation that aims to investigate changes in individual attitudes and the formation of public opinions over time through scale-free networks. The relationship between the distribution of final public opinions

and initial distribution is studied, and the impacts of intervention and social influence in opinion change are investigated.

Chapter 11 (by Xiang Zhong, Jingshan Li, Goutham Rao, and K.P. Unnikrishnan) talks about a study conducted to generate growth curves of American children from more recent datasets and compare them with CDC reference curves in 2000. The results show that children covered in the new datasets are heavier, at any given age, than the children included in the CDC dataset, and their adiposity rebound also occurs at an earlier age, or may not even exist. These findings suggest a progressive fattening of American children, and the growth charts generated in the past as standards for measuring growth might no longer be applicable to today's population.

We are grateful to anonymous reviewers for their comments that helped improve the quality of the contributing chapters. In addition, we express our deep gratitude to Catherine Yeo Man Ling, Yubing Zhai, and Allison McGinniss of World Scientific Publishing Company for their incredible support.

Jingshan Li (University of Wisconsin-Madison, USA),
Nan Kong (Purdue University, USA), and
Xiaolei Xie (Tsinghua University, China)
2017

Contents

1. Patient Appointment Scheduling

Joonyup Eun*, Sangbok Lee† and Yuehwern Yih‡

*Department of Anesthesiology, School of Medicine, Vanderbilt University Medical Center, Nashville, Tennessee, USA
†Department of Industrial and Management Engineering, Hansung University, Seoul, South Korea
‡Regenstrief Center for Healthcare Engineering and School of Industrial Engineering, Purdue University, West Lafayette, Indiana, USA

Abstract

This chapter discusses patient appointment scheduling that manages the patient inflow to the healthcare delivery system while satisfying patient needs. It directly affects the operations within the healthcare system and the matrices related to patient outcomes, patient safety, accessibility and timely care. Patient appointment scheduling is complex and challenging due to the uncertainties associated with patient demands, disease progression, treatments, procedures, supporting services and other environmental factors, such as regulations, reimbursement, etc. To manage these uncertainties, simulation or stochastic modeling techniques are frequently used to tackle this class of scheduling problems. This chapter provides an overview of both techniques and demonstrates each technique in

an outpatient clinic setting and in a surgery scheduling setting respectively.

1.1. Introduction

In general, patient appointment scheduling is defined as determining the sequence and the time for any activity in healthcare services that requires patient attendance. The common areas for patient appointment scheduling include outpatient procedure (e.g., primary care, chemotherapy, and radiotherapy), elective surgery, diagnostic imaging, and laboratory tests. Due to the uncertainties in patient demands, procedure duration, and supporting environments, simulation or stochastic modeling techniques are frequently used to tackle such problems.

Patient appointment scheduling directly affects the overall quality of healthcare delivery systems in many ways. First, it affects how soon the patient can see the provider or get the service s/he needs, which contributes to providing timely care. A delay in appointment not only causes patient dissatisfaction but also harms the patient due to missed opportunity to treat. Second, it affects the allocation and usage of medical resources, such as providers, medical staff, equipment, patient rooms, and operating rooms (ORs). The allocation of these resources contributes to the overall healthcare expenditures. The poorly generated schedules increase costs as they do not utilize the resources efficiently and waste resources [1].

The knowledge of appointment scheduling on a couple of application areas (such as outpatient procedures, elective surgery, diagnostic imaging, and laboratory tests) can be extended to the rest of the application areas as the problems related to patient appointment scheduling share the similarities such as inherent scheduling complexity of calculating factorials, limited medical resources, and uncertain procedure durations.

This chapter focuses on two types of patient appointment scheduling problems that have gained increasing attention from researchers in healthcare operations management. Those are appointment scheduling in outpatient clinics and surgery scheduling in hospitals.

In this chapter, we will provide more detailed literature review related to each specific scheduling problem. Further, we will present a simulation-based approach for the outpatient appointment scheduling problem and a stochastic optimization approach for the surgery scheduling problem.

1.2. Appointment Scheduling in Outpatient Clinics

Appointment scheduling in outpatient clinics is one of the most important drivers to reduce escalating healthcare costs [2]. Institute of Medicine [3] expected $80 billion would be saved through the efficient use of clinical resources. Appointment scheduling in outpatient clinics aims at reducing unnecessary or inefficient use of clinical resources and, at the same time, increasing the quality of care (patient satisfaction and acute care).

Patients make appointments with doctors on a regular basis or when they feel sick. In the traditional appointment system, patients need to wait for a long time due to fully occupied appointment slots. During the waiting time, the condition of patients may get better, stay the same, or get worse. If the condition gets worse, they visit emergency department for acute care or seek for other alternatives. In any case, long waiting time causes high likelihood of no-shows and late cancellations. As no-shows and late-cancellations may not be predicted correctly, the doctor's slots are under-utilized despite being fully booked. Considering patient health as well as the high cost of using the emergency department, we need to meet patient needs better.

Open access (OA) and overbooking (OB) are alternative scheduling systems in outpatient clinics. OA leaves the majority of slots for the same day appointments. The same day appointments are made for patients who want to meet doctors on the same day because they feel sick. The remaining slots are intentionally designated for returning patients according to the requirements of the patients or the doctor [4–6]. OB allows double-booking or multiple-booking in the same slot, based on rough prediction of no-shows. For example, if the mean no-show rate of an outpatient clinic is 20%, the clinic will

open double-booking on 20% of its regular slots. Lee *et al.* [7] describes different roles of OA and OB in appointment scheduling system with Fig. 1.1. As shown in the figure, OA increases resource utilization by reducing no-shows, while OB increases it by increasing scheduled appointments. OB can be simply implemented by accepting more appointments than available slots. OB is expected to reduce waiting delays (the period time between the day an appointment is made and the actual appointment day), and thus, it helps reducing the no-show rate. OA also reduces the no-show rate; however, its implementation takes extra effort, such as reducing backlogs of appointments, which have been piled up in the traditional scheduling system, and changing appointment lead times from many days to one or a few days.

There is a large number of literature on appointment scheduling systems (Cayirli and Veral [8] provide a good review about this topic). Many research papers, particularly which considered the

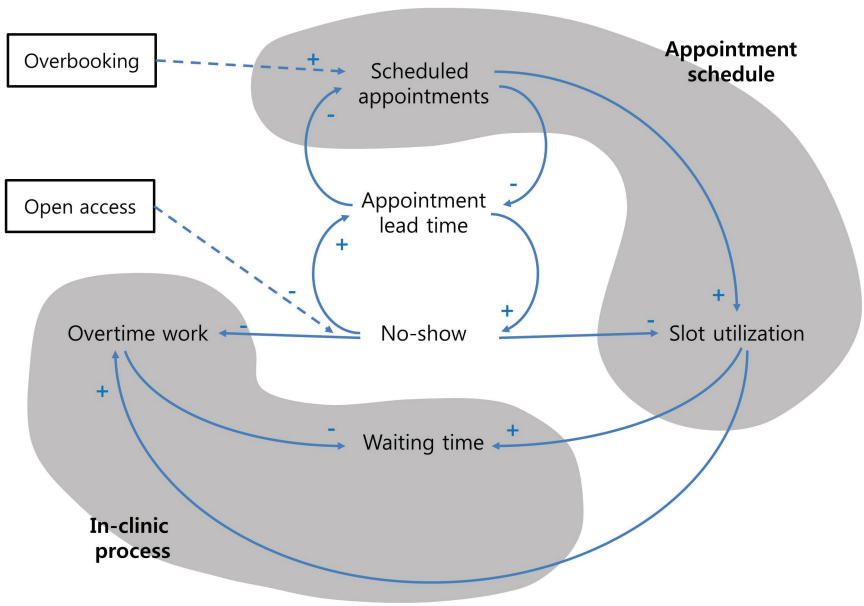


Figure 1.1. Different roles of open access and overbooking.

stochastic arrival process, have focused on modifying the appointment slot intervals and the number of patients in a slot (multiple-blocking allows the appointment time for multiple patients, while single-blocking allows the appointment of only one patient in a block). As more attention has been paid to the efficiency and the effectiveness of the clinical process, more research work has been done on advanced appointment scheduling, including OA and OB. Many case studies of OA have been investigated in real clinical settings [6, 9–10]. The optimal implementation strategy for OA has also been widely studied [11–13]. There are many analytical studies of OB in which the location and the number of OB slots are determined [14, 15]. Lee *et al.* [7] compare OA and OB to suggest a better solution for various clinic settings.

1.2.1. *Appointment scheduling in outpatient clinics with simulation*

This subsection briefly describes a simulation study for advanced appointment scheduling systems in outpatient clinics. OA and OB are renovated systems that have been recently adopted by many clinics. Although these ideas have been successfully implemented, they are not free from restrictions. For example, implementing OA requires an equilibrium between patient-demand and doctor-capacity [2, 16, 17]. OB also needs frequent overtime work by clinical personnel, particularly under over-capacity demand environments. A study to compare OA and OB performances under different demand-capacity settings can be useful to understand both systems and can also be used as a basis for designing a mixed scheduling system of OA and OB, which could perform better than OA or OB alone. A simulation study about this topic is presented in this section.

A discrete-event simulation model is developed for the comparison between OA and OB. Patients come as scheduled, but if the other patients or the doctor arrives early or late, the checked-in patients can see the doctor as soon as s/he is ready. However, the late arrival of the patient for his appointment, with the doctor waiting for him, is surely waste of costly resources. No walk-in patients are

modeled under OA as it is unlikely to have them in clinics. There is only one doctor in the model under an assumption of patient and primary-care doctor matching.

OA, when implemented in the field, mixes the same day appointment (SDA) and the long-term scheduling (LTS), while OB allows only LTS. For a fair comparison, both systems have the same number of patient demand, and the OA simulation model separates it into SDA and LTS with a certain probability. Here, the patient demand refers to patient appointment calls. In the simulation model, uncertainties for both appointment calls and actual time of arrival are considered. Daily appointment calls are modeled using normal distributions with different means, demand and capacity-equilibrium, 14% larger demand, and 28% larger demand (the latter two cases are test environments that are above the limits of system capability as well as the base condition for implementing OA). The appointment calls on a day are modeled as shown in Fig. 1.2. Since there are two peaks in real data, they are generated by two normal distributions (a bi-modal distribution) in the model. Modeling call-arrivals on a day is particularly important for OA since it has SDA. Patient earliness and lateness are modeled using a normal distribution.

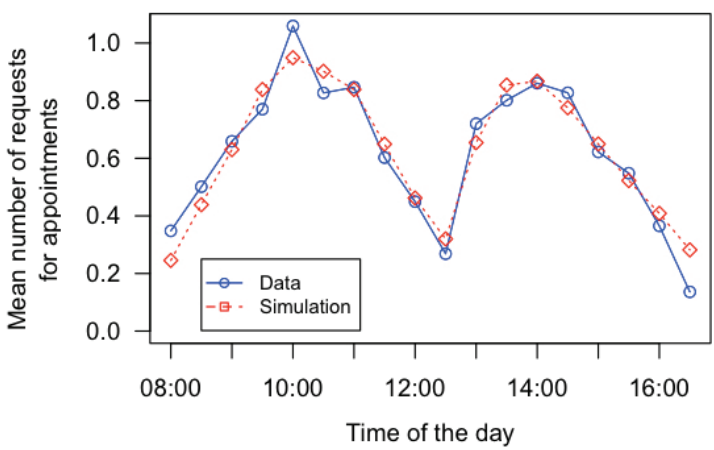


Figure 1.2 Patient calls per day from the data and the simulation model.

Patient no-show is an important issue in appointment scheduling studies. There is a large amount of literature that considers waiting time is dependent on no-show probabilities [7, 12, 18, 19]. It means that no-show rate increases as the appointment lead time (the period between the day an appointment is requested and the actual appointment day) increases. One simulation example of this idea can be obtained from Kopach *et al*. [19]. They developed the following no-show decay function along with appointment lead time, which was hinted from the exponential decay in the physical processes such as radioactive decay.

$$f_{ns}(x) = N_s \cdot (1 - 0.5 \cdot e^{-k \cdot x}) \tag{1.1}$$

where: N_s: estimated no-show probability
x: appointment lead time
k: exponential decay constants

Figures 1.3 and 1.4 describe OA and OB simulation processes. From the identically-generated daily appointment demand, individual calls are distributed with the bi-modal distribution, and then, the scheduler assigns patients by following two different flows. If any patient fails to be scheduled in any slot of a planning horizon, s/he will be classified as "unscheduled", and this case is penalized in evaluating the performance (although, in reality, they may try another time, walk-in, or go to different clinics, including emergency departments).

As many stakeholders are involved in outpatient clinics, only one performance measure is not enough to evaluate scheduling performance. The four measures defined are in-clinic waiting time, proportion of scheduled patients, appointment slot utilization, and overtime work (in minutes). The former two measures are in the patient's perspective and the latter two are in the clinic's perspective.

In-clinic waiting time refers to the period between the time a patient arrives at the clinic and the actual time s/he is seen by a doctor. The appointment delay (the period between the appointment call and the scheduled day) is not considered because OA always

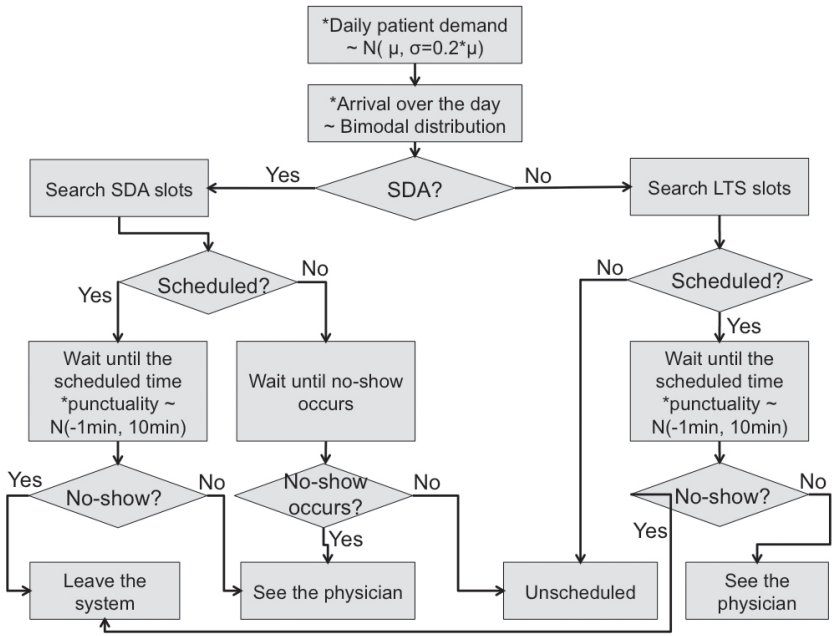


Figure 1.3. Flowchart of OA in the simulation model [7].

*Shared process with OB.

performs well with it. The proportion of scheduled patients is represented by the ratio of unmet requests (the unscheduled patients). It implies appointment backlogs or a potential burden for the society because the unscheduled patients should be taken care of. Appointment slot utilization is a straightforward measure, and it does not account for a doctor's idle time in a service slot. With the overtime work measure, early completion does not compensate for the overtime work hours.

An integrated metric, which is a linear combination of the four performance measures, is developed. There are discrepancies in the unit of the performance measures. Overtime work and waiting time are measured in minutes, while unmet appointment request and slot utilization are measured in ratio. To combine them on the same standard (ratio), overtime work and waiting time are divided by regular length of work-day. In addition, since the utilization implies positive

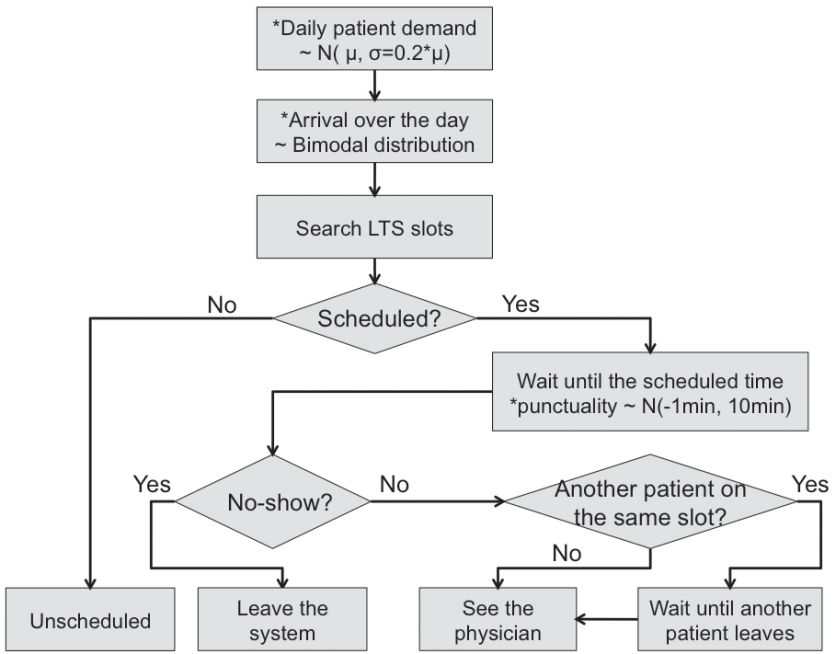


Figure 1.4. Flowchart of OB in the simulation model [7].

*Shared process with OA.

aspect, and other measures are negative aspects, utilization is replaced with (1–*utilization*). As a result, all the coefficients in the linear function are positive, and the integrated metric becomes a cost function.

The coefficients of the integrated metric are not fixed at a certain value. Various combinations of the coefficients are tested with simulation runs. Ideas about correlations among the four measures can be used to set ranges of the coefficients. For example, the US regulation stated that the overtime cost should be paid at least 50% more than the regular hours. Thus, the coefficient of the overtime work can be 1.5 times more than that of under-utilization (1–*utilization*). The unmet appointment request must be satisfied by additional work, and thus, its cost could be at least as much as that of overtime work.

The simulation runs 260 days (the number of working days per year). Since the schedule is empty at the beginning of the simulation,

the analysis of 260-day simulation run is performed after 100-day warm-up period, which is the time when the system is sufficiently in a steady state from preliminary simulation runs.

From this simulation study, for each test setting (clinic environment), the best OA and OB scheduling policies are selected. The comparison is conducted with the best policies, and then, suggestions on the scheduling policy and scheme are provided. The basic fundamental for the best OA policies is to prepare open slots as much as the proportion of the same day requests per day. However, as daily demand exceeds the capacity, reducing some same-day slots from the fundamental may be necessary to improve scheduling performance. The decision on overbooking policies is as follows. When the demand and capacity are in equilibrium, no overbooking slots are necessary. In case, the demand is above capacity, and no-show probability is relatively low (10%), it is the best to allow as many overbooking slots as the no-show rate times daily capacity. When the no-show rate is large under the high demand, e.g., 30% and 50%, clinics must add less than 30% and 50% overbooking slots, respectively (10% or 20% for 30% no-shows and 20% or 30% for 50% no-shows, according to the simulation study). In addition, as demand increases, the proportion of overbooking slots must be reduced a little (up to 10%).

Compared to OA scheduling scheme, OB, generally, performs better particularly when unmet-request (appointment backlog) cost is more expensive than overtime work cost (it is reasonable because unmet-request must be satisfied by emergency department or other medical services in the society). From the simulation study, OB outperforms OA even when the demand and the capacity are in equilibrium, and more than 80% of patients want the same-day appointments. It is very interesting because OA is known to function well in the demand-capacity equilibrium condition. Also, although the majority of patients want to be seen on the same day, OB, which does not particularly prepare the same-day slots, functions better.

In case of higher demand than the system's capacity, better scheduling scheme can be determined by the cost ratio between appointment backlog and overtime work. If the cost ratio is one, OA

is better. If appointment backlog is more expensive than overtime work, OB can be the choice. It is a predictable result as OB takes the risk of accumulating appointments every day. When the demand is much higher than the capacity (around 50% more), comparing two scheduling schemes is meaningless because the system cannot handle the case at all. For more details, the authors recommend to see Lee *et al.* [7].

1.3. Surgery Scheduling

Operating rooms (OR) are the most cost-intensive area in hospitals. Surgery operations comprise more than 40% of the expenses of hospitals [20–22, 12] due to the involvement of expensive resources (e.g., surgeons, anesthesiologists, nurses, surgical equipments, and ORs). Furthermore, the schedule for surgery operations has a significant impact on perioperative patient flow in hospitals [23]. Therefore, hospitals are under pressure to develop efficient surgery schedules that reduce costs for expensive resources and the patient flow delay.

Surgery scheduling is very demanding since many resources, their activities, and uncertainty of those activities need to be considered. For this reason, researches are increasingly paying attention to surgery scheduling problems. In the following subsections, solution methodologies for surgery scheduling problems are examined and an overall procedure of the sample average approximation (SAA) method, which deals with uncertainty in surgery durations, is discussed. More general reviews on surgery scheduling problems can be found in Magerlein and Martin [24]; Przasnyski [25]; Blake and Carter [26]; Cardoen *et al.* [27]; May *et al.* [28]; Hulshof *et al.* [29].

1.3.1. *Methodologies*

This subsection examines solution methodologies used to solve stochastic surgery scheduling models in the papers published in or after the year 2000. In this section, mixed integer programming (MIP) also includes pure integer programming.

As uncertainty related to surgery durations and patient arrivals is incorporated in surgery scheduling problems, stochastic models comprise a great portion of the recent surgery scheduling literature.

Table 1.1 classifies the literature based on the solution methodologies used to solve its stochastic models: MIP-/linear programming (LP)-based approach, dynamic programming (DP)-based approach, heuristic, and simulation. There may be no clear-cut classification between MIP-/LP-based approach and heuristic, and between DP-based

Table 1.1. Solution methodologies used to solve stochastic models.

	Solution Methodologies			
Paper	**MIP[a]-/ LP[b]-based Approach**	**DP[c]-based Approach**	**HEU[d]**	**SIM[e]**
Bowers and Mould [33]				X
Cardoen and Demeulemeester [34]				X
Denton and Gupta [30]	X			
Denton *et al.* [20]	X		X	
Denton *et al.* [31]	X		X	
Dexter [35]				X
Dexter [36]				X
Dexter and Traub [37]				X
Epstein and Dexter [38]				X
Gupta [39]	X			
Hans *et al.* [40]			X	X
Herring and Herrmann [41]		X		
Lamiri *et al.* [42]	X			
Lamiri *et al.* [43]	X			
Lebowitz [44]				X
Lee and Yih [45]			X	
Mancilla and Storer [32]	X			

(*Continued*)

Table 1.1. (*Continued*)

Paper	Solution Methodologies			
	MIP[a]-/ LP[b]-based Approach	DP[c]-based Approach	HEU[d]	SIM[e]
Marcon and Dexter [46]				X
Marcon *et al.* [47]	X			X
Min and Yih [48]		X		
Min and Yih [22]	X			
Pérez Gladish *et al.* [49]	X			
Sciomachen *et al.* [50]				X
Tyler *et al.* [51]				X
Wullink *et al.* [52]				X

[a]Mixed integer programming.
[b]Linear programming.
[c]Dynamic programming.
[d]Heuristic.
[e]Simulation.

approach and heuristic. It is because, in most solution methodologies for stochastic models, the optimality of a solution is not easily guaranteed even though the solution methodologies are based on the exact algorithms for deterministic MIP/LP or DP. Therefore, in this section, solution methodologies that employ MIP/LP formulations are classified as MIP-/LP-based approaches and those that employ DP formulations are classified as DP-based approaches.

Among MIP-/LP-based approaches, a prevalent way to tackle the problems is to convert the stochastic models to the deterministic-equivalent models using scenarios (i.e., realized sets of random parameters) and, then, solve the deterministic models using well-established techniques, such as the L-shaped decomposition [30, 31], the bender's decomposition [32], and the total unimodularity [22].

A few papers employ the DP framework, in which solution structures are analysed, and a Bellman equation is constructed [41, 48].

In many cases, MIP-/LP-based and DP-based approaches are time-consuming. Therefore, to solve large-size problems, time-efficient heuristics are also presented and evaluated [20, 31, 41].

If the problems under consideration are too complex or have a lot of stochasticity, it is hard to formulate the problems as mathematical programming models. In those cases, simulation serves as a good modeling tool. Simulation allows researchers to describe their problems in detail and easily change the details. For these reasons, simulation has been successfully and frequently used in the literature.

1.3.2. *Surgery scheduling with SAA method*

This subsection describes an overall procedure of the SAA method to solve a well-known surgery scheduling problem. The theoretical background of the SAA method can be found in Ahmed and Shapiro [53]; Kleywegt *et al.* [54].

Example problem

Denton and Gupta [30] introduced a surgery scheduling problem. The objective of the problem is to find a surgery schedule that minimizes the expected total cost for patient waiting time, OR idle time, and OR overtime in a particular OR on a particular day (or block). The same problem with slight modifications on the mathematical formulation was also studied by Denton *et al.* [20]; Mancilla and Storer [32]. The mathematical formulation suggested by Mancilla and Storer [32] is used in this section as an example to explain the SAA. A stochastic mixed integer program to formulate the problem is as follows:

$$\min E\left[\sum_{i=1}^{n}\sum_{j=1}^{n} c_j^w W_{ij} + \sum_{i=1}^{n}\sum_{j=1}^{n} c_j^s S_{ij} + c^l L\right] \tag{1.2}$$

subject to

$$t_{i+1} - t_i + \sum_{j=1}^{n} W_{i+1,j} - \sum_{j=1}^{n} W_{ij} - \sum_{j=1}^{n} S_{ij} = \sum_{j=1}^{n} Z_j x_{ij}, i = 1,\ldots,n-1 \tag{1.3}$$

$$t_n + \sum_{j=1}^{n} W_{nj} + \sum_{j=1}^{n} Z_j x_{nj} - L + G = d, \tag{1.4}$$

$$\sum_{i=1}^{n} x_{ij} = 1, \quad j = 1, \ldots, n \tag{1.5}$$

$$\sum_{j=1}^{n} x_{ij} = 1, \quad i = 1, \ldots, n \tag{1.6}$$

$$S_{ij} \le M x_{ij}, \quad i = 1, \ldots, n; j = 1, \ldots, n \tag{1.7}$$

$$W_{ij} \le M x_{ij}, \quad i = 1, \ldots, n; j = 1, \ldots, n \tag{1.8}$$

$$t_1 = 0 \tag{1.9}$$

$$S_{ij} \ge 0, W_{ij} \ge 0, \quad i = 1, \ldots, n; j = 1, \ldots, n \tag{1.10}$$

$$L \ge 0, G \ge 0 \tag{1.11}$$

$$t_i \ge 0, \quad i = 2, \ldots, n \tag{1.12}$$

$$x_{ij} \in \{0, 1\}, \quad i = 1, \ldots, n; j = 1, \ldots, n \tag{1.13}$$

Notations

- j: surgery index, $j = 1, \ldots, n$
- i: position index (i.e., ith position) in the surgery sequence, $i = 1, \ldots, n$
- c_j^w: patient waiting time penalty for surgery j
- c_j^s: OR idle time penalty for surgery j
- c^l: OR overtime penalty
- d: available time duration in which surgeries can be performed without OR overtime
- M: sufficiently large number
- Z_j: random surgery duration for surgery j
- t_i: scheduled starting time for the surgery in position i
- L: OR overtime
- G: slack variable that means the earliness with respect to d

$$x_{ij}: \begin{cases} 1 \text{ if surgery } j \text{ is assigned to position } i \\ 0 \text{ otherwise} \end{cases}$$

W_{ij}: patient waiting time when surgery j is assigned to position i
S_{ij}: OR idle time when surgery j is assigned to position i

The expected total costs for patient waiting time, OR idle, and OR overtime are minimized in objective function (1.2). Constraint (1.3) defines W_{ij} and S_{ij} (see Fig. 1.5), and constraint (1.4) defines L and G (see Fig. 1.6). Constraints (1.5) and (1.6) ensure that each surgery is assigned to one position and each position accommodates one surgery. Constraints (1.7) and (1.8) force S_{ij} and W_{ij} to be zero if surgery j is not assigned to position i. Constraint (1.9) ensures that the starting time of the first surgery is zero.

SAA model

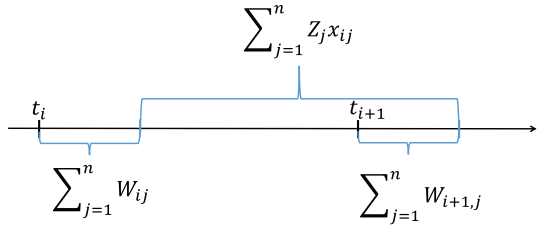
For the exact evaluation of objective function (1.2) in the stochastic mixed integer program, the deterministic-equivalent mixed integer program for every realization of the uncertain surgery durations needs to be solved, which is computationally prohibitive [53]. Therefore, a number of samples have been taken from the surgery duration distribution for each patient and plugged into the deterministic-equivalent mixed integer program (i.e., SAA model).

The SAA model for the original stochastic mixed integer program is as follows:

$$\min \frac{1}{m}\sum_{k=1}^{m}\left[\sum_{i=1}^{n}\sum_{j=1}^{n}c_j^w w_{ij}(\xi_k)+\sum_{i=1}^{n}\sum_{j=1}^{n}c_j^s s_{ij}(\xi_k)+c^l l(\xi_k)\right] \tag{1.14}$$

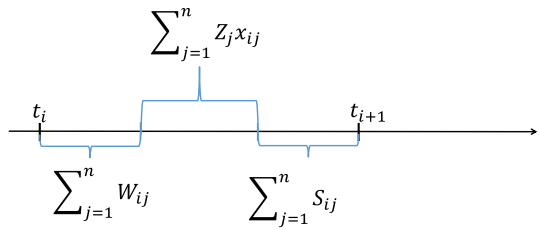
subject to

$$t_{i+1}-t_i+\sum_{j=1}^{n}w_{i+1,j}(\xi_k)-\sum_{j=1}^{n}w_{ij}(\xi_k)-\sum_{j=1}^{n}s_{ij}(\xi_k)=\sum_{j=1}^{n}z_j(\xi_k)x_{ij}, \quad i=1,\ldots,n-1; k=1,\ldots,m \tag{1.15}$$



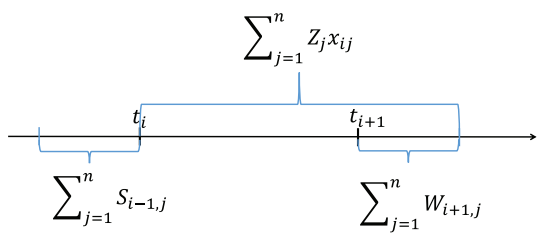
$$t_{i+1} - t_i + \sum_{j=1}^{n} W_{i+1,j} - \sum_{j=1}^{n} W_{i,j} = \sum_{j=1}^{n} Z_j x_{ij}$$

(a) Case 1



$$t_{i+1} - t_i - \sum_{j=1}^{n} W_{ij} - \sum_{j=1}^{n} S_{ij} = \sum_{j=1}^{n} Z_j x_{ij}$$

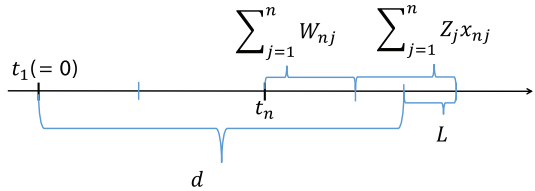
(b) Case 2



$$t_{i+1} - t_i + \sum_{j=1}^{n} W_{i+1,j} = \sum_{j=1}^{n} Z_j x_{ij}$$

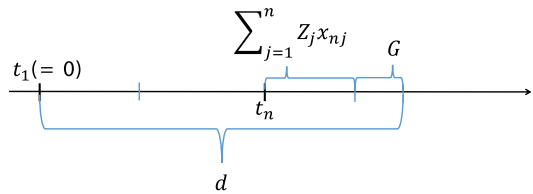
(c) Case 3

Figure 1.5. Relationship among $t_i, \Sigma_{j=1}^{n} W_{ij}, \Sigma_{j=1}^{n} W_{i+1,j}, \Sigma_{j=1}^{n} S_{ij}$, and $\Sigma_{j=1}^{n} Z_j x_{ij}$.



$$t_n + \sum_{j=1}^{n} W_{nj} + \sum_{j=1}^{n} Z_j x_{nj} - L = d$$

(a) Case 1



$$t_n + \sum_{j=1}^{n} Z_j x_{nj} + G = d$$

(b) Case 2

Figure 1.6. Relationship among $t_n, \Sigma_{j=1}^{n} W_{nj}, \Sigma_{j=1}^{n} Z_j x_{ij}, L$ and G.

$$t_n + \sum_{j=1}^{n} w_{nj}(\xi_k) + \sum_{j=1}^{n} z_j(\xi_k) x_{nj} - l(\xi_k) + g(\xi_k) = d, k = 1, \ldots, m \tag{1.16}$$

$$\sum_{i=1}^{n} x_{ij} = 1, \quad j = 1, \ldots, n \tag{1.17}$$

$$\sum_{j=1}^{n} x_{ij} = 1, \quad i = 1, \ldots, n \tag{1.18}$$

$$s_{ij}(\xi_k) \le M x_{ij}, \quad i = 1, \ldots, n; j = 1, \ldots, n; k = 1, \ldots, m \tag{1.19}$$

$$w_{ij}(\xi_k) \le M x_{ij}, \quad i = 1, \ldots, n; j = 1, \ldots, n; k = 1, \ldots, m \tag{1.20}$$

$$t_1 = 0 \tag{1.21}$$

$$s_{ij}(\xi_k) \geq 0, w_{ij}(\xi_k) \geq 0, \quad i = 1,...,n; j = 1,...,n; k = 1,...,m \quad (1.22)$$

$$l(\xi_k) \geq 0, g(\xi_k) \geq 0 \quad k = 1,...,m \quad (1.23)$$

$$t_i \geq 0, \quad i = 2,...,n \quad (1.24)$$

$$x_{ij} \in \{0,1\}, \quad i = 1,...,n; j = 1,...,n \quad (1.25)$$

New notations

ξ_k : kth scenario that defines kth realization of the surgery duration verctor $\vec{Z} = (Z_1, Z_2, ..., Z_n),\ k = 1,...,m$

$z_j(\xi_k)$: element of scenario ξ_k that defines surgery j's duration

$l(\xi_k)$: realization of L under scenario ξ_k

$g(\xi_k)$: realization of G under scenario ξ_k

$w_{ij}(\xi_k)$: realization of w_{ij} under scenario ξ_k

$s_{ij}(\xi_k)$: realization of s_{ij} under scenario ξ_k

The SAA model is able to be solved using optimization software, like AIMMS (http://www.aimms.com/), AMPL (http://www.ampl.com/), and GAMS (https://www.gams.com/), to obtain an surgery schedule (i.e., SAA solution) and its objective value (1.14). The objective value (1.14) of an SAA solution approximates the objective value (1.2) in the stochastic mixed integer program, and the quality of the SAA solution is examined by statistical analysis.

One-sided confidence interval on the optimality gap

To obtain statistical results that evaluate the quality of SAA solutions, the SAA model is solved several times, changing the sets of scenarios.

Let r be the number of replications of the SAA model and w_k^p be the element of the scenario set $\{\omega_k^p \mid k = 1,...,m\}$ used for the pth SAA replication. A feasible schedule X is defined as a feasible solution set $\{t_i, X_{ij} \mid i = 1,...,n; j = 1,...,n\}$. Let X^* be the optimal schedule of the

stochastic mixed integer program, X^{p*} be the optimal schedule of the pth SAA replication, $\phi(X)$ be the objective value (1.2) of X, and $\theta(X,\omega_k^p)$ be the objective value (1.14) of X under the scenario set $\{\omega_k^p \mid k=1,...,m\}$.

For a given schedule $\hat{X}$, the optimality gap is defined as

$$\phi(\hat{X}) - \phi(X^*). \tag{1.26}$$

$\phi(\hat{X})$ is estimated by

$$\bar{v}(\hat{X}) := \frac{1}{r}\sum_{p=1}^{r}\left[\frac{1}{m}\sum_{k=1}^{m}\theta(\hat{X},\omega_k^p)\right]. \tag{1.27}$$

$\phi(X^*)$ is estimated by

$$\tilde{v} := \frac{1}{r}\sum_{p=1}^{r}\left[\frac{1}{m}\sum_{k=1}^{m}\theta(X^{p^*},\omega_k^p)\right]. \tag{1.28}$$

Therefore, $\phi(\hat{X}) - \phi(X^*)$ is estimated by

$$\bar{v}(\hat{X}) - \tilde{v}. \tag{1.29}$$

The variance of $\bar{v}(\hat{X}) - \tilde{v}$ is

$$\frac{1}{r(r-1)}\sum_{p=1}^{r}\left[\left\{\frac{1}{m}\sum_{k=1}^{m}\theta\left(\hat{X},\omega_k^p\right) - \frac{1}{m}\sum_{k=1}^{m}\theta\left(X^{p^*},\omega_k^p\right)\right\} - \left\{\bar{v}(\hat{X}) - \tilde{v}\right\}\right]^2 \tag{1.30}$$

Note that since $\bar{v}(\hat{X})$ is always greater than or equal to $\bar{v}$ regardless of $(\hat{X})$, the $100(1-\alpha)\%$ one-sided confidence interval on the optimality gap (CIOOG) can be constructed for sufficiently large r, by the central limit theorem, as follows:

$$\phi(\hat{X}) - \phi(X^*) + z_\alpha\sqrt{\frac{1}{r(r-1)}\sum_{p=1}^{r}\left[\left\{\frac{1}{m}\sum_{k=1}^{m}\theta\left(\hat{X},\omega_k^p\right) - \frac{1}{m}\sum_{k=1}^{m}\theta\left(X^{p^*},\omega_k^p\right)\right\} - \left\{\bar{v}(\hat{X}) - \tilde{v}\right\}\right]^2} \tag{1.31}$$

where z_α is the value such that $Pr(Z > z_\alpha) = \alpha$, and Z is a standard normal random variable.

Since $\tilde{v}(\hat{X})$ is an unbiased estimator of $\phi(\hat{X})$, the SAA method usually selects an SAA solution that yields the lowest value (1.27) among several SAA solutions obtained using different sets of scenarios [22]. However, SAA solutions can be further analyzed by generating more scenarios [54] and/or examining each solution's confidence interval [55].

Numerical example

We present a simple numerical example and solve it with the SAA model.

Suppose that three patients are to be scheduled in an OR on the same day for laparoscopy and tubal cautery, inquinal hernia repair, and laparoscopic cholecystectomy. It is assumed that surgery durations (i.e., Z_j) follow the lognormal distribution [56], and their means and standard deviations are given in Table 1.2. To make the example simple, other parameters are set to be as follows: c_j^w (patient waiting time penalty) = c_j^s (OR idle time penalty) = c^l (OR overtime penalty) = 1 for all j, d (available time duration) = 480 minutes, and m (number of scenarios) = 10.

The SAA model is solved 30 times (i.e., $r = 30$) by GAMS 24.1.3, changing the sets of scenarios that are taken from the surgery duration distributions.

Table 1.2. Statistics for surgery durations.

		Surgery Duration (In Minutes)	
j	*Description	Mean*	Standard Deviation*
1	Laparoscopy and tubal cautery	105	27.4
2	Inquinal hernia repair	143	38.5
3	Laparoscopic cholecystectomy	219	47.2

*Statistics from Strum *et al.* [57].

Table 1.3 shows the 30 SAA solutions obtained, their objective values (1.27), and 95% CIOOGs (1.31). Furthermore, average patient waiting time, OR idle time, and OR overtime are calculated for each SAA solution. As aforementioned, an SAA solution that yields the lowest objective value (1.27) is typically selected. It is the solution obtained during the 25th replication. The solution indicates the following three points: 1) the second and third surgeries should be scheduled to start about 96 minutes and 236 minutes, respectively, after the first surgery starts, 2) surgeries 1, 2, and 3 should be performed in that order, and 3) if the schedule is implemented, the average patient waiting time, OR idle time, and OR overtime are about 47 minutes, 12 minutes, and 23 minutes, respectively.

Note that selecting a solution with the lowest patient waiting time, OR idle time, or OR overtime is not a good decision. In the parameter setting, patient waiting time, OR idle time, and OR overtime are equally weighted (i.e., $c_j^w = c_j^s = c^l = 1$ for all j). The SAA solutions are obtained under the assumption: if OR managers consider a measure more important than the others, its coefficient needs to be declared heavily in the parameter setting. This way optimizes the trade-off between the performance measures.

1.4. Summary

This chapter presents two types of patient appointment scheduling problems (i.e., appointment scheduling in outpatient clinics and surgery scheduling in hospitals), focusing on methodologies used to solve these problems. There are a great number of papers that deal with patient appointment scheduling problems in healthcare delivery systems. The scheduling problems are especially complex and challenging due to the dynamic nature of patient demand, procedures, and environments, in addition to the inherent scheduling complexity of calculating factorials. Therefore, in recent years, many researchers in healthcare operations management have applied advanced solution methodologies (e.g., simulation-based optimization techniques, such as; the SAA method) to solve those problems. We believe that this trend will continue and more advanced solution

Table 1.3. SAA solutions to the numerical example.

p	SAA Solution ($\hat{X}$)	Objective Value (1.27)	Patient Waiting Time*	OR Idle Time*	OR Overtime*	95% CIOOG (1.31)
1	$\hat{t}_2 = 106.76, \hat{t}_3 = 233.71, \hat{x}_{11} = 1, \hat{x}_{22} = 1, \hat{x}_{33} = 1$	83.47	48.17	10.99	24.31	14.70
2	$\hat{t}_2 = 114.62, \hat{t}_3 = 248.15, \hat{x}_{11} = 1, \hat{x}_{22} = 1, \hat{x}_{33} = 1$	85.37	41.43	16.38	27.55	16.98
3	$\hat{t}_2 = 87.59, \hat{t}_3 = 213.26, \hat{x}_{11} = 1, \hat{x}_{22} = 1, \hat{x}_{33} = 1$	90.79	64.80	5.04	20.95	24.10
4	$\hat{t}_2 = 174.46, \hat{t}_3 = 304.92, \hat{x}_{12} = 1, \hat{x}_{21} = 1, \hat{x}_{33} = 1$	130.11	34.32	43.02	52.77	66.51
5	$\hat{t}_2 = 102.53, \hat{t}_3 = 230.11, \hat{x}_{11} = 1, \hat{x}_{22} = 1, \hat{x}_{33} = 1$	83.17	50.11	9.67	23.39	14.58
6	$\hat{t}_2 = 129.39, \hat{t}_3 = 260.22, \hat{x}_{12} = 1, \hat{x}_{21} = 1, \hat{x}_{33} = 1$	91.96	41.12	22.96	27.88	24.43
7	$\hat{t}_2 = 216.26, \hat{t}_3 = 338.22, \hat{x}_{13} = 1, \hat{x}_{21} = 1, \hat{x}_{32} = 1$	93.42	43.16	23.14	27.12	26.11
8	$\hat{t}_2 = 82.63, \hat{t}_3 = 225.43, \hat{x}_{11} = 1, \hat{x}_{22} = 1, \hat{x}_{33} = 1$	87.95	58.50	8.12	21.34	20.76
9	$\hat{t}_2 = 97.38, \hat{t}_3 = 231.36, \hat{x}_{11} = 1, \hat{x}_{22} = 1, \hat{x}_{33} = 1$	82.31	49.29	10.20	22.82	13.85

(*Continued*)

Table 1.3. (*Continued*)

p	SAA Solution ($\hat{X}$)	Objective Value (1.27)	Patient Waiting Time*	OR Idle Time*	OR Overtime*	95% CIOOG (1.31)
10	$\hat{t}_2 = 204.60, \hat{t}_3 = 349.96, \hat{x}_{13} = 1, \hat{x}_{22} = 1, \hat{x}_{31} = 1$	101.02	62.13	16.43	22.46	34.56
11	$\hat{t}_2 = 106.66, \hat{t}_3 = 334.76, \hat{x}_{11} = 1, \hat{x}_{23} = 1, \hat{x}_{32} = 1$	88.81	36.88	24.49	27.45	20.99
12	$\hat{t}_2 = 92.77, \hat{t}_3 = 252.35, \hat{x}_{11} = 1, \hat{x}_{22} = 1, \hat{x}_{33} = 1$	85.20	39.25	20.19	25.76	17.05
13	$\hat{t}_2 = 111.47, \hat{t}_3 = 255.34, \hat{x}_{11} = 1, \hat{x}_{22} = 1, \hat{x}_{33} = 1$	84.80	36.31	20.19	28.30	16.86
14	$\hat{t}_2 = 122.42, \hat{t}_3 = 218.39, \hat{x}_{12} = 1, \hat{x}_{21} = 1, \hat{x}_{33} = 1$	92.43	66.53	4.99	20.92	26.36
15	$\hat{t}_2 = 139.78, \hat{t}_3 = 248.83, \hat{x}_{12} = 1, \hat{x}_{21} = 1, \hat{x}_{33} = 1$	86.51	44.44	16.14	25.93	18.23
16	$\hat{t}_2 = 91.38, \hat{t}_3 = 254.08, \hat{x}_{11} = 1, \hat{x}_{22} = 1, \hat{x}_{33} = 1$	86.50	39.05	21.32	26.13	18.51
17	$\hat{t}_2 = 130.74, \hat{t}_3 = 371.59, \hat{x}_{12} = 1, \hat{x}_{23} = 1, \hat{x}_{31} = 1$	97.91	45.06	27.39	25.46	30.73
18	$\hat{t}_2 = 85.99, \hat{t}_3 = 220.46, \hat{x}_{11} = 1, \hat{x}_{22} = 1, \hat{x}_{33} = 1$	87.85	60.13	6.58	21.15	20.85
19	$\hat{t}_2 = 138.57, \hat{t}_3 = 231.59, \hat{x}_{12} = 1, \hat{x}_{21} = 1, \hat{x}_{33} = 1$	88.28	55.60	9.20	23.48	20.56

20	$\hat{t}_2 = 105.32, \hat{t}_3 = 252.92, \hat{x}_{11} = 1, \hat{x}_{22} = 1, \hat{x}_{33} = 1$	83.10	37.07	19.12	26.91	14.86
21	$\hat{t}_2 = 101.55, \hat{t}_3 = 315.30, \hat{x}_{11} = 1, \hat{x}_{23} = 1, \hat{x}_{32} = 1$	87.52	46.07	17.49	23.96	19.76
22	$\hat{t}_2 = 78.11, \hat{t}_3 = 321.22, \hat{x}_{11} = 1, \hat{x}_{23} = 1, \hat{x}_{32} = 1$	94.53	51.98	19.95	22.59	27.48
23	$\hat{t}_2 = 114.26, \hat{t}_3 = 271.07, \hat{x}_{11} = 1, \hat{x}_{22} = 1, \hat{x}_{33} = 1$	92.14	30.35	28.44	33.35	25.79
24	$\hat{t}_2 = 110.76, \hat{t}_3 = 258.51, \hat{x}_{11} = 1, \hat{x}_{22} = 1, \hat{x}_{33} = 1$	85.49	34.52	22.02	28.95	17.86
25	$\hat{\mathbf{t}}_2 = 95.90, \hat{\mathbf{t}}_3 = 236.11, \hat{\mathbf{x}}_{11} = 1, \hat{\mathbf{x}}_{22} = 1, \hat{\mathbf{x}}_{33} = 1$	**81.76**	**46.82**	**11.83**	**23.11**	**13.13**
26	$\hat{t}_2 = 91.67, \hat{t}_3 = 347.38, \hat{x}_{11} = 1, \hat{x}_{23} = 1, \hat{x}_{32} = 1$	95.51	32.71	33.77	29.04	28.22
27	$\hat{t}_2 = 94.99, \hat{t}_3 = 234.63, \hat{x}_{11} = 1, \hat{x}_{22} = 1, \hat{x}_{33} = 1$	82.04	47.76	11.40	22.88	13.54
28	$\hat{t}_2 = 208.86, \hat{t}_3 = 372.35, \hat{x}_{13} = 1, \hat{x}_{22} = 1, \hat{x}_{31} = 1$	101.28	50.46	24.79	26.03	34.60
29	$\hat{t}_2 = 73.34, \hat{t}_3 = 232.97, \hat{x}_{11} = 1, \hat{x}_{22} = 1, \hat{x}_{33} = 1$	93.86	60.67	11.35	21.84	26.56
30	$\hat{t}_2 = 107.43, \hat{t}_3 = 333.68, \hat{x}_{11} = 1, \hat{x}_{23} = 1, \hat{x}_{32} = 1$	88.84	37.76	23.69	27.39	21.03

*In minutes.

methodologies will be applied and developed for patient appointment scheduling problems.

In addition, for future research, patient condition should also be considered in patient appointment scheduling. Patient condition has hardly been considered in literature as it has been believed to be subjective. However, there are already many measures for patient condition, such as Karnofsky grade [58], model for end-stage liver disease (MELD) score [59], and dyspnea index [60]. The analysis of the trajectories of these measures will allow us to incorporate patient condition in patient appointment scheduling to improve patient safety.

References

1. Litvak, E (2005). Optimizing patient flow by managing its variability. In *Front Office to Front Line: Essential Issues for Health Care Leaders*. Oakbrook Terrace, IL: Joint Commission Resourses.
2. Gupta, D and B Denton (2008). Appointment scheduling in health care: Challenges and opportunities. *IIE Trans*, 40(9), 800–819.
3. Institute of Medicine (2007). Health care: Billions wasted. Business week posted by Cathy Arnst on October 07.
4. Murray, M and C Tantau (1999). Redefining open access to primary care. *Manage Care Q*, 7(3), 45–55.
5. Murray, M and D Berwick (2003). Advanced access: Reducing waiting and delays in primary care. *J Am Med Assoc*, 289(8), 1035–1040.
6. Murray, M and C Tantau (2000). Same-day appointments: Exploding the access paradigm. *Fam Pract Manage*, 7(8), 45–50.
7. Lee S, D Min, J Ryu and Y Yih (2013). A simulation study of appointment scheduling in outpatient clinics: Open access and overbooking. *Simul Trans Soc Model Simul Int*, 89(12), 1459–1473.
8. Cayirli, T and E Veral (2003). Outpatient scheduling in health care: A review of literature. *Prod Oper Manage.*, 12(4), 519–549.
9. O'Hare, CD and J Corlett (2004). The outcomes of open-access scheduling. *Fam Pract Manage*, 11(2), 35–38.
10. Belardi, FG, S Weir and FW Craig (2004). A controlled trial of an advanced access appointment system in a residency family medicine center. *Fam Med*, 36(5), 341–345.

11. Qu, X, R Rardin, J Williams and D Willis (2007). Matching daily healthcare provider capacity to demand in advanced access scheduling systems. *European J Oper Res*, 183(2), 812–826.
12. Lee, S and Y Yih (2010). Analysis of open access appointment scheduling system in outpatient clinics—A simulation study. *Simul Trans Soc Model Simul Int*, 86(8–9), 503–518.
13. Peng, Y, X Qu and J Shi (2014). A hybrid simulation and genetic algorithm approach to determing the optimal scheduling templates for open access clinics and admitting walk-in patients. *Comput Ind Eng*, 72, 282–296.
14. LaGanga, L and S Lawrence (2007). Clinic overbooking to improve patient access and increase provider productivity. *Decis Sci*, 38(2), 251–276.
15. LaGanga, L and S Lawrence (2012). Appointment overbooking in health care clinics to improve patient service and clinic performance. *Prod Oper Manage*, 21(5), 874–888.
16. Green, L and S Savin (2008). Reducing delays for medical appointments: A queueing approach. *Oper Res*, 56(6), 1526–1538.
17. Liu, N, Ziya S and V Kulkarni (2010). Dynamic scheduling of outpatient appointments under patient no-shows and cancellations. *Manuf Ser Oper Manage*, 12(2), 347–364.
18. Hofmann, P and J Rockart (1969). Implication of the no-show rate for scheduling OPD appointments. *Hosp Prog*, 50(8), 35–40.
19. Kopach, R, P DeLaurentis, M Lawley, K Muthuraman, L Ozsen, R Rardin, H Wan, P Intrevado, X Qu and D Willis (2007). Effects of clinical characteristics on successful open access scheduling. *Healthc Manage Sci*, 10, 111–124.
20. Denton, B, J Viapiano and A Vogl (2007). Optimization of surgery sequencing and scheduling decisions under uncertainty. *Healthc Manage Sci*, 10(1), 13–24.
21. Pham, D-N and A Klinkert (2008). Surgical case scheduling as a generalized job shop scheduling problem. *Eur J Oper Res*, 185(3), 1011–1025.
22. Min, D and Y Yih (2010). Scheduling elective surgery under uncertainty and downstream capacity constraints. *Eur J Oper Res*, 206(3), 642–652.
23. Amato-Vealey, EJ, P Fountain and D Coppola (2012). Perfecting patient flow in the surgical setting. *AORN J*, 96(1), 46–57.
24. Magerlein, JM and JB Martin (1978). Surgical demand scheduling: A review. *Health Ser Res*, 13(4), 418.

25. Przasnyski, ZH (1986). Operating room scheduling: A literature review. *AORN J*, 44(1), 67–82.
26. Blake, JT and MW Carter (1996). Surgical process scheduling: A structured review. *J Soc Health Syst,* 5(3), 17–30.
27. Cardoen, B, E Demeulemeester and J Beliën (2010). Operating room planning and scheduling: A literature review. *Eur J Oper Res,* 201(3), 921–932.
28. May, JH, WE Spangler, DP Strum and LG Vargas (2011). The surgical scheduling problem: Current research and future opportunities. *Prod Oper Manage,* 20(3), 392–405.
29. Hulshof, PJ, N Kortbeek, RJ Boucherie, EW Hans and PJ Bakker (2012). Taxonomic classification of planning decisions in health care: A structured review of the state of the art in OR/MS. *Health Syst,* 1(2), 129–175.
30. Denton, B and D Gupta (2003). A sequential bounding approach for optimal appointment scheduling. *IIE Trans,* 35(11), 1003–1016.
31. Denton, B, AJ Miller, HJ Balasubramanian and TR Huschka (2010). Optimal allocation of surgery blocks to operating rooms under uncertainty. *Oper Res,* 58, 4-part-1, 802–816.
32. Mancilla, C and R Storer (2012). A sample average approximation approach to stochastic appointment sequencing and scheduling. *IIE Trans,* 44(8), 655–670.
33. Bowers, J and G Mould (2004). Managing uncertainty in orthopaedic trauma theatres. *Eur J Oper Res,* 154(3), 599–608.
34. Cardoen, B and E Demeulemeester (2008). Capacity of clinical pathways a strategic multi-level evaluation tool. *J Med Syst,* 32(6), 443–452.
35. Dexter, F (2000). A strategy to decide whether to move the last case of the day in an operating room to another empty operating room to decrease overtime labor costs. *Anesth Analg*, 91(4), 925–928.
36. Dexter, F (2003). Operating room utilization: Information management systems. *Curr Opin Anesthesiol,* 16(6), 619–622.
37. Dexter, F and RD Traub (2002). How to schedule elective surgical cases into specific operating rooms to maximize the efficiency of use of operating room time. *Anesth Analg,* 94(4), 933–942.
38. Epstein, RH and F Dexter (2002). Uncertainty in knowing the operating rooms in which cases were performed has little effect on operating room allocations or efficiency. *Anesth Analg,* 95(6), 1726–1730.
39. Gupta, D (2007). Surgical suites' operations management. *Prod Oper Manage,* 16(6), 689–700.

40. Hans, E, Wullink G, Van Houdenhoven M and G Kazemier (2008). Robust surgery loading. *Eur J Oper Res,* 185(3), 1038–1050.
41. Herring, WL and JW Herrmann (2012). The single-day surgery scheduling problem: Sequential decision-making and threshold-based heuristics. *OR Spectrum*, 34(2), 429–459.
42. Lamiri, M, X Xie, A Dolgui and F Grimaud (2008). A stochastic model for operating room planning with elective and emergency demand for surgery. *Eur Oper Res,* 185(3), 1026–1037.
43. Lamiri, M, X Xie and S Zhang (2008). Column generation approach to operating theater planning with elective and emergency patients. *IIE Trans,* 40(9), 838–852.
44. Lebowitz, P (2003). Schedule the short procedure first to improve or efficiency. *AORN J,* 78(4), 651–659.
45. Lee, S and Y Yih (2014). Reducing patient-flow delays in surgical suites through determining start-times of surgical cases. *Eur J Oper Res,* 238(2), 620–629.
46. Marcon, E and F Dexter (2006). Impact of surgical sequencing on post anesthesia care unit staffing. *Healthc Manage Sci,* 9(1), 87–98.
47. Marcon, E, S Kharraja and G Simonnet (2003). The operating theatre planning by the follow-up of the risk of no realization. *Int J Prod Econ,* 85(1), 83–90.
48. Min, D and Y Yih (2010). An elective surgery scheduling problem considering patient priority. *Comput Oper Res,* 37(6), 1091–1099.
49. Pérez Gladish, B, M Arenas Parra, A Bilbao Terol and MV Rodrıguez Urıa (2005). Management of surgical waiting lists through a possibilistic linear multiobjective programming problem. *Appl Math Comput,* 167(1), 477–495.
50. Sciomachen, A, E Tanfani and A Testi (2005). Simulation models for optimal schedules of operating theatres. *Int J Simul,* 6(12–13), 26–34.
51. Tyler, DC, CA Pasquariello and C-H Chen (2003). Determining optimum operating room utilization. *Anesth Anal,* 96(4), 1114–1121.
52. Wullink, G, M Van Houdenhoven, EW Hans, JM van Oostrum, M van der Lans and G Kazemier (2007). Closing emergency operating rooms improves efficiency. *J Med Syst,* 31(6), 543–546.
53. Ahmed, S and A Shapiro (2002). *The sample average approximation method for stochastic programs with integer recourse.* Technical Report, School of Industrial and Systems Engineering. Atalanta, GA: Georgia Institute of Technology.

54. Kleywegt, AJ, A Shapiro and T Homem-de Mello (2002). The sample average approximation method for stochastic discrete optimization. *SIAM Journal on Optimization*, 12(2), 479–502.
55. Bayraksan, G and DP Morton (2006). Assessing solution quality in stochastic programs. *Math Program,* 108(2–3), 495–514.
56. Spangler, WE, DP Strum, LG Vargas and JH May (2004). Estimating procedure times for surgeries by determining location parameters for the lognormal model. *Health Care Management Science*, 201(3), 97–104.
57. Strum, DP, AR Sampson, JH May and LG Vargas (2000). Surgeon and type of anesthesia predict variability in surgical procedure times. *Anesthesiology*, 92(5), 1454–1466.
58. Oken, MM, RH Creech, DC Tormey, J Horton, TE Davis, ET Mcfadden and PP Carbone (1982). Toxicity and response criteria of the eastern cooperative oncology group. *Am J Clin Oncol,* 5(6), 649–656.
59. Kamath, PS and W Kim (2007). The model for end-stage liver disease (meld). *Hepatology*, 45(3), 797–805.
60. Mahler, DA and CK Wells (1988). Evaluation of clinical methods for rating dyspnea. *Chest J,* 93(3), 580–586.

2. A Simulation Model of French Emergency Medical Service

Aboueljinane Lina*,†, Sahin Evren‡,§ and Jemai Zied¶

*École Nationale Supérieure des Mines de Rabat,
Avenue Hadj Ahmed Cherkaoui-BP: 753,
Agdal, Rabat, Morocco
†AMIPS-Ecole Mohammadia d'Ingénieurs,
Université Mohamed 5 Avenue Ibn Sina B.P. 765,
Agdal, Rabat, Morocco
‡AP-HP, Hôpital Henri Mondor, Créteil Cedex,
94010, France
§Laboratoire Genie Industriel, CentraleSupélec,
Université Paris-Saclay, Grande Voie des Vignes,
92290 Chatenay-Malabry, France
¶University of Tunis El Manar, ENIT-OASIS, 1002,
Tunis, Tunisia

Abstract

The French Emergency Medical Service of the Val-de-Marne department, known as SAMU 94, is a public safety system responsible for the coordination of pre-hospital care under emergency conditions. Pre-hospital care is requested through emergency calls

and includes the stabilization of patient's condition and transportation, when needed, to an appropriate care facility. The current research aims at improving the organizational processes of SAMU 94 in order to meet the population's needs under limited resources. Hence, we develop a discrete event simulation (DES) model in order to model and assess the current performance of this complex system, as well as to investigate the effects of potential process changes that would lead to enhanced operational efficiency, in terms of response time performance (i.e. the period between the receipt of a call and the first arrival of a rescue team at the scene), which is a critical aspect for SAMU providers. The developed DES model was validated using historical data and used as a decision-support tool for comparing the relative benefits of several scenarios mainly related to the needed resource levels and static location of rescue teams throughout the Val-de-Marne area and their assignment to incoming calls. Sensivity analyses were also performed by changing values of some input parameters such as arrival rates of calls, travel times and service times.

2.1. Introduction

In France, emergency medical services (EMS) are known as SAMU, which is the French acronym for Urgent Medical Aid Service. They are public safety systems that coordinate the delivery of pre-hospital care to patients under medical emergency conditions. During pre-hospital care, care givers stabilize the patient's condition and transport him or her to an appropriate care facility in order to prevent further injury and reduce mortality. It is, however, very difficult to evaluate the interaction between the survival rate of patients and organizational changes in the EMS process. The difficulties are mainly associated to the determination of accurate quantitative measures that affect the patient's survival through the time [1, 2]. Thus, more readily measurable quality of care metrics, such as response time and coverage, has attracted the attention of both researchers and EMS providers. The response time is the period between the incident reporting and the arrival of a rescue team at the scene of accident, and the coverage can be defined as the percentage

of calls responded to within a specific target time. The association between low response time/high coverage and high survival rate of patients has been observed by several authors in the medical literature, especially in the case of life-threatening emergencies [3–7]. For instance, the likelihood of survival from cardiac arrest decreases by 7%–10% for each minute of delay in response time [8]. Therefore, a high coverage within a target time of 20 min is a common objective of SAMU systems. Achieving this timeliness objective requires a careful management of limited resources (such as physicians, rescue vehicles, and call center operators), especially considering the high level of uncertainty (in terms of factors such as the frequency and location of demand and the location and availability of vehicles) that characterizes these systems.

In this context, the current study has been proposed with the aim of using discrete event simulation (DES) tool to model, evaluate, and improve SAMU performance of a specific French department: the Val-de-Marne (SAMU 94). DES is an operational research tool that has been extensively used in recent years to design, investigate, improve efficiency, and reduce costs of various healthcare delivery systems. A DES model typically represents the patient flow as the movement of individual entities through a series of queues and care processes at discrete points in time, in order to assess the current performance, identify areas of improvement and waste, and predict the impact of several design and operational changes over different metrics (such as patient throughput, waiting times, and the length of stay) [9, 10]. To our knowledge, there have been several applications of DES in the healthcare sector that address a wide range of problems, such as capacity and hospital bed planning [11], design of the emergency department [12], patient flow and waiting [13], geographical locations of new healthcare services [14], and emergency medical services [15]. Thus, DES is a particularly well-suited technique, in the context of EMS systems, that describes the system in a high degree of detail and avoids simplifying assumptions required to obtain performance measure predictions when using methods such as mathematical programming or queuing theory [16].

The present study uses the potential of DES to help SAMU 94 managers formalize their current care process and compare various strategic and operational scenarios to answer questions such as:

- How many call center operators and medical teams should be hired?
- Where the fleet of rescue vehicles should be located throughout the Val-de-Marne department?
- How a rescue vehicle to an incoming call should be assigned?
- What policy should be adopted due to an increase in the demand in the coming years?

The remaining sections of chapters have been organized as follows: Section 2 gives a short literature review on the use of the simulation tool in the context of EMS. Section 3 presents the detailed methodology used to build and validate the SAMU 94 DES model. Section 4 describes the results of the DES model. Finally, Section 5 reports some concluding remarks and the directions for future research.

2.2. The Use of Simulation in the EMS Literature

In the literature of EMS, the use of simulation tool was initiated in the late sixties by Savas [17] to examine the cost-effectiveness of several changes in the number and location of ambulances in the New York ambulance service. This use was, then, intensified due to several advances in the simulation tool, such as improved ease of use, development of input and output analysis tools, and integration facilities to other software. Thus, authors in EMS literature used simulation to assess the impact of several potential changes, named scenarios, on several selected performance indicators [18]. The considered scenarios were mainly related to decisions at various levels of planning, such as:

- Adding/removing rescue teams, i.e., emergency vehicles staffed by one or several physician(s), nurse(s), and/or emergency medical technician(s) in the existing or new waiting positions, called bases, to adequately cover the service area [19–22]

- Determining the deployment strategy at the mid-term, i.e., the assignment of rescue teams to bases to reach patients promptly and to achieve a particular service-level objective [23–32]
- Determining the redeployment strategy at short term, i.e., the assignment of rescue teams to bases that can be adjusted according to changes in the temporal and geographical demand pattern during a time period (known as multi-period redeployment) or the real-time availability of rescue teams following the allocation or release of a team (known as dynamic redeployment) [33–37]
- Determining the working hours and location for both vehicles and crew to satisfy demand for rescue teams in each base specified in the deployment/redeployment strategy [21, 32]
- Selecting the dispatching rule, i.e., the assignment of the best rescue team to an incoming call, based on the geographic location of the fleet, to minimize the total distance or time for all rescues [35, 38, 39]
- Determining the best sequence for assigning injured victims of mass casualty incidents to appropriate destination hospitals in order to maximize the overall survival rate [40, 41]

Results pertaining to the above scenarios provided clear measure of the relative benefits of some alternatives against the others. In this regard, EMS managers used simulation results as a decision-making and communication tool particularly to face the evolution of some factors (e.g., demand and transportation times) or to anticipate the impact of new reforms (e.g., changes in legislation regarding the location of bases or the desired level of coverage). It is noteworthy to mention that none of the scenarios discussed in the literature outperformed others in absolute terms, and thus, the performance of each alternative was highly dependent on assumptions, operation rules, and initial data pertaining to the studied system. Therefore, we developed a DES model to represent the specificity of the SAMU 94 system in a detailed manner (sequence of events in the emergency process, the type and schedule of resources involved, and the characteristics of the Val-de-Marne transport network) and to pinpoint potential areas of improvement. The developed model provides SAMU 94 managers with a flexible tool that enables the investigation

of a wide variety of scenarios that will lead to enhanced performance before committing real resources.

2.3. The DES Model of SAMU 94

This section is structured in accordance with fundamental steps used in the literature to build a simulation study [42–44]. It discusses the formulation of the problem, data inputs, as well as model implementation and validation.

2.3.1. *The SAMU 94 process description*

In France, the SAMU system is managed at the department level (i.e., a French administrative division corresponding to a median area of 6,000 km^2 and a median population of approximately 510,000 inhabitants) and provides 24-hour service for each department. In this study, we focused on SAMU 94, which covers the Val-de-Marne department (South-east of Paris). With a population of more than 1,300,000 inhabitants, this small department (with area of 245 km^2) is among the most populated areas in France.

Compared to other EMS systems worldwide, the specificity of the French SAMU system consists of involving physicians in the whole process of an emergency treatment, from the evaluation of emergency calls till the realization of rescue missions. The objective is to guarantee efficient assistance and high advance care to victims either on the phone or at the scene of accidents. However, a higher quality of service involves extended time to process calls and to perform on-scene treatments.

In order to build the SAMU 94 model, the first step was to fully understand the SAMU 94 vehicle dispatch and care delivery process, based on independent empirical observations and discussions with SAMU 94 experts. This process is graphically summarized in Fig. 2.1 that identifies its two main operations:

1. *Central operations*: These operations are performed in the reception and regulation (R&R) center. They include providing phone

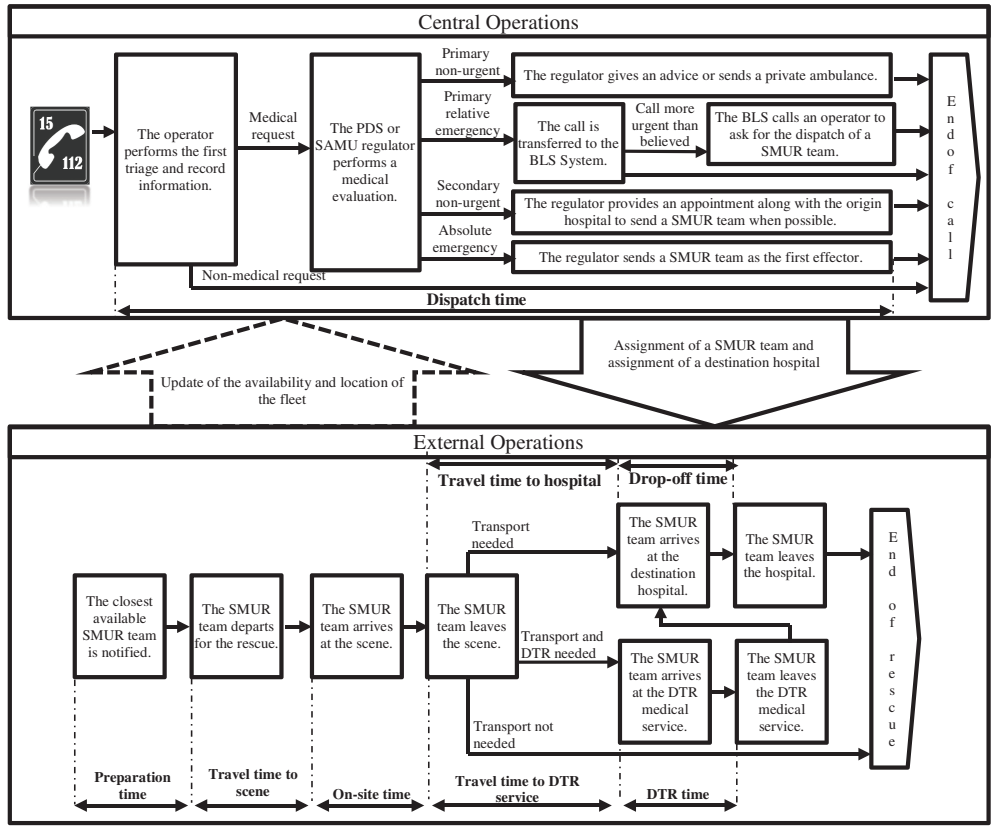


Figure 2.1. SAMU 94 process.

support and deciding the proper response for each emergency call received.

2. *External operations*: These operations consist of dispatching one or several mobile response vehicles, known as Mobile Emergency and Resuscitation Services (SMUR) teams, to perform either primary rescues that are related to major injuries or illnesses and require immediate medical assistance outside the hospital (e.g., cardiac arrest, trauma, and childbirth) or secondary rescues that correspond to the transport of patients from one hospital to another if medical staff assistance is required during the transfer.

Several types of human resources that are involved in central and external operations, such as operators, PDS regulators ("PDS" stands for the French acronym of "Permanent Care"), SAMU regulators, and rescue and SMUR teams, use two types of vehicles: mobile intensive care units (MICU) and medical vehicles (MV) (See Table 2.1).

When the R&R center gets calls for help, the operator performs the first triage to eliminate calls that are not medical requests and record basic information of the remaining calls. A medical evaluation of the calls is, then, performed by a regulator (an SAMU regulator for potentially high priority calls and a PDS regulator for other types of calls). This evaluation may lead to several decisions, such as:

- A simple advice is provided to the caller (in case of non-urgent primary calls).
- The call is transferred to basic life support (BLS) services, such as the fire department (primary relative emergency). After attending the scene, the BLS may call an operator to ask for a team because the incident is more urgent than originally believed (primary rescues by sending BLS as a first effector).
- A SMUR team is immediately dispatched for primay or secondary absolute emergencies (primary rescues by sending a SMUR team as a first effector).
- An appointment is planned to send a SMUR team in case of a non-urgent secondary call.

Table 2.1. SAMU 94 human resources.

	Role	Number in Weekdays	Number in Weekends
Operators	Answer calls Identify inappropriate calls Create medical files	4 (*9 p.m. to 7 a.m.*) 5 (*7 a.m. to 2 p.m.*) 6 (*2 p.m. to 9 p.m.*)	**4 (*9 p.m. to 7 a.m.*)** **5 (*7 a.m. to 2 p.m.*)** **6 (*2 p.m. to 9 p.m.*)**
PDS regulators (*General practitioners*)	Perform medical evaluation of low-priority calls	2 (*24 hours*)	**3 (*24 hours*)**
SAMU regulators (*Emergency physicians*)	Perform medical evaluation of high-priority calls	2 (*12:30 p.m. to 8 a.m.*) 1 (*8 a.m. to 12:30 p.m.*)	**2 (*24 hours*)**
Rescue	Operate either as SAMU regulators or as physicians on SMUR teams	1 (*8 a.m. to 5 p.m.*)	**1 (*8 a.m. to 5 p.m.*)**
SMUR teams located at central base (Henri-Mondor Hospital, HM)	Vehicles (MICU or MV) staffed by one physician, one driver, one nurse, and/or one emergency medical technician	5 (*12:30 p.m. to 7:30 p.m.*) 4 (*7:30 p.m. to 10:30 p.m.*) 3 (*10:30 p.m. to 10:30 a.m.*) 4 (*10:30 a.m. to 12:30 p.m.*)	**3 (*24 hours*)**
SMUR teams located at auxilliary base (Villeneuve-Saint-Georges, VSG)	Vehicles (MICU or MV) staffed by one physician, one driver, one nurse, and/or one emergency medical technician	1 (*24 hours*)	**1 (*24 hours*)**
Mobile intensive care units (MICU)	Well-equipped ambulances (can transport the patient)	5	**5**
Medical vehicles (MV)	Fast vehicles usually dispatched for the most serious calls (cannot transport the patient)	2	**2**

When a SMUR team needs to be dispatched, the regulator notifies the closest available team to perform the rescue, based on the proximity of the incident's address. If the SMUR team is located at the base, it spends some time preparing the rescue (which includes gathering the equipment and getting to the vehicle). After the arriving at the scene and accessing to the patient, the SMUR team performs the primary assessment and treatment and transports the patient, if needed, to the appropriate hospital that can provide further emergency care. Before getting transported to the destination hospital, the patient may need to be transported to a diagnostic or therapeutic radiography (DTR) service to perform an MRI or X-ray (if the destination hospital does not have the appropriate equipment or have long wait time). The SMUR team is dismissed after transferring the patient to the destination hospital and completing paperwork. It, then, returns to its home base and waits for the next mission. SMUR teams are currently located at two bases: one at the central base, located at the Henri-Mondor Hospital (HM), and another at the auxiliary base, located in the Villeneuve-Saint-Georges Hospital (VSG).

2.3.2. *Data collection and analysis*

The following call and rescue records of the SAMU 94 database, dated from October 1, 2010, to December 31, 2011, were collected and verified for the analysis:

1. The number of calls per hour of the day and day of the week
2. The type of call (without a dispatch of the SMUR team and primary/secondary rescue)
3. The first effector of primary calls (SMUR or BLS)
4. The priority of rescues, i.e., the classification of the rescue cause performed by the regulator on the phone (scale of 1–2) and by the SMUR team at the scene (scale of 0–3)
5. The location of the rescue within basic units of approximately 2,000 residents called IRIS (the French acronym for aggregated units for statistical information). There are 527 IRIS in the Val-de-Marne department).

6. The times for the following rescue segments: *dispatching time*, the interval between the time the call is received and the time a SMUR team is notified; *preparation time (PT)*, the interval between the time the SMUR team is notified and the time it leaves for the rescue; *on-site time (OST)*, the interval between the time the SMUR team arrives at the scene and the time it leaves the scene; *travel time to DTR service (TTD)*, the interval between the time the SMUR team leaves the scene and the time it arrives at the service; *diagnostic or therapeutic–radiography time (DTRT)*, the interval between the time the SMUR team arrives at the DTR medical service and the time it leaves the service; *travel time to hospital (TTH)*, the interval between the time the SMUR team leaves the scene or DTR service and the time it arrives at the hospital; and *drop-off time (DOT)*, the interval between the time the SMUR team arrives at the hospital and the time it leaves the hospital.

The Input Analyzer tool (Arena software) was used to choose the best-fitted distributions of the above data by using Kolmogorov-Smirnov and Chi-Square goodness-of-fit tests. Those tests provided low p-values. Therefore, we used empirical distributions to better capture the characteristics of the data [43]. The distributions of each frequency and activity time were fed into the corresponding distribution of the DES model.

Note that *travel time to scene (TTS)*, the interval between the time the SMUR team leaves for rescue and the time it arrives to scene, cannot be modeled using empirical distributions. Indeed, the DES model must include travel time data for currently unexplored road networks that could be used under alternative deployment strategies. Furthermore, the model must also consider changes in travel times that arise at various times of the day/week due to congestion levels and population activities. Hence, we used the shortest path algorithm to compute travel times for every possible origin, destination, period, and priority of calls. The origins and destinations correspond to the 527 IRIS of the Val-de-Marne area. The periods represented the degree of traffic load at various times of the day according to the six shifts that distinguish between weekdays

(6 a.m.–10 a.m., 10 a.m.–3 p.m., 3 p.m.–9 p.m., and 9 p.m.–6 a.m.) and weekends (12 p.m.–9 p.m. and 9 p.m.–12 p.m.). Based on the database of GPS traces of SAMU 94 vehicles, an average travel time per period was assigned to each section of the road network of the Val-de-Marne department, according to its typology (motorway, main road, minor road, and local street). We used this model to dynamically compute the shortest path for a combination of origin/destination IRIS and period whenever the simulation required such path. However, it turned out to be a time-consuming computation that considerably increase the simulation time. As a reasonable trade-off, we precomputed and stored the shortest path for any given combination of origin/destination IRIS and period using a sample of 10 pairs of the exact addresses that were randomly chosen within the two IRIS. For each pair, the travel time was computed by summing up the average travel times associated with the sections that form the shortest path between the two addresses. Finally, as SMUR teams can travel at all possible speed while responding to primary calls of priority 1, the related travel times were weighted by a regression factor estimated at 0.962 to decrease them compared to standard travel times. The resulting travel time matrices were used to compute TTS in the DES model and to choose the closest available SMUR team to assign to incoming calls.

2.3.3. *DES model design*

The previously described SAMU 94 rescue process and data were summarized in a written mathematical and logical representation of the system, known as conceptual model. This model was created and iteratively refined, based on discussions with SAMU 94 managers and physicians.

The conceptual model was computationally implemented using ARENA (Version 12, Rockwell Automation). This widely used DES software uses the SIMAN processor and simulation language for analysing diverse operation types (such as manufacturing, supply chain, healthcare, and military) and predicting system performance

under varying conditions and decision criteria. It has the following advantages:

1. It allows to capture process hierarchy, including activity-based costing and process logic.
2. It integrates the ease of use found in high-level simulators that provide graphical simulation modelling and analysis modules.
3. It provides high flexibility of simulation languages accessed in low-level modules and even general-purpose procedural languages, like Visual Basic or C/C++, to model any desired level of detail and complexity.
4. It can be integrated with other software, including reading from or outputting to spreadsheets and databases.
5. It combines process simulation with optimization technologies, using meta-heuristic analysis tools.

Figure 2.2 presents the first level of SAMU 94 Arena model, which is a sequence of blocks (flowcharts or data modules) and connectors through which entities (calls) move. Several *attributes* are assigned to entities in order to specify the characteristics of calls (such as, the first effector, priority, and IRIS). Flowchart modules are hierarchically organized using *submodels* that allow modular implementation of each part of the model separately for easier verification, better readability and maintainability, and less risk of errors. They include *create* and *dispose* blocks used as starting and ending points of each flowchart, flow-control blocks used to direct entities in the process and specify processing methods (such as sending calls of a given priority and effector to a dedicated process and performing on-site treatment for a given priority), and information import, export, and assignment blocks used to assign data values obtained from input files, assign modules to a list of variables or attributes, or write data to an output device (such as establishing the location of a call and writing order-response time distribution to a data file). The creation of rescues, the definition of variables (e.g., the matrix of bases' locations per period, vector of availability, and location of each vehicle),

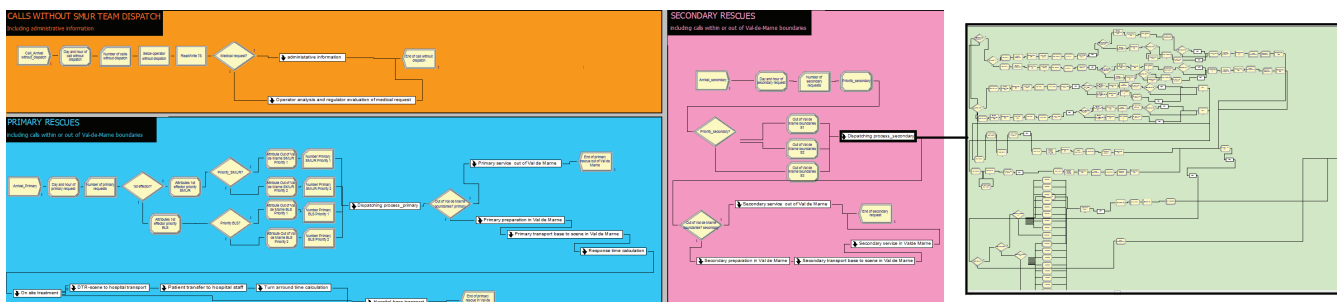


Figure 2.2. Overview of ARENA model for SAMU 94.

and the availability of resources, depending on their operating schedule, are controlled by data modules. Finally, integrating user-defined Visual Basic for Applications (VBA) functions in the model provided great flexibility in implementing dispatching rules, collecting various statistics per period, defining and assigning periods to calls, and reading travel time matrices (Microsoft Excel files) to obtain travel time for a given origin, destination, and period.

Different random number seeds were used to replicate the model 20 times in order to derive outcome variables. The replications' length corresponds to 15 months of operations with a warm-up period of one day. The warm-up period served to avoid any initialization bias in the estimate of the simulation steady-state parameters by filling the queues rather than starting with empty queues.

2.3.4. *DES model validation*

Model validation was performed with SAMU 94 specialists, who were asked to evaluate the conception (i.e., structural, logical, mathematical, and causal relationships) and output behavior of the model. They confirmed that the model ran the same way as the real-world system for the intended purpose of the study. Moreover, the DES model was validated by comparing the system's empirical input data (processing times and travel times) and output data (response times) per type, priority, and effector with the simulation-derived data. The results of this historical data validation showed that the outputs of the model were quite close to the observed distributions: the difference between the bounds of 95% confidence intervals (based on the 20 replications of the model) and the historical values range from 0% to 8.9%, with an average deviation of only 2.3%. This was a reasonable threshold to consider the DES model valid.

2.4. Analysis of DES Model Results

2.4.1. *Simulation strategy design*

Potential strategies were proposed by SAMU 94 managers to assess system performance under various changing conditions. Each strat-

egy considered different scenarios that are simulated, and the corresponding performances, resulting from the 20 replications of the model, were observed and compared to the initial scenario, which represents the current configuration of the system. Significant performance variables for SAMU 94 managers included coverage within 20 min for primary rescues with an SMUR team as the first effector and human ressources utilization rate (which is the total workload divided by the total operating time). The following strategies were analyzed in sequence:

- **Strategy A—Variation in the number of operators, regulators, and SMUR teams**: The objective of this first category of scenarios is twofold: First, to determine the effect of increasing number of resources during high-demand periods on the coverage performance in order to select the best trade-off between the additional cost of hiring staff and a better response time. Second, to assess the effect of reducing the number of different ressources on the utilization rate without decreasing the coverage performance within time periods that are consistent with personnel scheduling constraints.
- **Strategy B—Increase in the number of rescues**: In order to capture the sensitivity of SAMU 94 performance to factors such as demographic growth or aging of the population, we increased the arrival rate of rescues in the DES model from 20% to 100%, with steps of 20%. We also experimented increased demand scenarios by adding one more critical resource (SMUR teams) during the high-demand period 8 a.m.–8 p.m. in the HM base. The critical resource were determined as a result of analysing the waiting times for the assignment of different resources under increased demand scenarios.
- **Strategy C—Improved deployment of SMUR teams**: Besides the number of SMUR teams, their location was a decisive factor for designing efficient EMS systems that ensure equity of access to pre-hospital care throughout the entire service area. This strategy aimed to determine the effect of gradually relocating one to three SMUR teams initially located at the HM base in a decentralized

way to potential bases located throughout the Val-de-Marne department. In accordance with French legislation, potential bases correspond to three public hospitals located in the department are Saint-Camille (SC), Bicêtre (B), and Emile-Roux (ER) (Refer Fig. 2.3).

- **Strategy D—Increased travel times**: Based on historical data, we identified ten atypical days, characterized by exceptionally long travel times. These days correspond to sporting events, adverse weather conditions, or public transport strikes. Hence, three tendencies were observed corresponding to regular travel times increases of 100%, 150%, and 350%. This strategy is primarily concerned with quantifying potential benefits of relocation during these days. Therefore, we applied an increase in travel times to the initial scenario, as well as to optimal relocation scenarios obtained in Strategy C.
- **Strategy E—Alternative dispatching rule (regionalized response)**: In the current system, the closest available team is assigned to each call. In this strategy, we experimented a new dispatching rule known as regionalized response. It consists of assigning each SMUR team to serve a pre-specified service area. If the assigned SMUR team is busy, the closest available team must perform the rescue. This is a widely used dispatching rule in several EMS worldwide [20, 35, 39]. The advantages of this rule are the limited area the SMUR team has to traverse to reach call locations and the increased familiarity of the driver with the assigned zone.
- **Strategy F—Process improvement**: This strategy considers a decrease in the dispatching time, which may be driven by a number of actions, such as multitasking between regulation and preparation tasks and using more advanced technologies to perform call screening, medical file creation, or SMUR teams' notification.
- **Strategy G—Multiple changes**: The previous strategies have discussed several alternatives in isolation. However, a reasonable objective of the SAMU 94 system is to select a portfolio of strategies to maximize coverage performance at the minimal cost.

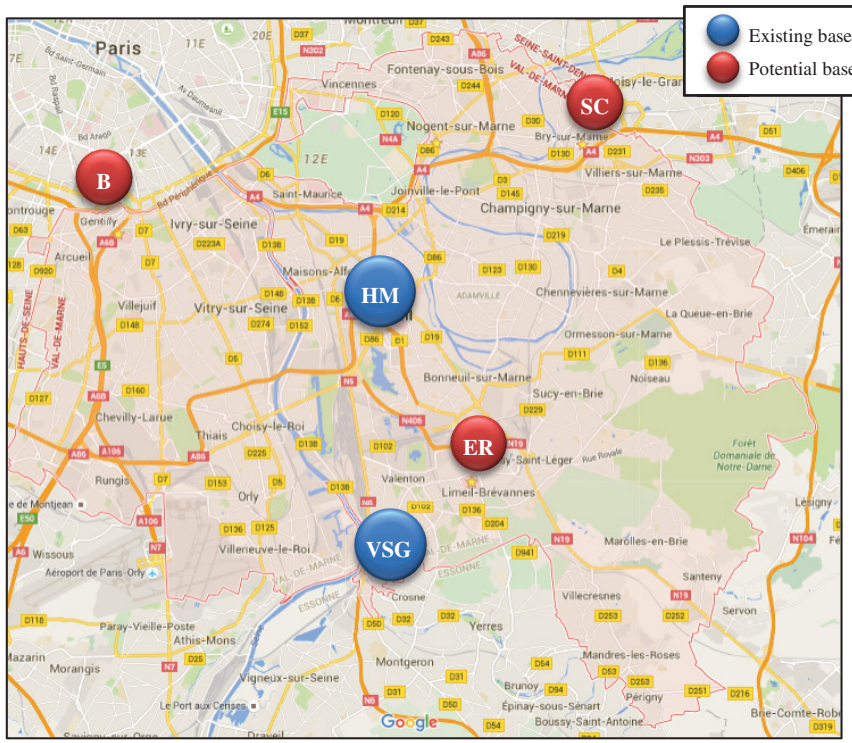


Figure 2.3. The existing and potential bases of the Val-de-Marne department.

Thus, we experimented several combinations of strategies, such as adding one more SMUR team (Strategy A) under decentralized deployment (Strategy C).

2.4.2. *Simulation results*

The comparison of results obtained from the simulation of scenarios described above allows establishing a hierarchy in advocating strategies to improve the SAMU 94 performances. Indeed, the most interesting strategy in terms of coverage will be the deployment of SMUR teams across the Val-de-Marne department (Strategy C).

Thus, according to simulation results obtained from this strategy, we recommend to adopt the following best deployment plan:

- Relocating two SMUR teams in Saint-Camille hospital and one SMUR team in Bicêtre hospital during weekdays. The corresponding average increase in coverage within 20 min between 8 a.m. and 8 p.m. will be 3.8% and 8.3% for priority 1 and 2, respectively.
- Relocating only one SMUR team in Saint-Camille hospital during weekends, which would lead to a 20-min coverage improvement of 5.3% and 3.1% for priority 1 and 2, respectively.

Implementing this deployment strategy is particularly appropriate during atypical days, with 100% increase in travel times, as the improvement in 20-min coverage can reach 10.9% on weekdays and 15% on weekends, compared to increased travel times under the current locations of SMUR teams (Strategy D).

Another strategy that seems promising to improve coverage performance is that of reducing dispatching time (Strategy F). Indeed, the DES model showed that a reduction of 40 seconds in this processing time would lead to a significant average improvement for 20-min coverage of 2.7% and 4.2% for priority 1 and 2, respectively. This suggests that similar improvements in other parts of the rescue process, such as preparation time and on-site time, may lead to additional efficiencies by reducing the response time and round-trip time (i.e., the period between the receipt of a call and the arrival of the SMUR team with the patient to the destination hospital) for an improved access to prompt professional medical treatment.

In contrast, less significant improvements in coverage performance will be achieved by varying the number of resources (Strategy A). Indeed, according to the DES model, an increase in the number of operators, PDS regulators, and SAMU regulators will have no significant effect on 20-min coverage, whereas adding one SMUR team will improve the 20-min coverage between 8 a.m. and 8 p.m.,

not exceeding 2.5% for both priorities 1 and 2. We, however, note that this latest coverage improvement is differently broken down by days of the week because it does not exceed 1.9% in weekdays but reaches 4.3% in weekends. Moreover, the impact of the additional SMUR team will be greater under some specific conditions, such as the decentralized deployment plan (Strategy G) and the increased demand for rescues (Strategy B) because the average increase in 20-min coverage performance will reach 9.1% when three SMUR teams are optimally relocated and 9.8% under 100% increase in demand. In the light of these observations, we recommend to focus on the recruitment of an additional SMUR team (as the system critical resource) only on weekends under the current configuration. However, if the demand increases or the recommended deployment plan has to be implemented, the recruitment of an SMUR team should be extended to all weekdays.

On the other hand, a reduction in the number of operators and PDS regulators seems to have a limited effect on coverage (Strategy A). This result suggests that the current coverage performance can be maintained by decreasing these resources. However, these savings are likely to be made at the expense of an increase in the stress level of the remaining resources and may lead to deteriorated quality of service and longer processing times, which may negatively impact the coverage performance.

Similarly, the DES model demonstrates that Strategy E of changing the dispatching rule into the alternative regionalized response has a limited effect on coverage performance under the current system configuration. If the recommended deployment strategy was adopted on weekdays, the implementation of the regionalized response would slightly improve the 20-min coverage performance by an additional 1.1% and 1.2% on average for priority 1 and 2, respectively. As a consequence of these limited benefits, we recommend not to change the current dispatching rule unless the regionalized response is judged to have operational advantages, compared to the closest available rule, such as easier implementation and reduced opportunity for errors.

2.5. Conclusions and Perspectives

In this study, we developed a discrete event simulation model in order to model and assess the current performance of the SAMU 94 system, as well as to investigate the effects of potential process changes that would lead to enhanced operational efficiency in terms of the target 20-min coverage performance of primary rescues. Several types of input data were used to specify model parameters. These data were obtained from historical records as well as interviews conducted with SAMU 94 managers. The developed DES model was validated using historical data and was used as a decision-support tool to compare the relative benefits of several strategies, mainly related to the needed resource levels and static location of rescue teams throughout the Val-de-Marne area and their assignment to incoming calls. Sensivity analyses were also performed by changing values of some input parameters, such as arrival rates of calls, travel times, and service times. Based on the results obtained, our recommendation for practitioners is to first focus on the optimal relocation of SMUR teams, which would significantly improve coverage with a minimal increase in costs. In addition, we recommend the recruitment of SMUR teams as the most critical resources of the system, particularly under conditions such as increased demand. Finally, we recommend additional studies to improve the dispatch and rescue process by removing non-value-added steps, such as duplicate processes and unnecessary procedures, based on principles such as the principles of the lean approach, which can be deployed in healthcare. Robinson *et al.*, 2012, [45] give some examples of such healthcare wastes as asking for patients' details several times, storing the frequently used equipment centrally instead of where it is used, or asking for unnecessary tests.

One main limitation of the current study is that relocation scenarios are based on the historical demand data of SAMU 94. One possible extension of this work can be devoted to the development of forecast models in order to predict the number and location of rescues and, therefore, to derive sufficiently robust relocation strategy of

SMUR teams that covers future demand at the desired service level. Another limitation is that financial aspects, including capital and operating costs, are not considered. Aboueljinane *et al.* [18] highlight the importance of performing a cost-effectiveness analysis in EMS simulation studies by comparing the costs of each alternative and the obtained increase in coverage to achieve the desired objectives at a low cost.

In the current study, the cost factor is not included due to the lack of detailed cost components associated with each studied strategy to support the analysis. Future work may include the enlargement of the scope to consider other pre-hospital care providers operating in the department (e.g., BLS, private ambulances, home cares, and general practitioners). Hence, modeling this integrated multi-facility system may greatly enhance the overall quality of care performance in the area. Finally, the deployment strategies discussed in the what-if analysis consider a fixed location of SMUR teams, regardless of the daily fluctuation in the volume and location of rescues. The simulation–optimisation technique can be used to determine an efficient multi-period redeployment plan that can further improve the coverage performance of SAMU 94.

References

1. Inoue, H, S Yanagisawa and I Kamae (2006). Computer-simulated assessment of methods of transporting severely injured individuals in disaster—case study of an airport accident. *Comput Methods Programs Biomed*, 81, 256–265.
2. Sacco, WJ, DM Navin, KE Fiedler, RK Waddell, WB Long and RF Buckman (2005). Precise formulation and evidence-based application of resource-constrained triage. *Acad Emerg Med*, 12, 759–770.
3. Cummins, RO (1989). From concept to standard-of-care? Review of the clinical experience with automated external defibrillators. *Ann Emerg Med*, 18, 1269–1275.
4. McLay, LA and ME Mayorga (2010). Evaluating emergency medical service performance measures. *Health Care Manag Sci*, 13, 124–136.
5. Sánchez-Mangas, R, A García-Ferrrer, A de Juan and AM Arroyo (2010). The probability of death in road traffic accidents. How important is a quick medical response? *Accid Anal Prev*, 42, 1048–1056. doi:10.1016/j.aap.2009.12.012

6. Vukmir, RB (2006). Survival from prehospital cardiac arrest is critically dependent upon response time. *Resuscitation*, 69, 229–234. doi:10.1016/j.resuscitation.2005.08.014
7. White, R, B Asplin, T Bugliosi and D Hankins (1996). High discharge survival rate after out-of-hospital ventricular fibrillation with rapid defibrillation by police and paramedics. *Ann Emerg Med*, 28, 480–485. doi:10.1016/S0196-0644(96)70109-9
8. American Heart Association (2005). 2005 American Heart Association Guidelines for Cardiopulmonary Resuscitation and Emergency Cardiovascular Care. *Circulation*, 112, IV–1–IV–211. doi:10.1161/CIRCULATIONAHA.105.166550
9. Zeng, Z, X Ma, Y Hu, J Li and D Bryant (2012). A simulation study to improve quality of care in the emergency department of a community hospital. *J Emerg Nurs JEN Off Publ Emerg Dep Nurses Assoc*, 38, 322–328. doi:10.1016/j.jen.2011.03.005
10. Tako, AA and S Robinson (2010). Model development in discrete-event simulation and system dynamics: An empirical study of expert modellers. *Eur J Oper Res*, 207, 784–794. doi:10.1016/j.ejor.2010.05.011
11. Holm, LB, H Lurås and FA Dahl (2013). Improving hospital bed utilisation through simulation and optimisation: With application to a 40% increase in patient volume in a Norwegian General Hospital. *Int J Med Inf*, 82, 80–89. doi:10.1016/j.ijmedinf.2012.05.006
12. Morgareidge, D, H CAI and J JIA (2014). Performance-driven design with the support of digital tools: Applying discrete event simulation and space syntax on the design of the emergency department. *Front of Archit Res*, 3, 250–264. doi:10.1016/j.foar.2014.04.006
13. Baril, C, V Gascon, J Miller and N Côté (2016). Use of a discrete-event simulation in a Kaizen event: A case study in healthcare. *Eur J Oper Res*, 249, 327–339. doi:10.1016/j.ejor.2015.08.036
14. Harper, PR, AK Shahani, JE Gallagher and C Bowie (2005). Planning health services with explicit geographical considerations: A stochastic location–allocation approach. *Omega*, 33, 141–152. doi:10.1016/j.omega.2004.03.011
15. Aboueljinane, L, E Sahin, Z Jemai and Marty (2014). A simulation study to improve the performance of an Emergency Medical Service: Application to the French Val-de-Marne Department. *Simul Model Pract Theor*, 47, 46–59.
16. Henderson, SG and AJ Mason (2005). Ambulance service planning: Simulation and data visualisation. In *Operations Research and Health*

Care, ML Brandeau, F Sainfort and WP Pierskalla (eds.), pp. 77–102. Boston: Kluwer Academic Publishers.

17. Savas, ES (1969). Simulation and cost-effectiveness analysis of New York's emergency ambulance service. *Manag Sci*, 15, 608–627. doi:10.1287/mnsc.15.12.B608
18. Aboueljinane, L, E Sahin and Z Jemai (2013). A review on simulation models applied to emergency medical service operations. *Comput Ind Eng*, 66, 734–750. doi:10.1016/j.cie.2013.09.017
19. Aboueljinane, L, Jemai Z and Sahin E (2012). Reducing ambulance response time using simulation: The case of Val-de-Marne department emergency medical service. In *Proc of the 2012 Winter Simulation Conference*, Edited by C Laroque, J Himmelspach, R Pasupathy, O Rose and AM Uhrmacher. Piscataway, New Jersey: IEEE.
20. Gunes, E and Szechtman R (2005). A simulation model of a helicopter ambulance service. In *Proc of the 2005 Winter Simulation Conference*, ME Kuhl, NM Steiger, FB Armstrong and JA Joines (eds.), pp. 951–957. Piscataway, New Jersey: IEEE.
21. Ingolfsson, A, E Erkut and S Budge (2003). Simulation of single start station for edmonton EMS. *J Oper Res Soc*, 54, 736–746.
22. Inakawa, K, T Furuta and A Suzuki (2010). Effect of ambulance station locations and number of ambulances to the quality of the emergency service. In *The 9th International Symposium on Operations Research and Its Applications (ISORA'10)*, pp. 340–347. Chengdu-Jiuzhaigou, China.
23. Henderson, SG and AJ Mason (2005). Ambulance service planning: Simulation and data visualisation. In *Operations Research for Healthcare*, ML Brandeau, F Sainfort and WP Pierskalla (eds.), pp. 77–102. Boston: Kluwer Academic Publishers.
24. Fitzsimmons, JA. An emergency medical system simulation model. In *Proc of the 1971 Winter Simulation Conference*, pp. 18–25. ACM, New York.
25. Goldberg, J, R Dietrich, JM Chen, M Mitwas, T Valenzuela and EA Criss (1990). A simulation model for evaluating a set of emergency vehicle base locations: Development, validation, and usage. *Socio-Econ Plan Sci*, 24, 125–141.
26. Berlin, GN and JC Liebman (1974). Mathematical analysis of emergency ambulance location. *Socio-Econ Plan Sci*, 8, 323–328.
27. Uyeno, DH and C Seeberg (1984). A practical methodology for ambulance location. *Simulation*, 43, 79–87.

28. Aringhieri, R, G Carello and D Morale (2007). Ambulance location through optimization and simulation: The case of Milano urban area. In *38th Annual Conference of the Italian Operations Research Society Optimization and Decision Sciences,* 1–29.
29. Fujiwara, O, T Makjamroen and KK Gupta (1987). Ambulance deployment analysis: A case study of Bangkok. *Eur J Oper Res*, 31, 9–18. doi:10.1016/0377-2217(87)90130-5
30. Harewood, SI (2002). Emergency ambulance deployment in Barbados: A multi-objective approach. *J Oper Res Soc*, 53, 185–192. doi:10.1057/sj.jors.2601250
31. Lee, T, H Jang, S-H Cho and JG Turner. A simulation-based iterative method for a trauma center: Air ambulance location problem. In *Proc of the 2012 Winter Simulation Conference*, C Laroque, J Himmelspach, R Pasupathy, O Rose and AM Uhrmacher (eds.). Piscataway, New Jersey: IEEE. doi: 10.1109/WSC.2012.6465042
32. Trudeau, P, J-M Rousseau, JA Ferland and J Choquette (1989). An operations research approach for the planning and operation of an ambulance service. *Inf Syst Oper Res*, 27, 95–113.
33. Peleg, K and JS Pliskin (2004). A geographic information system simulation model of EMS: Reducing ambulance response time. *Am J Emerg Med*, 22, 164–170. doi:10.1016/j.ajem.2004.02.003
34. Van Buuren, M, K Aardal, R Van der Mei and H Post (2012). Evaluating dynamic dispatch strategies for emergency medical services: Tifar simulation tool. In *Proc of the 2012 Winter Simulation Conference*, Edited by C Laroque, J Himmelspach, R Pasupathy, O Rose and AM Uhrmacher. Piscataway, New Jersey: IEEE.
35. Repede, JF and JJ Bernardo (1994). Developing and validating a decision support system for locating emergency medical vehicles in Louisville, Kentucky. *Eur J Oper Res*, 75, 567–581.
36. Gendreau, M, G Laporte and F Semet (2006). The maximal expected coverage relocation problem for emergency vehicles. *J Oper Res Soc*, 57, 22–28.
37. Maxwell, MS, SG Henderson and H Topaloglu (2009). Ambulance redeployment: An approximate dynamic programming approach. In *Proc of the 2009 Winter Simulation Conference*, Edited by MD Rossetti, RR Hill, B Johansson, A Dunkin and RG Ingalls, pp. 1850–1860. Piscataway, New Jersey: IEEE.
38. Koch, O and H Weigl (2003). Modeling ambulance service of the Austrian Red Cross. In *Proc of the 2003 Winter Simulation*

Conference, Edited by S Chick, PJ Sánchez, D Ferrin and DJ Morrice, pp. 1701–1706. Piscataway, New Jersey: IEEE.

39. Su, S and CL Shih (2003). Modeling an emergency medical services system using computer simulation. *Int J Oper Inform*, 72, 57–72.
40. Wears, RL and CN Winton (1993). Simulation modeling of prehospital trauma care. In *Proc of the 1993 Winter Simulation Conference*, Edited by GW Evans, M Mollaghasemi, EC Russel, WE Biles. Piscataway, New Jersey: IEEE.
41. Wang, Y, KL Luangkesorn and L Shuman (2012). Modeling emergency medical response to a mass casualty incident using agent based simulation. *Socioecon Plann Sci*, 46, 281–290. doi:10.1016/j.seps.2012.07.002
42. Baldwin, LP, T Eldabi and RJ Paul (2004). Simulation in healthcare management: A soft approach (MAPIU). *Simul Model Pract Theory*, 12, 541–557. doi:10.1016/j.simpat.2004.02.003
43. Kelton, WD, RP Sadowski and DT Sturrock (2008). *Simulation with Arena*, 4th Ed. New York: McGraw-Hill.
44. Law, AM and MG McComas (2001). How to build valid and credible simulation models. In *Proc of the 2001 Winter Simulation Conference*, Edited by BA Peters, JS Smith, DJ Medeiros and MW Rohrer, pp. 22–29. Piscataway, New Jersey: IEEE. doi:10.1109/WSC.2001.977242
45. Robinson, S, ZJ Radnor, N Burgess and C Worthington (2012). SimLean: Utilising simulation in the implementation of lean in healthcare. *Eur J of Oper Res*, 219, 188–197. doi:10.1016/j.ejor.2011.12.029

3. Modeling and Simulation of the Emergency Department of an Italian Hospital

Wanying Chen*, Alain Guinet† and Tao Wang‡

*Zhejiang Gongshang University, 310018, Hang Zhou, China
†INSA Lyon, 69621, Villeurbanne, France
‡University of Saint Etienne, 69621, Villeurbanne, France

Abstract

As the first or the second step of the admission process, the emergency department (ED) is one of the most important departments in the hospital. The very large scale of the case mix of pathologies makes the process of the ED's activities very complicated. Though a lot of research is devoted to this area, most studies are based on academic assumptions such as the assumption of independent resources or predefined pathways, and, therefore, research results are difficult to apply in the real world. This paper studies the ED of a large-sized Italian hospital in the normal situation and in the major accident situation (overcrowding situation), based on real cases. First, IDEF0 method is used to build the conceptual models to present the process of the activities of an ED in a normal situation as well as in the major accident situation. Then,

SIMIO is used to simulate the conceptual models in detail. In the part of experiments, factorial design is used to analyze the impact of resource dimensioning related to the total process time to treat all patients in the ED in different situations. With the help of the simulation model, one method is proposed to improve the current system. The rule to change the process of ED from normal situation to main influx situation is also defined.

3.1. Introduction

One of the most common problems for ED is overcrowding [1]. According to the experts at the American College of Emergency Physicians, overcrowding refers to a situation when the available institutional resources are insufficient to meet the basic service needs of emergency patients. In other words, overcrowding represents an obstacle to the safe and timely delivery of health care [2]. Overcrowding of ED, occurring in almost all big hospitals around the world, has led to crisis problems. The ability of an ED to deal with the suddenly increasing requirement of medical resources has been questioned by numerous researchers [3]. In the past few years, more and more EDs got exposed to the risks of meeting the overcrowding situation caused by terrorist attacks (such as the terrorist attacks on Paris on 13 November 2015). This led to hundreds of victims to be treated in EDs. To deal with this, care providers, administrators, and policy makers of EDs must find an efficient solution to handle a sudden afflux of patients injured by terrorist attacks. This highlights at least two major problems: first, health workers should know the process of ED activities clearly. This can help the workers know what needs to be done in a particular kind of situations. Second, since overcrowding always leads to the shortage of resources, decision-makers should have a thorough understanding of the effects of resource dimensioning on the number of people who can be received and the effective use of these limited resources to alleviate overcrowding.

A substantial body of literature focused on the study of EDs with the intent of managing the overcrowding problem in a better way.

But a lot of studies were theoretical, not based on real cases. For example, literature reviews that studied EDs did not follow the change of activity processes in the EDs or did not considered the teams assigned to activities. Moreover, due to the lack of a standard definition of overcrowding, some papers could not distinguish the study of EDs between normal and overcrowding situations. Both of the above research gaps are likely to be addressed by using our computer simulation model of ED. This paper studies an ED of a large-sized hospital in Italy in normal and main influx (overcrowding) situations, with the aim of formalizing, modeling, and improving the activity process.

The remaining part of this paper is organized as follows: Section 2 reviews the related literature briefly. Section 3 presents the stakeholders, the context, and the activity in ED, after describing the problem. Section 4 is devoted to our simulation models. Section 5 shows simulation experiments, analyzes the results, and proposes some suggestions for hospitals.

3.2. Literature Review

Many researchers [4–6] claimed that simulation is an effective and useful tool to study complex problems in ED. As we mentioned before, overcrowding, as a thorny problem in ED, can lead to a lot of problems, such as long waiting time and patients' dissatisfaction. To overcome these problems, many researchers studied EDs from different angles — such as employees' schedules [7, 8], patients' flow [9–13], and activity process [10, 14, 15] — with different objectives — such as to minimize the waiting time [7], to maximize patient throughput [13], and to optimize resource utilization [16]. This section presents a brief review of these papers.

Both Yeh and Lin [7] and Rossetti *et al.* [8] integrated simulation with other methods to study staff schedules. Yeh and Lin [7] used simulation and genetic algorithm to study nurse schedules. This simulation is developed to present patient flow through ED. With the aim of minimizing patients' waiting time, a genetic algorithm is applied to find an optimal nurse schedule. Rossetti *et al.* [8] used

computer simulation to test how different physician staff schedules in an ED impact patient throughput and resource utilization.

Hoot *et al.* [9] and Medeiros *et al.* [11] developed a discrete event simulation to study patients' flow in an ED. Brenner *et al.* [10] studied patient flow in the ED at the University of Kentucky Chandler Hospital with simulation. Both optimal numbers of human and equipment resources (i.e., nurses, physicians, and radiology technology) were investigated. Hung *et al.* [12] studied the patients' flow in a pediatric hospital to optimize resource utilization in the pediatric ED. Ahmed and Alkhamis [13] connected simulation with optimization to determine the optimal number of doctors, lab technicians, and nurses required to maximize patient flow and to reduce patients' stay time in an ED unit.

Ruohonen *et al.* [14] and Komashie and Mousavi [16] simulated the activity process in an ED. Ruohonen *et al.* [14] presented a simulation model to describe operations in the ED at the Central Hospital of Jyvaskyla, Finland. This simulation model can be used to test different process scenarios, allocate resources, and perform activity-based cost analysis. The simulation model developed by Komashie and Mousavi [16] was devoted to helping ED managers understand the hidden causes of excessive waiting time. This simulation served as a tool for assessing the impact of major departmental resources on key performance indicators and was also used as a cost-effective method to test various what-if scenarios for possible system improvement.

Since the activities of an ED vary in different countries, some authors studied EDs in specific countries. Duguay and Chetouane [17] described a simulation study on the ED of a Canadian hospital. The objective of this study was to reduce the patients' waiting time and to improve overall service delivery and system throughput. As the patients' waiting time is linked to resource availability, a number of alternatives were designed based on adding resource scenarios. Zeng *et al.* [18] presented how to use simulation to improve the quality of care of the ED at a community hospital in Lexington, Kentucky. The simulation model can evaluate the quality of care in terms of length of stay and waiting time. It was validated by experimenting with the data collected in the ED.

To sum up, more and more researchers use simulation to analyze ED to overcome the overcrowding problem. However, most of the research works were not based on a real scenario. Moreover, a lot of research works seem to mix overcrowding and normal situations, which made the study results difficult to put into practice due to different process and resource allocation. We have tried filling these gaps by using simulation approach to study the ED of an Italian hospital in normal as well as in overcrowding situation, caused by potential terrorist attack.

3.3. Problem Description

3.3.1. *Stakeholders*

The San Raffaele Hospital (OSR) is a university hospital located in Milan, Italy. It is spread across 300 thousand square meters and is composed of 11 buildings, with 49 specialty clinics and over 6,000 employees.

OSR is 9.3 km away from Milan Linate Airport (LIN), which served 9,031,855 passengers in 2014. The international airport is close to the center of Milan, only 7 km to the east of the city center, and is used as a base by Alitalia and Alitalia City Liner. It is highly likely to be attacked by terrorists, and the victims are likely to be admitted in OSR, due to its proximity.

3.3.2. *The context*

According to health workers in OSR, the activities of the ED follow different processes in different situations. Therefore, we intend to study activities of the ED in the normal situation, as well as in a major accident situation. The major accident situation we intend to study is provoked by a hypothetical terrorist attack on LIN. After the terrorist attack, the emergency management plan (EMP) is activated immediately. After discussing with the people working in OSR, we got to know that at the most 100 patients can be admitted by the ED at the event of triggered EMP. Meanwhile the EMP

activates, the ED will change its regular activity process (activity process in the normal situation) to the special activity process (the process of activities in the major accident situation). When the EMP is activated, the ED will switch to the special activity process (which is modelled by EMP) and will only accept the patients who are hurt by the terrorist attack on LIN.

3.3.3. *Process of ED activities*

The activities of the ED in the normal situation and in the major accident situation were modelled using the method IDEF0, based on the real practice. The IDEF0 method helped us elaborate a conceptual model, which offered a generic picture of the current situation.

IDEF0 is an IEEE Standard that is derived from the graphical language Structured Analysis and Design Technique (SADT) [19]. It is a method designed to model the events, data, and activities of an organization or a system. The structure of IDEF0 can help us model the existing system (as-is system) to understand its activities clearly. The standard can describe even a complex system easily [20]. Also, it can promote good communication between the analyst and the users.

The display of the IDEF0 model is based on a simple syntax. Each activity is described by a box. Inputs are shown as arrows entering the left side of the activity box, while the output is shown as exiting arrows on the right side of the box. Controls are displayed as arrows entering from the top of the box, and mechanisms are displayed as arrows entering from the bottom of the box. Fig. 3.1 presents the main features of IDEF0 with regard to analyzing the system by a collection of hierarchically organized diagrams with a limited number of elements: boxes, which represent activities, and arrows to model physical, information, and order flows.

The conceptual model can present the activity process in the ED clearly. Therefore, people can get a general idea of the sequence of events and activities to be performed by our IDEF0 model. In other words, the IDEF0 model can formalize the activity process of the ED

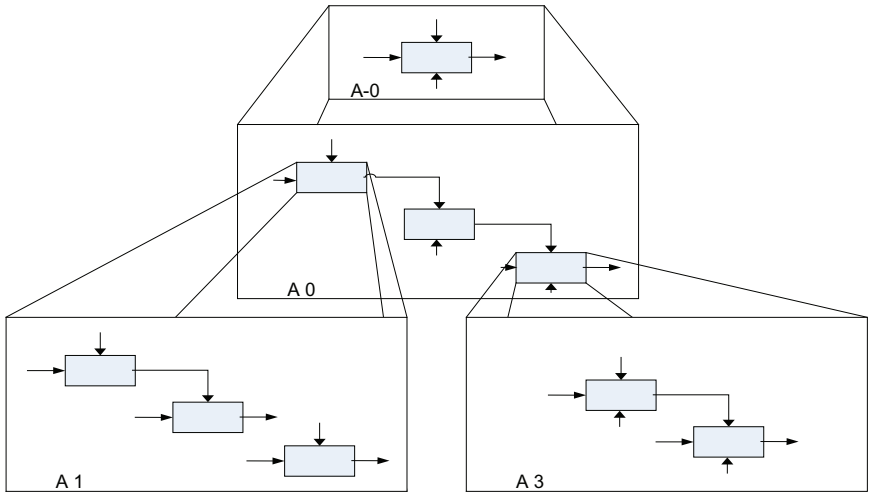


Figure 3.1. Hierarchical decomposition in IDEF0 method.

and enable people to know clearly what they should do in a certain situation.

Figures 3.2 and 3.3 show the activity process of the ED in the normal and the major accident situations, respectively. Figure 3.2 shows the normal situation, in which the patients should first finish the admission process after admitting in the ED. Then, according to the Emergency Severity Index, they will be triaged and will receive clinical treatments. In other words, after triage, patients will go to one of the three following treatment units: the shock room, the surgical area, or the medical area. The patients who are triaged to the shock room first will be transferred to the surgical room or the medical area for further treatment later. After being treated in the surgical area or the medical area, patients will either go to their home or be transferred to the related wards.

The hospital will change the activities of the ED if it experiences an influx of patients from outside (Fig. 3.3). In the major accident situation, when patients come to the ED, they will be triaged directly in the admission room to save time, and then, to other areas

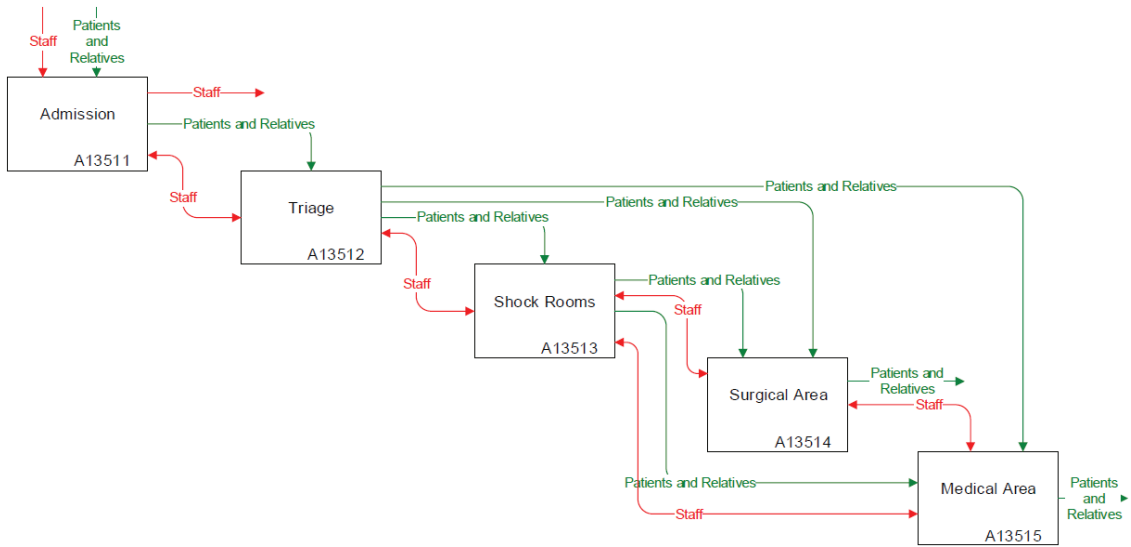


Figure 3.2. Activity process of ED in normal situation.

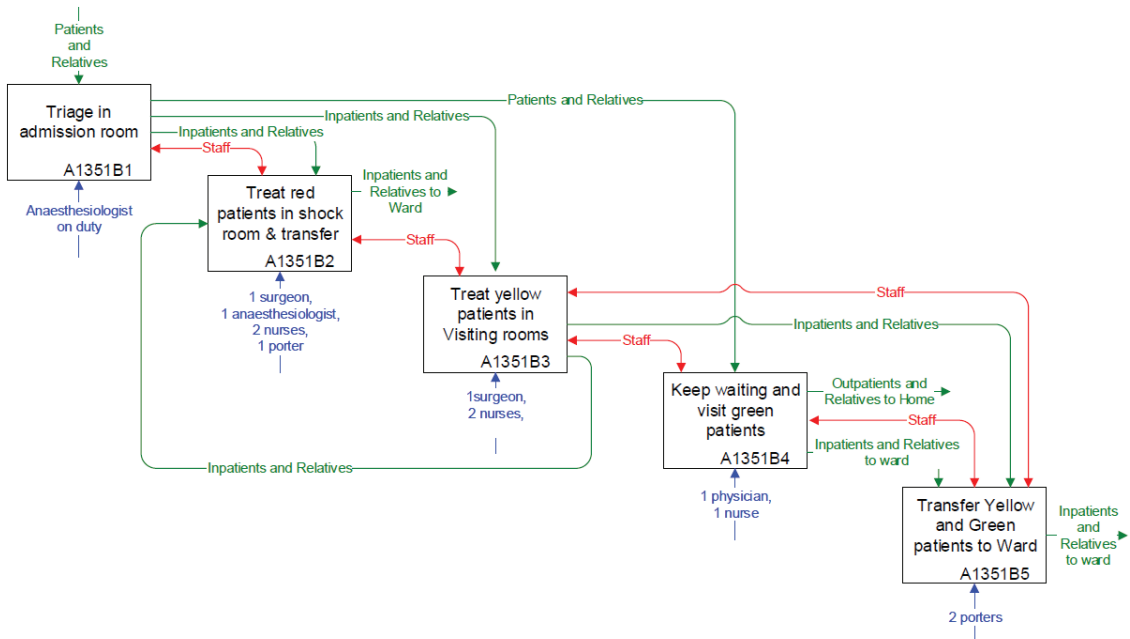


Figure 3.3. Activity process of ED in major incident situations.

according to their health condition. The red patients are patients with major life-threatening injuries and must receive treatments immediately. The yellow patients are the patients with major injuries whose treatments can be delayed until a given deadline. The green patients are the patients with minor injuries whose treatment can be delayed. The treatment of red patients consists of two parts: one includes the first assessment and stabilization and another covers additional tests, such as CT scan. After the initial treatment, the red patients will be transferred to an ICU or OT directly. After the treatment in the ED, part of yellow and green patients should be transferred to the related wards on stretchers.

3.4. Simulation Model

Based on our literature review, computer simulation seems to be one of the most suitable methods to deal with problems related to complex and uncertain real-world situations [21]. Computer simulation can capture the stochastics nature of data explicitly [22]. It can help anticipate different configurations and parameter settings of the systems easily, which is useful for complex problems.

Among different computer simulation packages, we chose SIMIO to build our simulation model. Oriented to agent-based modeling, SIMIO can support the object modeling paradigm and both discrete and continuous systems. It is also powerful at 3-D simulation, which is useful for presenting the simulation executions and results. The user-friendly interface of SIMIO allows novices to develop complex model quickly and simplifies the modelling tasks.

After selecting the simulation tool, we built our simulation model based on the IDEF0 model, proposed previously. This simulation model presents the activities based on the IDEF0 model in detail, considering stochastic characteristics. Our intention was neither to give a detailed description of the simulation model nor to get into all technical aspects. Our aim was to show how the simulation model can present an uncertain world taking into account the stochastics data. In the following paragraphs, we present the main activities in both situations.

3.4.1. *Activities of ED in the normal situation*

The activity is represented by system blocks called servers. A server includes a queue where entities are stocked before being processed. The process time and the processing capacity can be constant or stochastic, as chosen by the user. Several processors can be activated in a server at the same time.

Figure 3.4 presents the simulation model of the activities of the ED in the normal situation. The activities consist of five servers, which represent five activities. An entity is used to present a patient. The first activity, which is represented by the first server, simulates the patients' admission. The second server represents the activity of triage. After proper triage, the patients will be admitted in the shock room, the surgical area, or the medical area, according to the medical intervention they need. According to the data from OSR, in the normal situation, about 20 patients visit the ED in morning (from 8 a.m. to 12 a.m.). In the real-life, the arrival pattern of patients is always stochastics. Therefore, we suppose the interval time of patients' arrival follows the exponential distribution [23]. Some 20% of 20 patients should be transferred to the shock room. As much as 30% and 50% of them should be transported to the surgical area and the medical area, respectively. Therefore, at the output point of the server "triage", a routing logic function, which can

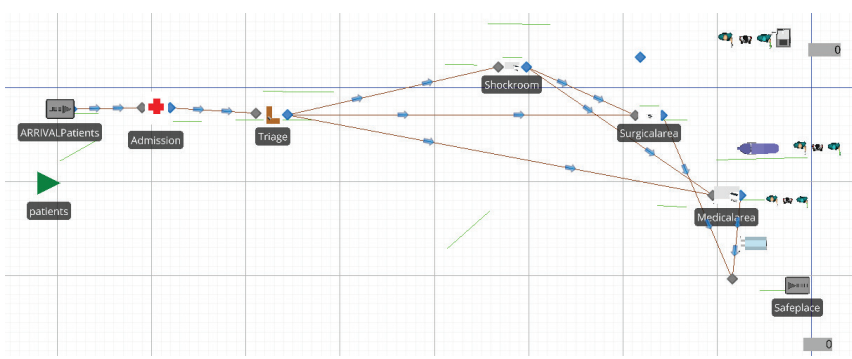


Figure 3.4. Simulation model of activities of an ED in the normal situation.

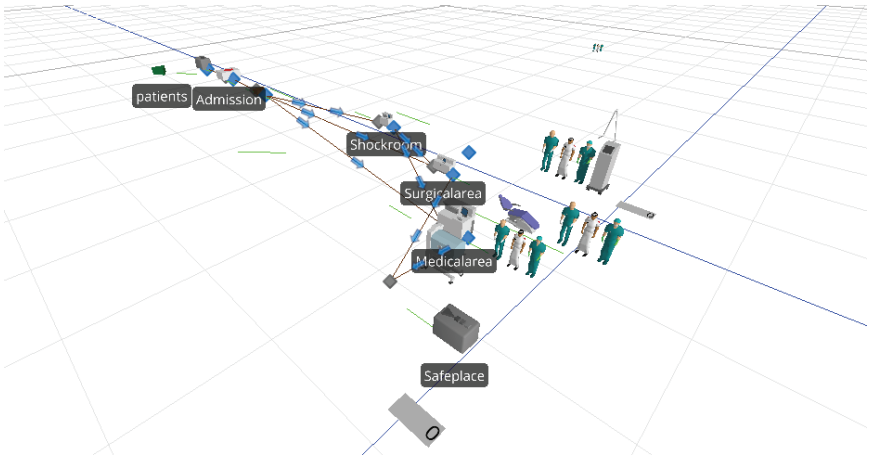


Figure 3.5. Simulation model of activities of ED in the normal situation (3-D effect).

classify the patients according to different weights, has been trigged. Three parallel servers simulate the treatment in the shock room, the surgical area, and the medical area. Two paths have been used to connect the transition between the shock room and other two rooms. Figure 3.5 shows the 3-D effect of this model.

As mentioned before, our purpose was not to show the detail of the whole situation but to present our tool. Taking into account the warm up period, we simulated the situation of the ED from 7 a.m. to 12 a.m. We had defined one statistics element to count the number of patients and used a monitor to alarm the processor when the number of patients would reach 20. Once the number of patients reached 20, a process was used to stop this model.

3.4.2. *Activities of ED in the major accident situation*

Figure 3.6 shows the activities of the ED in the major accident situation. Two vehicles are used to model the two porters of stretchers. A vehicle is an object that can be used to define a dynamic population

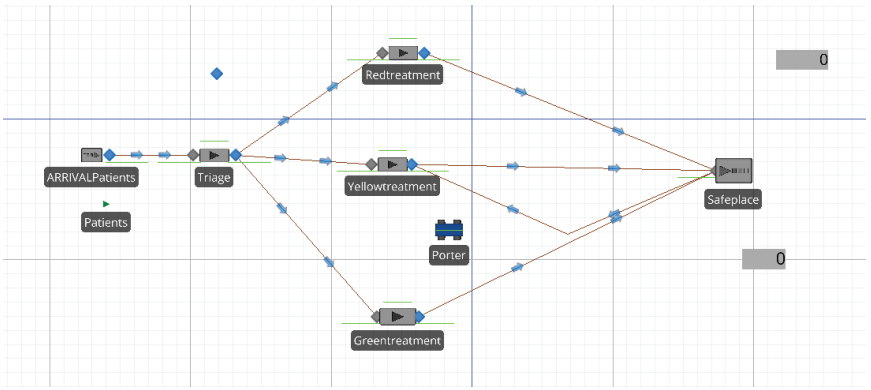


Figure 3.6. Simulation model of activities of the ED in the major accident situation.

of moveable unit resources. It can pick up entity objects at a location, carry those objects through a network of links or free space, and then drop the entities off at a destination location. Therefore, it is used to present porters. Again, at the output point of the server "triage", a routing logic function is used to classify the patients.

In this simulation model, four servers are used to deal with the following four activities: triage, treatment for red patients, treatment for yellow patients, and treatment for green patients. The output of servers "yellow treatment" and "green treatment" is also designed as the starting point of the porters. Again, statistics element and monitor have been used to limit the number of patients. Three processes have been used to adjust the color of patients in order to present the 3-D effect of the SIMIO model. In this way, this simulation model will be much easier to be understood.

3.5. Simulation Experiments

The purpose of this part is to assess the total process time to treat all patients in the normal situation and in the major accident situation

and to evaluate how the total process time evolves according to significant changes in model parameters or system settings. Concretely, the behavior of the simulation model can be evaluated through the comparison of the total process time when human resources (e.g., the number of medical teams) dedicated to EMP and the process time to complete activities (e.g., the time of the triage) vary. The process time to treat all patients of the ED can be predicted by our simulation models. With the help of our simulation results, decision makers can get a comprehensive idea about the impacts of resource dimensioning on the total process time of completing main activities and of the time used to treat all patients of ED.

Normally, nine steps are necessary for simulation experiments [24]: (1) problem description, (2) setting of objectives and overall project plan, (3) model conceptualization, (4) model translation, (5) data collection, (6) validation and verification, (7) experimental design, (8) execution and results analysis, and (9) documenting and reporting. Steps (1)–(4) have already been discussed in the above paragraphs. The following sections pay attention to steps (5)–(7).

3.5.1. *Data collection*

The effectiveness of the results of simulation depends on the accuracy of input data [25]. Therefore, it is important to select input parameters that are suitable for the current scenario. Real data in OSR indicates the presence of seven health teams usually. In the normal situation, two medical teams are in-charge of the shock rooms, three medical teams take care of the surgical area, and two medical teams handle the medical area. In the major accident situation, four medical teams take care of red patients, two medical teams take the responsibility for yellow patients, and one medical team look after green patients. In addition, two porters are involved in the transportation of yellow and green patients from the ED to the related wards, one for yellow and one for green. However, as one of the most important steps, collecting data for the process time of activities, especially for the major accident scenario, is rather difficult due to the lack of historical records about the activities of the ED.

Since the triangular distribution is often used where the distribution is only vaguely known, we invited the health care workers of OSR to estimate the upper process time, the lower process time, and the mode for each activity and, then, chose the triangular distribution to present the process time of each activity.

3.5.2. *Validation and verification*

There are two sub-steps for this step: validation of the conceptual model and verification of the simulation model. Validation helps in determining if the theories and assumptions underlying the conceptual model are correct and the representation of the problem, the model's logic structure, and mathematical and causal relationships are reasonable [26]. We presented our IDEF0 model to the health workers of OSR and, then, our conceptual model was well-validated.

Verification of a simulation model is performed to ensure that the programming of computer simulation and the implementation of the conceptual model are correct. After the thorough verification and debugging of the SIMIO model, we presented our simulation results to experts from the hospital. To treat 20 patients in the normal situation, 4.6653 hours would be needed. On average, patients spend 48 minutes in the ED. Since human resources are enough, on average, the waiting time of patients would not be more than 5 minutes. But, in the worst case, some patients would wait for 18 minutes to be treated in the surgical area or the medical area. In the major accident situation, 10.326 hours would be needed to treat 100 patients. On average, patients spend 3.175 hours in the ED, and red patients, yellow patients, and green patients wait for 0.2321 hours, 1.7985 hours, and 1.3549 hours, respectively. In the worst case, red patients, yellow patients, and green patients wait for 1.2016 hours, 4.1167 hours, and 3.9659 hours respectively. So, the waiting time for yellow patients is the longest. This result seems very logical because the number of yellow patients takes 30% of the total number of patients, while just two medical teams treat yellow patients. Moreover, taking into account the transportation time, the treatment time of yellow patients and red patients is the same.

3.5.3. *Design of experiments*

The experimental design is very useful in situations where the input parameters have to be specified and shows which input parameters have the biggest influence on the responses. It can be used to guide decision makers on devoting time/money to improve the responses. Since our main goal was to predict the average time used to treat all patients, we chose factorial design to plan our experiments to test the effect of resource dimensioning.

The factorial design is the experimental plan that considers k factors. These k factors are independent variables. There are only two levels for each factor: high level and low level. The choice of the two different levels should obey three principles: first, they are feasible; second, the difference between these two levels is big enough to trigger differences; and third, these two levels are close enough to assure that the system response is approximately linear over the range of the fact. The output result is called a response or the dependent variable. So, the effect of each independent variable on the dependent variable and the effect of interactions between the independent variable can be studied. In the normal situation, our experiment has the following independent variables: number of medical teams in the shock room (S), number of medical teams in the surgical area (T), and number of medical teams in the medical area (M), which lead to different combinations to be designed. In the major accident situation, the independent variables are the number of medical teams for red patients (R), the number of medical teams for yellow patients (Y), and the number of medical teams for green patients (G). The assignment of these health teams to different kinds of patients can be found in Tables 3.1 and 3.2.

Because 100 replications are enough to distinguish the mean of each scenario, we executed 100 replications for both of our sub-models, i.e., 800 runs for each sub-model and 1,600 runs for total. In SIMIO, a boxplot called SIMIO Measure of Risk and Error (SMORE) (Fig. 3.7) is used to present the simulation results. In the plot, SMORE, the mean, and the maximum and the minimum value obtained by a set of replications are presented. Meanwhile the

Table 3.1. Different design points and particular value taken by each scenario in the normal situation.

Design Point	S	T	M
1	2	3	2
2	2	3	4
3	2	6	2
4	2	6	4
5	4	3	2
6	4	3	4
7	4	6	2
8	4	6	4

Table 3.2. Different design points and particular value taken by each scenario in the major accident situation.

Design Point	R	Y	G
1	4	2	1
2	4	2	2
3	4	4	1
4	4	4	2
5	8	2	1
6	8	2	2
7	8	4	1
8	8	4	2

confidence interval of mean can be calculated as well. In our case, the confidence level is set to be 95%. Upper and lower percentile value can also be set. Here, the upper and lower percentile are set to be 75% and 25%, respectively. The confidence level of upper and lower percentile is set to be 95%.

The simulation results of the normal situation and the major accident situation are presented in Figs. 3.8 and 3.9. Each figure

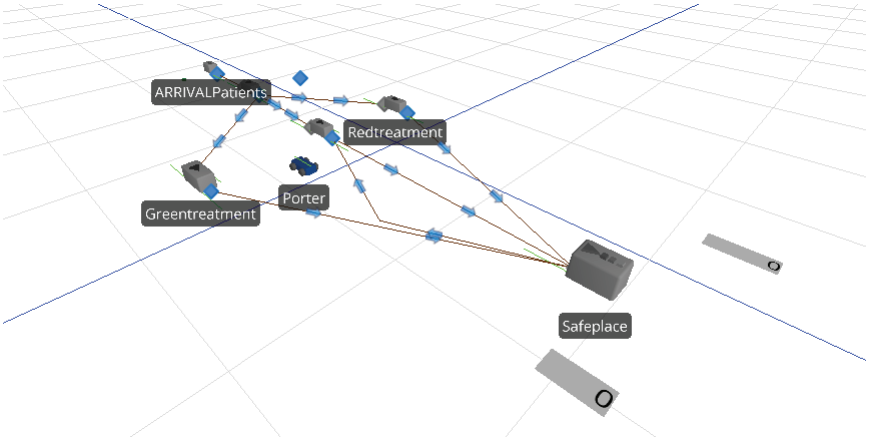


Figure 3.7. SIMIO Measure of Risk and Error Plot.

contains all possible combinations of possible resources, related to the design points of the factorial design. According to our independent variables, the process time to treat 22 patients in the normal situation and 100 patients in the major accident situation has been presented in these two figures, respectively. The less used time should point out the best choice. Taking this into account, in Fig. 3.8 (in the normal situation), scenarios 2, 4, 6, and 8 (corresponding to design points 2, 4, 6, and 8) clearly outperform others. The process time for scenarios 2, 4, 6, and 8 are 4.3276, 4.31235, 4.31067, and 4.28928 hours, respectively. However, the confidence intervals on the mean for scenarios 2, 4, 6, and 8 overlap each other. All these scenarios have the same number of teams in the medical area. Therefore, we can get our preliminary conclusion that the number of teams in the medical area has the biggest effect on the process time to treat all patients in the normal situation.

Figure 3.9 presented the process time to handle 100 patients injured by a terrorist attack. From Fig. 3.9, it can be found that scenario 4 and scenario 8 are better than others. The used periods for scenario 4 and 8 are 7.4 and 6.8 hours, respectively. The number of teams for green patients and yellow patients is the same. Therefore, we can get the preliminary conclusion that the effect of the number

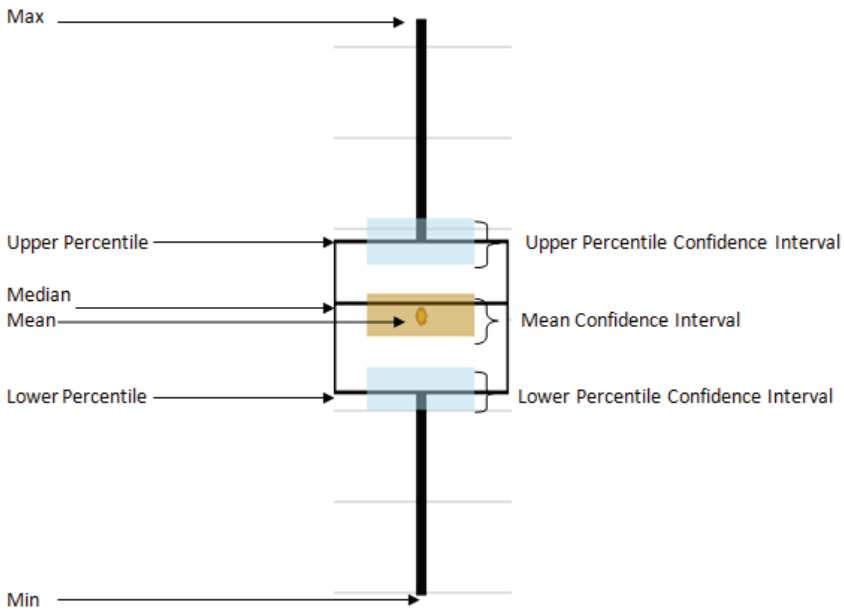


Figure 3.8. Simulation results in the normal situation.

of the medical teams for green patients and yellow patients is higher than the number of the medical team for red patients.

To prove our preliminary conclusion from a quantitative point of view, we will calculate the effect of each factor. Moreover, in order to help healthcare workers, just selecting the best scenario among different choices is far from enough. For decision makers, a comprehensive decision does not only mean the result is the best. It should also consider the impact of different resources on the output because the quantity of used resources has a direct relationship with the cost. Therefore, performing a sensibility analysis of the experiments is necessary. Based on our simulation result, the mean effect of each factor is calculated. We define the mean effect of an independent variable *a*. This mean effect is caused by the change of the value of independent variables, moving from the low level to the high level. The interaction between two independent variables *a* and *b* are denoted as E_a or E_b. This means the half of the difference

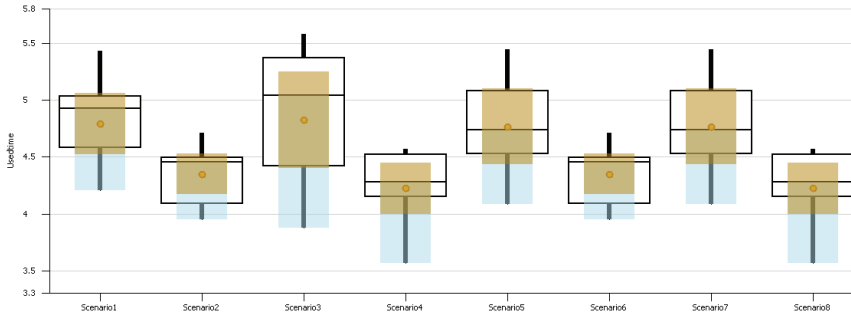


Figure 3.9. Simulation results in major accident situation.

Table 3.3. Sensitivity analysis in normal situation.

Factor(s)	Effect
E_T	–0.48
Es	–1.24
E_m	–1.33
Es_T	–0.96
Es_M	–0.09
E_{TM}	–0.05
Es_{TM}	–0.009

between the average effect of factor a when factor b takes its high level (and all the others factors are held constant) and the average effect of factor a when factor b takes its low level. We also computed the mean effect among all independent variables, denoted as E_{ab}. The results, sorted by higher relevance to lower relevance, are given in Tables 3.3 and 3.4.

According to Table 3.3, in the normal situation, the most effective factor (i.e., the one provoking the highest decrease in the used time when it moves from its low value (resources) to its high value (resources), the other factors being unchanged) is the number of teams in the medical area. Therefore, increasing the number of teams

Table 3.4. Sensitivity analysis in major accident situation.

Factor(s)	Effect
E_R	–0.094
E_Y	–0.407
E_G	–0.107
E_{RG}	–0.594
E_{RY}	–0.184
E_{YG}	–0.038
E_{RYG}	–0.03

in the medical area seems to be the most effective way to reduce the used time as a large number of patients need the treatment in the medical area. Based on Table 3.3, we can also find that the interaction effect between the number of teams in the shock room and the number of teams in the surgical area can also have a big effect on the used time. It is very logical because after the treatment in the shock room, some patients will be transferred to the surgical area for further treatment.

From Table 3.4, in the major accident situation, it can be found that the number of yellow medical teams has the biggest effect on the process time. Therefore, we can get a conclusion that, to reduce the used time, increasing the number of health workers to treat yellow patients may be the best choice. It is because the number of yellow patients accounts for a large percentage, and lot of time is required to treat yellow patients. Also, increasing the number of health workers to treat green patients can also have a good effect on the process time because green patients account for the largest proportion of patients.

3.5.4. *Improvements and suggestions for the hospital*

In this part, we will first propose one method to improve the current situation of the ED in the normal and the major accident situations

and, then, discuss the situation in which we should change the process of ED from the normal situation to the major accident situation.

1) The Improvement Method: For the ED, both in the normal situation and in the major accident situation, the capacities of different activities are fixed. The percentage of the utilization of resources is not equal. In other words, during simulation, the percentage of the utilization of resources can reach 100% for some activities. But, for some other activities, the percentage of the utilization of resources does not reach 80%. Therefore, we propose employing a coordinator that can assign the medical teams to different units, based on the current situation. Moreover, with the help of this coordinator, we can simulate from the process of the normal situation to the process of the major accident situation. We can treat the normal patients and patients in the major accident situation at the same time with the help of a coordinator.

   Figure 3.10 shows that, with the help of the coordinator, human resources can be assigned appropriately. In Fig. 3.10, there are two sub-models: one sub-model presents the process of activities in the normal situation and another presents the process of activities in the major accident situation. As the basic structure of these two sub-models are as same as the models we presented in Section 4, the description of the structure of these two

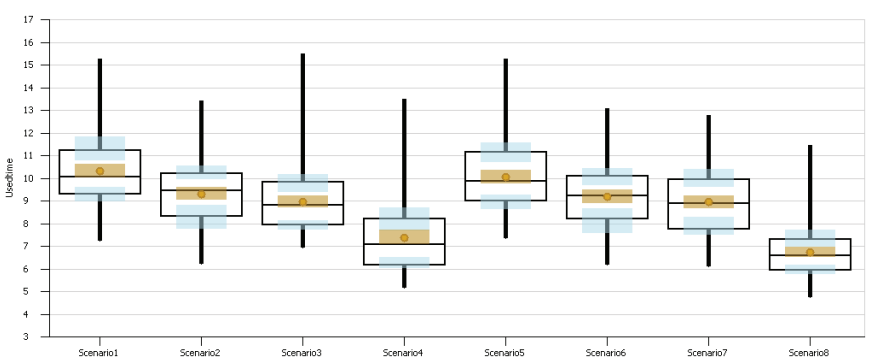


Figure 3.10. Improvement method.

sub-models is not required. The main difference between this new simulation model and the aforementioned two models is the way we simulate the resources. In old models, human resources are modeled by the servers with certain capacities. In the new model, human resources are presented by "Resource". In SIMIO, "Resource" can simulate human resources (the medical teams) with certain capacities that can be released and seized by other objects. Here, we use a "Table" to present the decisions of a coordinator about which human resource should be assigned to certain activities. We use "Table" to store the available human resources. If a "Server" needs a "Resource" to carry on certain activities, it should demand the available "Resource" in the "Table". If there is an available "Resource" in the "Table", the "Server" will get hold of it according to the first-come-first-service principle. When the activity is finished, the "Resource" will be released to the "Table". Based on the simulation results, if we use the coordinator to treat 100 patients under the major accident situation, we just use 8.2 hours, saving 2.126 hours.

2) The Condition to Change the Process: Until now, the hospital does not have a clear rule about the kind of situation under which the ED should take the normal process and the major accident process. Since in the normal situation the number of patients who visit the ED is about 20, therefore, we define this rule with 20 patients. In the major accident situation, it will cost 1.95 hours to treat 20 patients. To treat 20 patients in the normal situation, 4.6653 hours will be needed, and 2.7 hours will be saved. Though the use of the major accident process can save our time, it is better for us to keep the normal process in most of the cases due to the following reasons:

   i. The process of the normal activity and the process of the major accident activity do not use the same resources.
   ii. The number of patients that should be treated is not the same.
   iii. The objectives of these two situations are not the same.

   Usually, the process time to change the process from the normal situation to the major accident situation is less than 2.7 hours.

And, in the normal situation, the patients' arrival rate is bigger than the service capacity rate. Hence, some patients who are waiting to be treated are always there. Therefore, as long as the patients' arrival rate is more than 20 per hour, we should change the current process to the major accident process.

3.6. Conclusion

This paper studied the ED of a large-sized Italian hospital in the normal situation and in the overcrowding situation caused by supposed terrorist attack. A conceptual model, which can help health workers know when to do what in which kind of situation, was built by using the IDEF0 method. With the help of our conceptual model, the activity process was formalized, and then, SIMIO was used to present our conceptual model in detail. The SIMIO model can predict the time used to treat all patients of the ED in different resource dimensioning conditions. The factorial design was done to analyze the simulation results. Based on our simulation, we conclude that the number of medical teams should be increased to reduce the time used to treat all in the normal situation. In the major accident situation, we should increase the number of medical teams for yellow patients to reduce the used time. Also, we proposed to employ coordinators to improve the use of human resources as they can help save 2.126 hours in the normal situation.

References

1. Derlet, RW, JR Richards and RL Kravitz (2001). Frequent overcrowding in US emergency departments. *Acad Emerg Med*, 8(2), 151–155.
2. Warden, G, RB Griffin, SM Erickson, M Mchugh, B Wheatley, AS Dharshi, Madhani SJ and C Trenum (2006). Hospital-based emergency care: At the breaking point. *Natl Acad Sci Eng Med*, 1–8.
3. Trzeciak, S and EP Rivers (2003). Emergency department overcrowding in the United States: An emerging threat to patient safety and public health. *Emerg Med J*, 20(5), 402–405.
4. López-Valcárcel, BG and PB Pérez (1994). Evaluation of alternative functional designs in an emergency department by means of simulation. *Simulation*, 63(1), 20–28.

5. Kelton, WD and AM Law (2000). *Simulation Modeling and Analysis*. Boston: McGraw Hill.
6. Edmonds, MI and HM O'Connor (October 1999). The use of computer simulation as a strategic decision-making tool: A case study of an emergency department application. *Healthc Manage Forum,* 12(3), pp. 32–38.
7. Yeh, JY and WS Lin (2007). Using simulation technique and genetic algorithm to improve the quality care of a hospital emergency department. *Expert Syst Appl*, 32(4), 1073–1083.
8. Rossetti, MD, Trzcinski GF and S Syverud (1999). Emergency department simulation and determination of optimal attending physician staffing schedules. In *IEEE Simulat Conf Proc, 1999 Winter*, Vol. 2, pp. 1532–1540.
9. Hoot, NR, LJ LeBlanc, I Jones, SR Levin, C Zhou, CS Gadd and D Aronsky (2008). Forecasting emergency department crowding: A discrete event simulation. *Ann Emerg Med*, 52(2), 116–125.
10. Brenner, S, Z Zeng, Y Liu, J Wang, J Li and PK Howard (2010). Modeling and analysis of the emergency department at University of Kentucky Chandler Hospital using simulations. *J Emerg Nurs*, 36(4), 303–310.
11. Medeiros, DJ, E Swenson and C DeFlitch (December 2008). Improving patient flow in a hospital emergency department. In *Proceedings of the 40th Conference on Winter Simulation*, pp. 1526–1531. Winter Simulation Conference.
12. Hung, GR, SR Whitehouse, C O'Neill, AP Gray and N Kissoon (2007). Computer modeling of patient flow in a pediatric emergency department using discrete event simulation. *Pediatr Emerg Care*, 23(1), 5–10.
13. Ahmed, MA and TM Alkhamis (2009). Simulation optimization for an emergency department healthcare unit in Kuwait. *Eur J Oper Res*, 198(3), 936–942.
14. Ruohonen, T, P Neittaanmaki and J Teittinen (2006). Simulation model for improving the operation of the emergency department of special health care. In *Simulation Conference, 2006. WSC 06. IEEE Proc of the Winter*, pp. 453–458.
15. McGuire, F (1994). Using simulation to reduce length of stay in emergency departments. In *IEEE Simul Conf Proc, 1994 Winter*, pp. 861–867.
16. Komashie, A and A Mousavi (2005). Modeling emergency departments using discrete event simulation techniques. In *Proc of the 37th Conference on Winter Simulation*, pp. 2681–2685. Winter Simulation Conference.

17. Duguay, C and F Chetouane (2007). Modeling and improving emergency department systems using discrete event simulation. *Simulation*, 83(4), 311–320.
18. Zeng, Z, X Ma, Y Hu, J Li and D Bryant (2012). A simulation study to improve quality of care in the emergency department of a community hospital. *J Emerg Nurs*, 38(4), 322–328.
19. Ross, DT (1977). Structured Analysis (SA): A language for communicating ideas. *IEEE Trans Softw Eng*, SE-3, 16–34.
20. Shimada, Y and HA Gabbar (2008). Development of activity models of integrated safety and disaster management for industrial complex areas. In *Knowledge-Based Intelligent Information and Engineering Systems*, Lecture Notes in Computer Science, 5179, pp. 1–8. Springer, Germany.
21. Longo, F (2010). Emergency simulation: State of the art and future research guidelines. *SCS M&S Mag*, 1(4), 2010–2024.
22. Chen, W, A Guinet and A Ruiz (2015). Modeling and simulation of a hospital evacuation before a forecasted flood. *Oper Res Health Care*, 4, 36–43.
23. McGuire, F (1998). Simulation in healthcare. In *Handbook of Simulation*, J Banks (ed.), pp. 605–627. Charlotte: Premier, Inc.
24. Banks, J, JS Carson, BL Nelson and D Nicol (2004). *Discrete-Event System Simulation*, 4th Ed. New Jersey: Prentice Hall.
25. Raffo, DM and MI Kellner (2000). Empirical analysis in software process simulation modeling. *J Syst Softw*, 53(1), 31–41.
26. Runeson, P and M Wiberg (2005). Simulation of Experiments for Data Collection — A replicated study. In *5th Conference on Software Engineering Research and Practice*. Sweden, October 2005.

4. Stochastic and Dynamic Programming for Improving the Reservation Process of MRI Examinations

Na Geng

Department of Industrial Engineering and Management, Shanghai Jiao Tong University, China

Abstract

Quick diagnosis is critical for stroke patients but relies on expensive and heavily used imaging facilities such as Magnetic Resonance Imaging (MRI). To reduce the stroke patients' waiting time for MRI, this book chapter proposes a new reservation process. A certain number of appropriately distributed contracted time slots (CTS) are reserved for stroke patients. Except for CTS, the time slots by regular reservation (RTS) are still possible for stroke patients. The implementation of this new process need to determine the number of CTS and its distribution, the patient assignment policy to assign patients to either CTS or RTS, and the advance cancellation policy to cancel CTS in advance. Stochastic, dynamic programming, and local search methods are combined to solve these problems. In this new process, stroke patients assigned to RTS have to wait for about 35 days, much longer than those

assigned to CTS. In order to improve waiting time distribution, three other CTS implementation strategies, called RTS reservation strategies, are proposed which still make use of CTS and reserve RTS for stroke patients without directly assigning patients to RTS. All patients are scheduled to both CTS and RTS in the first-in, first-out (FIFO) order. Numerical experiments show that these new strategies greatly reduce the longest waiting time of stroke patients and avoid unlucky patients.

4.1. Introduction

This chapter introduces a new magnetic resonance imaging (MRI) examination reservation process, studied in [1–5], to reduce the waiting time of high-priority patients without degrading the utilization of MRI. This series of research is conducted in collaboration with a large French university teaching hospital in order to reduce the length of stay (LoS) of stroke patients treated in the neurovascular department.

As shown in Fig. 4.1, a stroke (sometimes called an acute cerebrovascular attack) is a sudden loss of brain function due to the block of blood supply to the brain (ischemic stroke) or the rupture

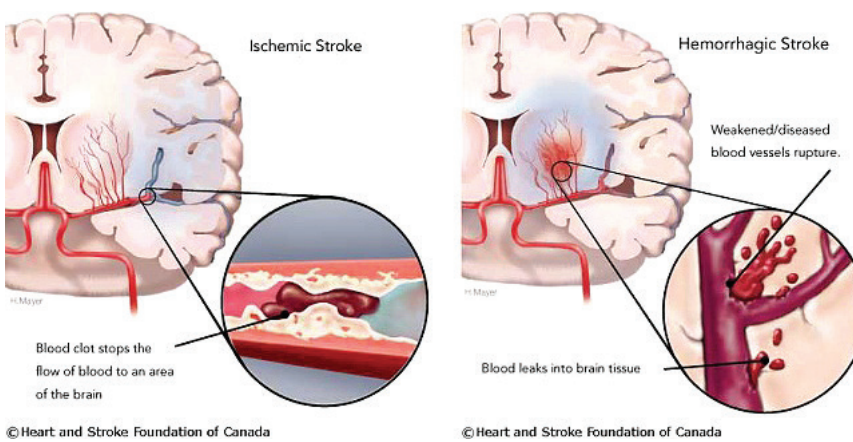


Figure 4.1. Photos for ischemic and hemorrhagic stroke.

Source: http://www.strokegenomics.org/index.php?page=about-stroke-genetics

of a blood vessel in the brain (hemorrhagic stroke), which leads to inability to move one or more limbs of one side of the body, understand or speak clearly, or see one side of the visual field. Stroke patients need the treatment as soon as possible, following a number of necessary examinations. Field observations in the collaborated hospital indicate that patients face significant delays in treatment as many key examinations rely on expensive and heavily-used imaging facilities such as MRI, as shown in Fig. 4.2. However, a new MRI scanner is very expensive (about $2 million), with a commensurate cost for building and preparing the space it needs. Therefore, hospital managers are under high pressure to reduce the LoS of stroke patients by reducing their waiting time for MRI examinations, without degrading the utilization of MRI scanners.

A six-month field observation was performed in the neurovascular department to collect data concerning patient arrival, medical examinations requested for each patient, delays of the examinations, and LoS of the patient. A detailed analysis of the historical data, as shown in Fig. 4.3, reveals that the neurovascular department has rather stable weekly demand for medical examinations.

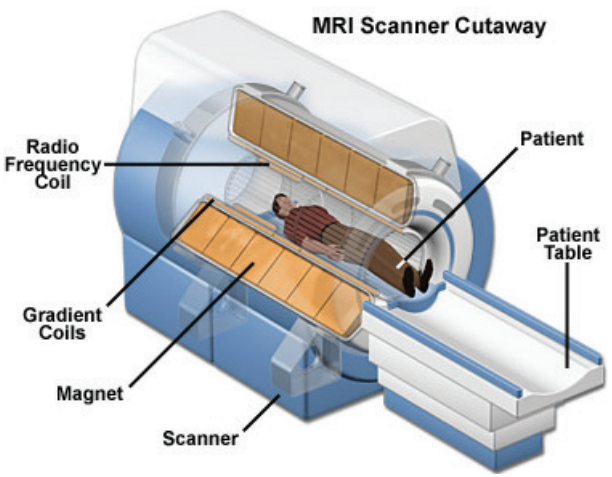


Figure 4.2. The photo of a MRI scanner.

Source: http://www.magnet.fsu.edu/education/tutorials/magnetacademy/mri/

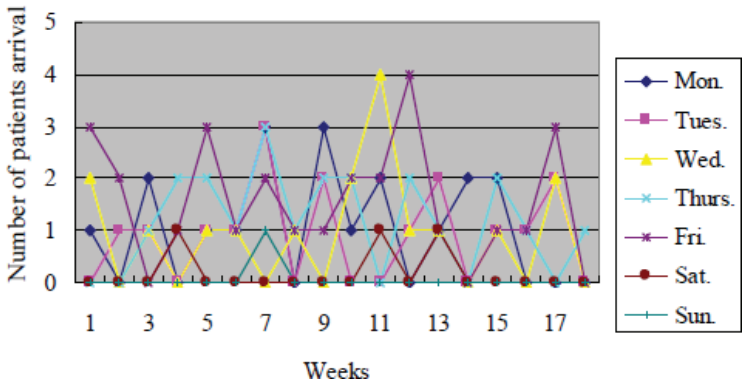


Figure 4.3. Historical data collected from the neurovascular department [1].

The neurovascular department is actually the largest customer of the imaging department. Furthermore, the MRI examination of stroke patients takes nearly the same time which is one time slot of about 30 minutes. Patients need to wait for about 30–40 days for MRI examinations.

Based on the observation in the neurovascular department of the target hospital, Geng [1], Geng *et al.* [2, 3, 5], and Geng and Xie [4] propose a contract-based MRI examination reservation process. A certain number of appropriately-distributed contracted time slots (CTS) are reserved for the patients with high priority, who usually suffer from more urgent illness than others. In addition, these patients can reserve regular time slots (RTS) via the regular reservation process. The contract-based examination reservation process is characterized by the following decisions and control policies:

1. Contract decisions, i.e., the number of CTS and its distribution over time.
2. Patient assignment control policy, which assigns patients to either CTS or RTS (if the patient is assigned to CTS, s/he needs to wait for CTS in the following days, otherwise, his/her examination is reserved through regular MRI examination reservation process).

3. Advance CTS cancellation policy, which cancels the CTS in advance when there are no enough stroke patients to fill CTS.
4. Improvement of patients' waiting time distribution, which tries to reduce the waiting time of stroke patients assigned to RTS.

A stochastic programming model is proposed to simultaneously determine the contract decisions, i.e., the number of the CTS and the patient assignment policy to assign patients to either CTS or RTS. To solve this model, an average-cost Markov decision process (MDP) approach is used to identify structural properties of the optimal control policy. A Monte Carlo approximation approach combining with a local search is used to determine the number and the distribution of CTS. The new reservation process greatly reduces the average waiting time of high-priority patients, with high underutilization ratio of CTS and much longer waiting time of patients assigned to RTS. To reduce the unused CTS, one-day and two-day advance cancellation of the slots are considered. Structural properties of optimal control policies are established via the average-cost MDP. Numerical experiments show that the appropriate advance cancellation of CTS greatly reduce the unused CTS with nearly the same waiting time. To reduce the waiting time of patients assigned to RTS, an improved reservation process is proposed to reserve an RTS according to three different criteria, without directly assigning the patient to RTS. This improved reservation process is proven to be able to reduce the unused time slots and improve the patients' waiting time distribution.

The remaining part of this chapter is organized as follows: Section 2 reviews the state-of-the-art methods and approaches used for the related problems. Section 3 presents the stochastic and dynamic programming model for the determination of CTS and the patient assignment control policy. Section 4 introduces the one- and two-day advance cancellation control policies. Section 5 describes three different improved reservation processes. Finally, section 6 concludes the chapter by highlighting some open-research problems.

4.2. Literature Review

The operational management of diagnostic devices, such as computer tomography (CT) and MRI scanners, includes capacity planning, capacity allocation, and scheduling. Capacity planning usually determines how many facilities should be purchased for the hospital. Capacity allocation helps in allocating the capacity to different patient groups usually to reduce the waiting time of some high-priority patients by sacrificing the waiting time of some low-priority patients. The capacity is measured in terms of the number of patients who can be examined in one day. Patients can be grouped according to the urgency level, the clinical department, and medical constraints. Patient scheduling helps reducing the patients' waiting time in three levels: advance scheduling, appointment scheduling, and real-time scheduling. Advance scheduling determines the number of patients that are scheduled to a particular day within a time horizon. Appointment scheduling helps in determining the appointment rule and assigning each patient to a time slot in a day. Real-time scheduling puts into sequence patients of different priority in an online way or to determine the patient that needs to be served next. In this chapter, we will focus on the literature on capacity allocation in diagnostic facilities.

The capacity allocation of imaging facilities has received limited coverage, with the earliest contribution being of Vasanawala and Desser [6]. The queuing theory is used to predict the optimal number of schedule slots that are served for urgent CT and ultrasonography. Green *et al.* [7] address how to match the demand with imaging diagnosis capacity by considering inpatients, outpatients, and emergency patients. The outpatient appointment schedule and the dynamic priority rules for admitting patients into service are considered. A finite-horizon dynamic program model is formulated, and the optimal control policy is established for admitting patients into services.

Patrick and Puterman [8] propose a simple approach for dividing the available diagnosis capacity between emergency patients and inpatients, on the one hand, between emergency patients and

outpatients on the other hand. A certain amount of capacity is reserved for the emergency by carrying over a percentage of the non-emergency inpatient demand to the next day. An MDP model is proposed in Patrick *et al.* [9] to schedule multi-priority patients to a diagnostic facility by considering the patients' waiting time targets. Waiting time targets are defined as the maximal allowable waiting time for each priority patient. An approximate dynamic programming approach is proposed to overcome the state space explosion problem.

Kolish and Sickinger [10] allocate the capacity of two CT scanners to three patient groups with different arrival patterns and cost structures. The problem is formulated as an MDP with the objective of maximizing the expected total reward. Sickinger and Kolisch [11] pursue the previous work to determine the optimal number of outpatients to be scheduled and assign the outpatients to a variable-block/fixed-interval appointment schedule. Schütz and Kolisch [12] propose a continuous-time MDP to solve the problem of accepting or rejecting the reservation of different services by different classes of customers. The solution strategy is proposed by combining simulation-based approximate dynamic programming (ADP) and discrete event simulation. Numerical experiments show that the heuristic ADP algorithm performs very well in terms of objective function value, solution time, and memory requirements. Schütz and Kolisch [13] propose an MDP approach to decide whether to accept requests for MRI examinations from patients with different priorities, such as inpatients and outpatients. Different examination types, cancellations, no-shows and overbooking, and same-day demand are considered. Patients' behaviors, such as preferences and no-shows and cancellations, are seldom considered in the capacity allocation level. They are usually considered in the patient scheduling level, e.g., Chakraborty *et al.* [14], Feldman *et al.* [15], and Laganga and Lawrence [16].

Zonderland and Timmer [17] employ the generic Bayesian game approach to deal with the fairness of capacity allocation in MRI scanners that depends on the quality of information provided by

hospital departments. The disclosure of true demand is stimulated, and then, the capacity is fairly allocated.

4.3. The Determination of CTS and the Optimal Patient Assignment Policy

This section introduces the contract design and the optimal patient assignment policy by Geng *et al*. [2]. The problem is formulated as a stochastic programming model from the perspective of stroke patients or high-priority patients with the objective of minimizing the patients' waiting time and unused CTS cost. The structural properties of the optimal patient assignment policy are identified via MDP approach. The Monte Carlo approximation is used to solve the stochastic programming model. The solution is improved by the local search.

We now present the assumptions and notations used in Geng *et al*. [2, 3, 5] and Geng and Xie [4]. Hereinafter, we use "patients", "stroke patients" and "high-priority patients" to refer to those who are urgent and "regular patients" to refer to those who are not urgent.

Assumption A1: Only MRI examination is considered, and one MRI time slot is required by each patient. Each patient can be assigned to either one CTS or one RTS.

Assumption A2: Emergency stroke patients are not considered, and all other stroke patients have the same priority.

Assumption A3: Patient arrival varies during a week but is stationary from one week to another. Furthermore, the number of arrivals on one day is independent of the arrivals on other days.

Assumption A4: The same contract is used for different weeks, i.e., $n_t = n_{t+7}$ for all t. As a result, the contract can be represented by a 7-entry integer-valued vector $\mathbf{n} = \{n_1, \ldots, n_7\}$.

Note that the number of CTS plus RTS should equal to the capacity of MRI facilities, i.e., the number of patients examined by MRI facilities in a day.

In the beginning of day t or equivalently weekday i, with $i = \mathrm{d}(t)$, number x_{t-1} of patients waiting for CTS, with x_0 a given constant,

and number a_t of patients arrives. By Assumption A3, daily arrivals a_t for $t \in IN$ are mutually independent random variables, and weekly arrivals (a_{7j+1}, a_{7j+2}, …, a_{7j+7}) are identically distributed for all $j = 0$, 1, …. Thus, the arrival process is characterized by the probability matrix $\boldsymbol{P} = [P_{ij}]$ for $i = 1, \ldots, 7$ and for all $j \geq 0$, with P_{ij} denoting the probability of j arrivals in weekday i. The contract decision is n_t, which means number $\min(x_{t-1} + a_t, n_t)$ of high-priority patients could be examined in day t. At the end of day t, number x_t of patients are waiting for the CTS in the following days. Penalty 1 or T^R is charged if 1 patient is waiting for CTS or RTS for one day with $T^R > 1$. c is penalty factor of an unused CTS. It serves as a weighting factor to balance the waiting times and unused MRI time slots.

4.3.1. *Model formulation*

Considering the above notation, let the contract decision be n_t and patient assignment x_t be decision variables. A stochastic programming (SP) model formulated in Geng *et al.* [2] is as follows:

Model-SP:

$$\underset{\mathbf{n},f}{MIN}\ \mathrm{E}\left[\lim_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T}\left(T^R y_t + x_t + c(n_t - x_{t-1} - a_t)^+\right)\right] \tag{1}$$

subject to:

$$y_t = f_t + (x_{t-1} + a_t) \leq x_{t-1} + a_t \tag{2}$$

$$x_t = (x_{t-1} + a_t - y_t - n_t)^+ \tag{3}$$

$$(n_1, n_2, \ldots, n_7) \in IN^7, f_t : IN \to IN. \tag{4}$$

where $y_t = f_t(x_{t-1} + a_t) : IN \to IN$: number of patients directed to RTS in day t.

The objective function (1) is the expected cost of waiting time and unused CTS. Constraint (2) defines the control policy for the use of RTS. Constraint (3) updates the CTS queue length. Model-SP is

impossible to solve mainly due to the underlying optimal control policies in constraint (2). Therefore, the patient assignment policy is first explored by assuming the known CTS and, then, the contract is determined by an optimization method. An infinite-horizon average cost MDP is proposed with the same objective of minimizing the expected stroke patients' waiting and unused CTS cost. Structural properties of the average cost MDP are established via discounted cost MDP by value iteration and by using relations between these two MDP models.

4.3.2. *Exploration of the optimal patient assignment policy via MDP*

The state of the system is represented by $z_t = x_{t-1} + a_t$, i.e., the state variable after patient arrivals. The control policy $\pi = \{\pi_1, \pi_2, ...\}$ is defined as $x_t = \pi_t(z_t)$ with $0 \le x_t \le (z_t - n_t)^+$. Note that the new definition of the control policy is equivalent to that of relation (2) as a result of relations (2)–(3).

The objective is to minimize over all policies $\pi = \{\pi_1, \pi_2, ...\}$ the average cost

$$J_\pi(i,z) = \lim_{T\to\infty} \frac{1}{T} E\left\{ \sum_{t=i}^{T+i} g_{d(t)}\left(z_t, x_t\right) \middle| z_i = z \right\} \tag{5}$$

or the α-discounted total cost with $0 < \alpha < 1$

$$J_{\alpha,\pi}(i,z) = \lim_{T\to\infty} E\left[\sum_{t=i}^{T} \alpha^{t-i} g_{d(t)}\left(z_t, x_t\right) \middle| z_i = z \right] \tag{6}$$

for any given initial state $z_i = z$ with $i = 1, \ldots 7$, where $g_{d(t)}(z_t, x_t)$ is the stage cost incurred at day t

$$g_{d(t)}\left(z_t, x_t\right) = c\left(n_{d(t)} - z_t\right)^+ + x_t + T^R\left[\left(z_t - n_{d(t)}\right)^+ - x_t\right] \tag{7}$$

Consider the following optimal cost function:

$$V_\alpha\left(i,z\right) = \underset{\pi}{MIN}\, J_{\alpha,\pi}\left(i,z\right)$$

The index α is omitted for simplicity in this section. From Bertsekas [18], since all stage costs $g_t(z_t, x_t) \geq 0$ and the control constraint set is finite for each z_t as $x_t \leq z_t$, the optimal cost function is a solution of the following optimality equations:

$$V(i,z) = \min_x \left\{ c(n_i - z)^+ + x + T^R (z - n_i - x)^+ + \alpha \sum_a P_{i+1,a} V(i+1, x+a) \right\} \quad \forall i = 1,\ldots,7 \tag{8}$$

with $7 + 1 \equiv 1$, which means Sunday is followed by Monday.

The optimal control policy is given by the argument x that reaches the minimum in (8), and the optimal cost function is the limiting function of the following value iteration:

$$V^t(z_t) = c(n_t - z_t)^+ + T^R (z_t - n_i)^+ + \min_{0 \leq x_t \leq (z_t - n_t)^+} \left\{ U^{t+1}(x_t) - (T^R - 1) x_t \right\} \tag{9}$$

where

$$U^{t+1}(x_t) = \alpha \sum_a P_{t+1,a} V^{t+1}(x_t + a) \tag{10}$$

$$V^0(z) = 0 \tag{11}$$

for $t = 0, -1, -2, \ldots$ As a result,

$$V(i,z) = \lim_{n \to \infty} V^{-7n+i}(z) \tag{12}$$

The major theoretical result of this chapter is based on the following properties of the value iteration by (9)–(11):

(i) the optimal x_t is non-decreasing in z_t,
(ii) $c \leq V^t(z_t + 1) - V^t(z_t) \leq T^R$, for any z_t and t,
(iii) $V^t(z_t)$ is convex in z_t, and $U^{t+1}(x_t)$ is convex in x_t.

Theorem 1 (Geng *et al.* [2], Theorem 1): The value functions $V(i, z)$ and $U(i, x)$ are convex functions in z and x, respectively, for

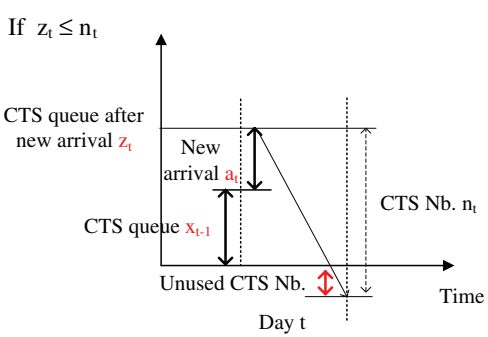
all $i = 1, \ldots 7$. Further the optimal control policy is of the following form:

$$x_i^* = \begin{cases} 0 & \text{if} \quad z - n_i \le 0 \\ z - n_i & \text{if} \quad 0 < z - n_i \le L_i \\ L_i & \text{if} \quad z - n_i \ge L_i \end{cases} \tag{13}$$

Structural properties of the optimal patient assignment control policy are established for the discounted-cost MDP model. The optimal control is to keep stroke patients in CTS queue if the CTS queue length at the end of a day is below some threshold L_i ($i = 1, 2, \ldots, 7$) for different days in a week, otherwise to send the remaining patients to RTS to make sure that the CTS queue ends at L_i. The average-cost MDP model has the same optimal control policy in the case of bounded patients' new arrival by using Proposition 4.2.6 in Bertsekas [18] and in the case of unlimited patients' new arrival by using Theorem 8.10.7 of Puterman [19].

The existence of optimal threshold control makes the implementation easy. According to the relation (13), the implementation of the L control policy can be divided into three cases [1], as shown in Figs. 4.4a–c:

Case 1: If the CTS queue after new patients' arrival z_t is smaller than the same-day CTS number n_t, then there are number $n_t - z_t$ of unused CTS and no patients waiting at the end of day t.



No assignment to RTS and empty CTS queue.

Figure 4.4(a). The optimal control if $z_t \le n_t$ [1].

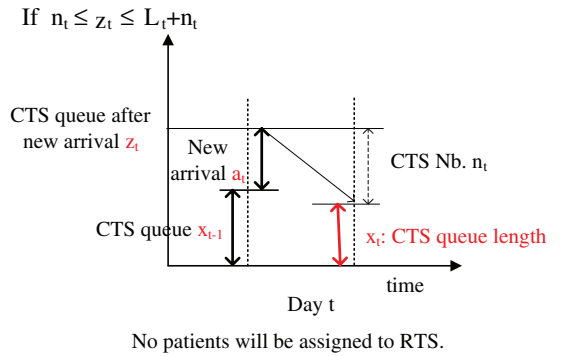


Figure 4.4(b). The optimal control if $n_t \le z_t \le L_t + n_t$ [1].

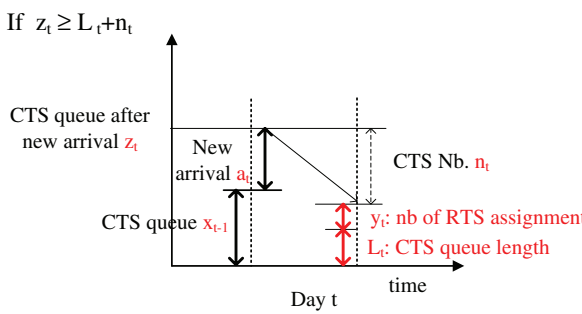


Figure 4.4(c). The optimal control if $z_t \ge L_t + n_t$ [1].

Case 2: If z_t is between the values of n_t and $n_t + L_t$, then all the remaining patients, i.e., number $z_t - n_t$ of patients are waiting for the following CTS and no patients are assigned to RTS at the end of day t.

Case 3: If z_t is greater than the values of $n_t + L_t$, then the number L_t of patients are kept in the CTS queue, and the remaining patients are assigned to RTS at the end of day t.

4.3.3. *Contract optimization via Monte Carlo approximation*

From the optimal patient assignment control policy (13), constraint (2) could be rewritten as:

$$y_t = \left(x_{t-1} + a_t - n_{d(t)} - L_{d(t)}\right)^+ \tag{14}$$

Due to the unknown $L_{d(t)}$, the model-SP with (14) replacing (2) is still impossible to solve. Therefore, this constraint is omitted. Furthermore, the uncertain demand is approximated by a deterministic optimization problem by using a single given but long enough sample path of patient arrivals. The nonlinear constraint (3) is reformulated as a linear constraint by adding one variable u_t, i.e., the number of unused time slots, whose reduction leads to the reduction of both x_t and y_t and, hence, the reduction of the objective value. The Monte Carlo approximation model provides a lower bound (LB) for the original model-SP.

Model-LB:

$$LB(\mathbf{a}) = \min\left(\sum_{t=1}^{T} T^R y_t + \sum_{t=1}^{T+D} x_t + c\sum_{t=1}^{T} u_t\right) / T$$

s.t.

$$\begin{aligned}
&x_{t-1} + u_t \geq n_t - a_t, && \forall t = 1,\ldots,T\\
&x_t - x_{t-1} + y_t - u_t = a_t - n_t, && \forall t = 1,\ldots,T\\
&x_{t-1} + u_t \geq n_t, && \forall t = T+1,\ldots,T+D\\
&x_t - x_{t-1} - u_t = n_t, && \forall t = T+1,\ldots,T+D\\
&n_t = n_{t+7}, && \forall t = 1,\ldots,T+D-7\\
&x_t, y_t, u_t, n_t \in \mathrm{IN}, && \forall t = 1,\ldots,T+D
\end{aligned}$$

The constraint matrix of the left-hand side of the constraints of the model LB is proven to be total unimodular. As a result, the integrity constraints of variables x_t, y_t, u_t can be relaxed. This model is easy to solve as it only involves seven-integer decision variables (n_1, n_2, n_3, …, n_7) and is expected to produce a good contract for Model-SP, which has been confirmed by the numerical experiments.

Starting from the initial contract solution of the LB model, a local search algorithm is proposed to improve the initial solution $\mathbf{n}^0$. Value iteration is used to find the exact objective value for contract decision $\mathbf{n}^0$ by considering the optimal patients' assignment policy. At each iteration, three different types of neighborhood are searched: $\mathbf{n} + \mathbf{e}_k$ (increasing one time slot in period k), $\mathbf{n} - \mathbf{e}_k$ (reducing one time

slot in period k), and $\mathbf{n} - \mathbf{e}_k + \mathbf{e}_j$ (moving one time slot from period k to period j). Value iteration is used to determine the best neighbor solution. This process repeats until no improvement can be found.

Numerical results show that although the lower bound from Monte Carlo approximation is not always tight, the contract decision is very close to the optimal one, with at most two local moves away from the optimal contract decision in all numerical experiments. The average waiting time of high-priority patients is greatly reduced, less than 5 days in most cases and less than 10 days even when the idle penalty is very large. However, about 8% of the CTS are unused, and 4% of the patients are assigned to RTS in the base case, when unused penalty $c = 15$ and RTS waiting penalty $T^R = 35$.

4.4. Joint Patient Assignment and Advance CTS Cancellation

In this section, we introduce the one-day advance cancellation of CTS by Geng *et al*. [3] and one-and two-day advance cancellation by Geng & Xie [4]. It is interesting to note that the optimal patient assignment and the advance cancellation policies are explored together. An average-cost MDP is separately proposed, and the optimal control policies are established via discounted-cost MDP, with the objective to minimize the expected stroke patients' waiting, MRI underutilization, and CTS advance cancellation cost. The local search algorithm, similar to the one proposed in Geng *et al*. [2], is used to improve the contract decisions. We now present the details of the approach.

4.4.1. *One-day advance CTS cancellation*

In Geng *et al*. [3], there are two control policies in this problem: patient assignment and one-day advance CTS cancellation policies. History-dependent policies are first considered, and then, optimal control policies are proved to be stationary Markov deterministic policies. The objective is to minimize the long-run average cost over the customer assignment policy by $\boldsymbol{\pi} = \{\pi_1, \pi_2, \ldots\}$ and the CTS

cancellation policy by $\boldsymbol{\mu} = \{\mu_1, \mu_2, \ldots\}$. We will see later that the former policy depends on the number of patients in the morning, i.e., z_t, and the latter depends on the number of patients at the end of a day i.e., x_t.

It is proved that there exists an optimal average cost policy such that $x_t \le \bar{x}$ for all $t > 0$, with $\bar{x} = [(T^R + c)n^*]$, where $[\bullet]$ is the least integer greater or equal to $\bullet$ and $n^* = \text{MAX}\{n_1, \ldots, n_7\}$. Structural properties are proved via discount-cost MDP. Thanks to this result, the following assumption is made without the loss of generality:

Assumption A5*:* $x_t \le \bar{x}$ for all $t > 0$.

In the following, we show the process of proof via discounted-cost MDP.

Since the set of states (i, z) is countable, and the control constraint set is finite as $x_t \le z_t$ and the number of cancellation $w_t \le n_{t+1}$ for each z_t, Theorem 6.10.4 in Puterman [19] implies that the optimal control policy is stationary deterministic and is given by the argument $\boldsymbol{w}$ and $\boldsymbol{x}$ that reach the minimum in (15)–(16), and the optimal cost function is the limiting function of the following value iteration:

$$U^t(z_t) = c(n_t - z_t)^+ + T^R(z_t - n_i)^+ + \min_{x_t \in [0,(z_t - n_t)^+ \wedge \bar{x}]} \left\{ V^t(x_t) - (T^R - 1)x_t \right\} \quad (15)$$

$$V^t(x_t) \min_{w_t \in [0,(n_{t+1} - x_t)^+]} \left\{ bw_t + \alpha \sum_a P_{t+1,a} U^{t+1}(x_t + w_t + a) \right\} \quad (16)$$

where $x \wedge y = \min(x, y)$, b is the one-day advance cancellation penalty, and w_t is the number of CTS for day t+1 cancelled in day t

$$U^0(z) = 0 \quad (17)$$

for $t = 0, -1, -2, \ldots$

In relation to (15)–(16), the following bound is proved: $-c \le U^t(z_t) - U^t(z_t - 1) \le T^R$ and $-b \le V^t(x_t) - V^t(x_t - 1) \le T^R$. Then, $U^t(z_t)$ is proved to be convex in z_t, and $V^t(x_t)$ is proved to be convex in x_t.

The major theoretical results are follows:

Theorem 2. The optimal value functions $U^t(z_t)$ and $V^t(x_t)$ are convex in z_t and x_t, respectively. Furthermore, the optimal control policy for the problem is characterized by the following form:

$$x_t^* = \begin{cases} 0 & \text{if} \quad z_t - n_t \le 0 \\ z_t - n_t & \text{if} \quad 0 \le z_t - n_t \le L_t \\ L_t & \text{if} \quad z_t - n_t \ge L_t \end{cases} \tag{18}$$

$$w_t^* = \begin{cases} S_{t+1} - x_t & \text{if} \quad x_t \le S_{t+1} \\ 0 & \text{if} \quad x_t \ge S_{t+1} \end{cases} \tag{19}$$

where

$$L_t = \underset{0 \le x_t \le \left((z_t - n_t)^+ \wedge \bar{x}\right)}{\arg\min} \left(V^t(x_t) - \left(T^R - 1\right)x_t\right)$$

$$S_t = \underset{y \ge 0}{\arg\min} \left\{ by + \alpha \sum_a P_{t+1,a} U^{t+1}(Y + a) \right\}$$

Proposition 4.2.6 in Bertsekas [18] and Theorem 8.10.7 of Puterman [19] are separately used to prove the same stationary optimal control policy for the average-cost MDP under the assumption of limited and unlimited new patients' arrival.

The existence of optimal control policies makes the implementation easy. For day t, the implementation of the optimal patient assignment policy depends on state variable z_t, while that of the optimal one-day advance cancellation policy depends on x_t, the CTS queue length at the end of day t. Therefore, the patient assignment is first made, then the CTS is cancelled for the next day. The implementation of the patient assignment policy is similar to Figs. 4.4a–c. The implementation of the optimal one-day advance cancellation control policy depends on the ending CTS queue at the end of the same day, which can be divided into two cases:

Case 1: As shown in Fig. 4.5a, if the ending CTS queue at day t x_t is smaller than S_{t+1}, then the number of CTS cancelled for day $t + 1$ is $w_t = S_{t+1} - x_t$.

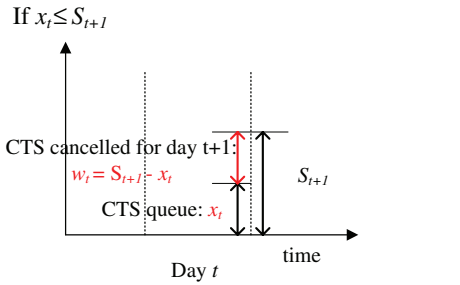


Figure 4.5(a). The optimal one-day advance CTS cancellation control if $x_t \leq S_{t+1}$ [1].

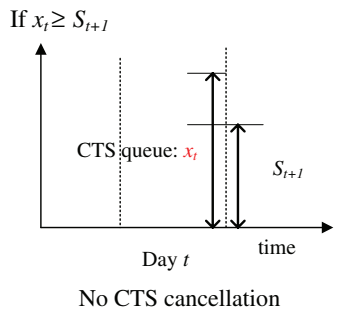


Figure 4.5(b). The optimal one-day advance CTS cancellation control if $x_t \geq S_{t+1}$ [1].

Case 2: As shown in Fig. 4.5b, if the ending CTS queue at day t x_t is greater than S_{t+1}, then there is no cancellation.

Numerical experiments show that the consideration of one-day advance CTS cancellation could greatly reduce the unused CTS ratio to less than 10% when the cancellation cost was smaller than half of the idle cost or the idle cost was large. On the contrary, the expect patients' waiting in the CTS queue was slightly increased.

4.4.2. *Joint patient assignment and one-day and two-day advance CTS cancellation policies*

Geng and Xie [4] explore the optimal patient assignment, one-day and two-day advance CTS cancellation policies are explored together.

The optimal control under a given contract is formulated as an average cost MDP in order to minimize patients' waiting, unused CTS, and CTS cancellation. The following notations are introduced in this paper:

w_t^1 and w_t^2 are the number of CTS cancelled in one day or two days advance at the end of day t, respectively, and $b_1(b_2)$ is the unit cost of one-day (two-day) advance cancellation

$$u_t = x_t + w_{t+1}^2, \quad y_t = u_t + w_{t+1}^1 = x_t + w_{t+1}^1 + w_{t+1}^2,$$
$$z_t = y_{t-1} + a_t = x_{t-1} + w_t^1 + w_t^2 + a_t$$

$$n^* = \mathrm{MAX}\{n_1, \ldots, n_D\}, \; a^* = \mathrm{MAX}\{\mathrm{E}[a_1], \ldots, \mathrm{E}[a_D]\}$$

Apart from the off-line contract decisions n_t, there are three on-line control decisions: customer assignment, one-period advance CTS cancellation, and two-period advance cancellation. At the end of period t, before making decisions $\left(x_t, w_{t+1}^1, w_{t+2}^2\right)$, the system state can be represented by $\left(z_t, w_{t+1}^2\right)$, i.e., the combination of the number of remaining customers and the number of period $t + 1$ CTS cancelled.

The objective is to minimize the long-run average cost of all contracts $\boldsymbol{n}$ and all history-dependent policies $(\boldsymbol{\pi}, \boldsymbol{\mu}^1, \boldsymbol{\mu}^2)$, which includes the penalties for CTS cancellation and unused CTS plus waiting time for CTS and RTS. History-dependent policies are first proved to be stationary Markov deterministic policies, then the structural properties are established via the corresponding discounted cost problem.

The optimal control policy is stationary deterministic and is given by any argument $(\boldsymbol{x}, \boldsymbol{w}^1, \boldsymbol{w}^2)$ that reaches the minimum in (20)–(22). The optimal cost function is the limiting function of the following value iteration:

$$U^t\left(z_t, w_{t+1}^2\right) = \min_{0 \le x_t \le (z_t - n_t)^+ \wedge \bar{x}} \left\{ c\left(n_t - z_t\right)^+ + x_t + T^R\left(z_t - n_t - x_t\right)^+ + V^t\left(x_t, w_{t+1}^2\right) \right\} \quad (20)$$

$$V^t\left(x_t, w_{t+1}^2\right) = \min_{x_t + w_{t+1}^2 \le y_t \le \left(\left(x_t + w_{t+1}^2\right) \vee n_{t+1}\right)} \left\{ W^t\left(y_t\right) + b_1\left(y_t - x_t - w_{t+1}^2\right) \right\} \quad (21)$$

$$W^t(y_t) = \min_{0 \le w^2_{t+2} \le n_{t+2}} E\left[b_2 w^2_{t+2} + \alpha U^{t+1}\left(y_t + a_{t+1}, w^2_{t+2}\right)\right] \qquad (22)$$

where for $t = 0, -1, -2, \ldots$ where $0 \le x_t \le \left((z_t - n_t)^+ \wedge \bar{x}\right)$, $w^1_{t+1} \le \left(n_{t+1} - x_t - w^2_{t+1}\right)^+$, $w^2_{t+2} \le n_{t+2}$, and $\wedge$ denotes component-wise minimization, and $\vee$ denotes component-wise maximization.

$U^t(z_t, w^2_{t+1})$, $V^t(x_t, w^2_{t+1})$, and $W^t(y_t)$ identify the optimal policies for customer assignment, one-period, as well as two-period advance cancellation, respectively.

The right-hand side of relation (21) is a function of u_t, with $u_1 = x_t + w^2_{t+1}$. Relation (21) becomes

$$V^t(u_t) \min_{u_t \le y_t \le (u_t \vee n_{t+1})} \left\{ W^t(y_t) + b_1 y_t \right\} - b_1 u_t \qquad (23)$$

The main result of this paper is the following theorem, which proves the convexity of the optimal cost functions and the structure properties of optimal control policies, i.e., the optimality of threshold control policies. The proof of this theorem relies on the technical analysis of different convexity properties of the optimal cost functions in the value iteration process defined by equations (20)–(22). In a summary, it is first proved that if $W^t(y_t)$ is convex in y_t, then $V^t(u_t)$ is convex in u_t. Then assume $W^t(y_t)$ is convex in y_t, $U^t(z, w)$ is supermodular and superconvex, and thus $U^t(z, w)$ is convex in z and w. Finally, it is proved that $W^{t-1}(y_{t-1})$ is convex in y_{t-1} if $W^t(y_t)$ is convex in y_t.

Definition 1 (**Koole [20]**): A function $f(x)$ is supermodular, denoted Super, if

$$f(x) + f\left(x + e_i + e_j\right) \ge f\left(x + e_i\right) + f\left(x + e_j\right)$$

or equivalently

$$f(x \vee y) + f(x \wedge y) \ge f(x) + f(y)$$

where $(x \vee y) = \max(x, y)$.

Definition 2 (Koole [20]): A function $f(x)$ is superconvex w.r.t. (i, j), denoted Super C (i, j) if

$$f\left(x+e_i\right)+f\left(x+e_i+e_j\right)\leq f\left(x+e_j\right)+f\left(x+2e_i\right)$$

From the above definitions, it can be proved that a function $f(x)$ that is Super and Super C (i, j) is convex in i.

Theorem 3. The optimal value function $U(z, w)$ in (20) is convex in z and w. $V(u)$ in (23) is convex in u. $W(i, y)$ in (22) is convex in y. Furthermore, the optimal control policies for the problem are as follows:

$$x_i^*=\begin{cases} 0 & \text{if} \quad z_i-n_i\leq 0 \\ z_i-n_i & \text{if} \quad 0\leq z_i-n_i\leq\left(L_i-w_{i+1}^2\right)^+ \\ \left(L_i-w_{i+1}^2\right)^+ & \text{if} \quad z_i-n_i\geq\left(L_i-w_{i+1}^2\right)^+ \end{cases} \tag{24}$$

$$w_{i+1}^{1*}=\begin{cases} S_{i+1}^1-x_i-w_{i+1}^2 & \text{if} \quad x_i+w_{i+1}^2\leq S_{i+1}^1 \\ 0 & \text{if} \quad x_i+w_{i+1}^2\geq S_{i+1}^1 \end{cases} \tag{25}$$

$$w_{i+2}^{2*}\left(y_i\right)=S_{i+2}^2\left(y_i\right) \tag{26}$$

where

$$L_i=\underset{w\leq u\leq w+\left(z-n_i\right)^+}{\arg\min}\left(V\left(i,u\right)-\left(T^R-1\right)u\right),\ S_{i+1}^1=\underset{u\leq y\leq\left(u\vee n_{i+1}\right)}{\arg\min}\left\{W_i\left(y\right)+b_1y\right\}$$

$$S_{i+2}^2\left(y\right)=\underset{0\leq w\leq n_{i+2}}{\arg\min}\left\{E\left[b_2w+U_{i+1}\left(y+a_{i+1},w\right)\right]\right\}$$

Structural properties of the optimal control policies for the average-cost MDP are established via the corresponding discounted-cost MDP problem. The local search was proposed to improve the contract.

The existence of optimal control policies makes the implementation easy. At the end of day t, the implementation of the optimal patient assignment policy first determines the CTS queue length x_t, which depends on state variable z_t and w^2_t. The second step is to determine the number of CTS cancelled for day t + 1, i.e., w^1_{t+1}, which depends on w^2_{t+1} and x_t. The final step is to make the two-day advance cancellation decision. The number of CTS cancelled for day t + 2 depends on state variable $y_t = x_t + w^1_{t+1} + w^2_{t+1}$.

Step 1: The implementation of the optimal patient assignment control policy can be divided into three cases:

Case 1: As shown in Fig. 4.6, if the state variable z_t is smaller than n_t, there exists the number $n_t - z_t$ of unused CTS, and no patient is waiting for the incoming time slots.

Case 2: As shown in Fig. 4.7, if state variable z_t is greater than n_t but smaller than $L_t + n_t - w^2_t$, then all the remaining patients are kept in the CTS queue, and no patients are assigned to RTS.

Case 3: As shown in Fig. 4.8, if state variable z_t is greater than $L_t + n_t - w^2_t$, then the number of patients assigned to CTS is kept at $L_t - w^2_t$, and the other remaining patients are assigned to RTS.

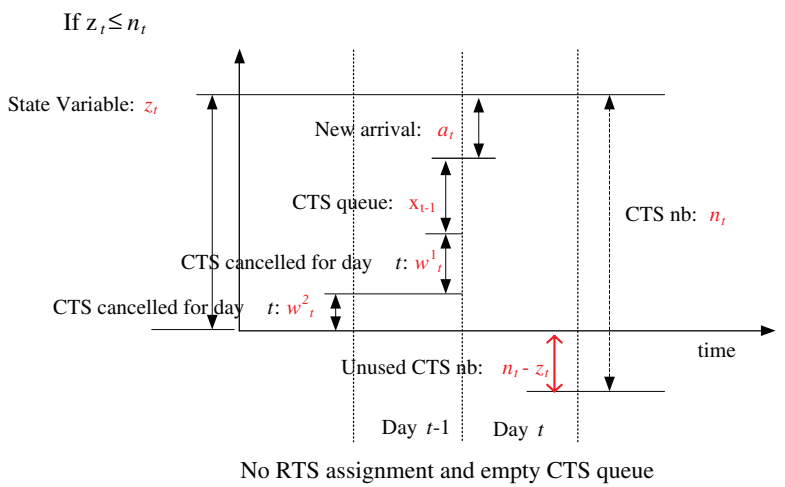


Figure 4.6. The optimal patient assignment control if $z_t \le n_t$ [1].

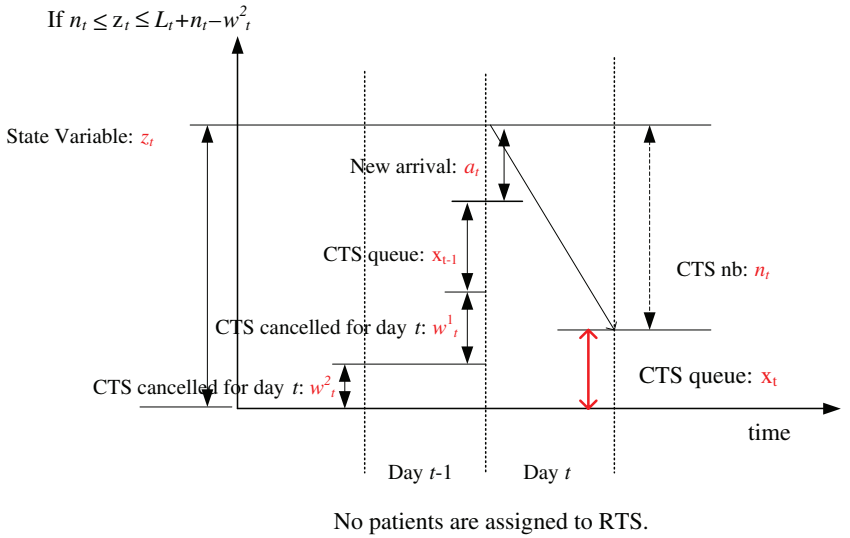


Figure 4.7. The optimal patient assignment control if $n_t \le z_t \le L_t + n_t - w^2_t$ [1].

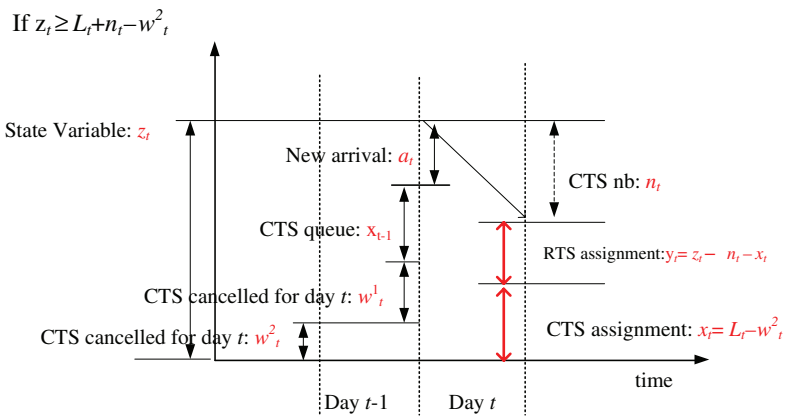


Figure 4.8. The optimal patient assignment control if $z_t \ge L_t + n_t - w^2_t$ [1].

Step 2: The implementation of one-day advance cancellation control can be divided into two cases:

Case 1: As shown in Fig. 4.9, if the ending CTS queue at day t plus two-day advance cancellation for day $t + 1$, w^2_{t+1}, is smaller

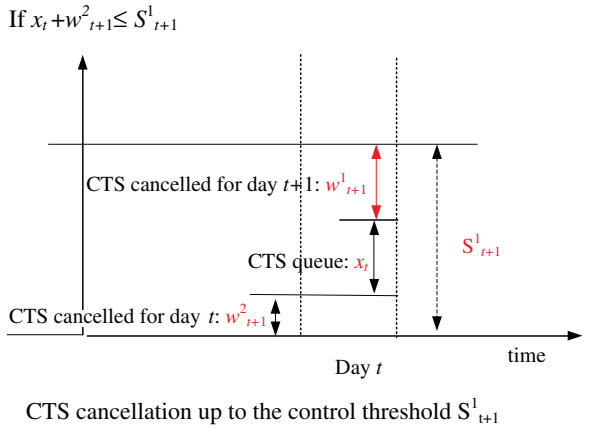


Figure 4.9. The optimal one-day advance CTS cancellation control if $x_t + w^2_{t+1} \leq S^1_{t+1}$ [1].

than S^1_{t+1}, then the number of CTS cancelled for day $t + 1$ is $w^1_{t+1} = S^1_{t+1} - x_t - w^2_{t+1}$.

Case 2: If $x_t + w^2_{t+1} \geq S^1_{t+1}$, no CTS is cancelled for day $t + 1$.

Step 3: The number of two-day advance cancellation depends on state variable $y_t = x_t + w^1_{t+1} + w^2_{t+1}$, which becomes known now, i.e., $S^2(y)$.

The numerical results show that the consideration of two-day advance cancellation and the local search further reduced the criterion values and improved the performance indicators.

4.5. Implementation Strategies

In the Geng *et al.* [2, 3] and Geng and Xie [4], stroke patients assigned to RTS have to wait for about 35 days, much longer than those assigned to CTS. It is unfair for the stroke patients using RTS. In order to have better waiting time distribution, Geng *et al.* [5] propose and analyze three other CTS implementation strategies, called RTS reservation strategies, without considering advance CTS

cancellation. The new strategies still make use of CTS. As shown in Fig. 4.10, these strategies reserve RTS for stroke patients without directly assigning patients to RTS. All patients are scheduled to both CTS and RTS in the first-in, first-out (FIFO) order. These new strategies are expected to reduce the longest waiting time of stroke patients and avoid unlucky patients. We now present the details of Geng *et al.* [5].

There are four implementation strategies: $\mathbf{P}_j$ with $j = 0, 1, 2, 3$, including one RTS assignment policy denoted as $\mathbf{P}_0$ and three RTS reservation policies $\mathbf{P}_1$, $\mathbf{P}_2$ and $\mathbf{P}_3$. Each policy $\mathbf{P}_j$ is associated with the following notations:

Y_{jt} number of patients directed to RTS or the number of RTS reserved at the end of day t

U_{jt} number of unused time slots in day t

x_{jt} total number of patients waiting for a time slot at the end of day t, including those directed to RTS but not yet served. It is called **global queue length.**

d_{jt} number of patients having received their time slots and, hence, left in day t.

The capital letter of each notation denotes the cumulative total from 0 to t. Notation A_t, D_{jt}, and U_{jt} will be used. The following notation is also used:

$$q_{jt} = x_{jt} - Y_j\left(t - T^R + 1,\, t - 1\right) - y_{jt} \tag{27}$$

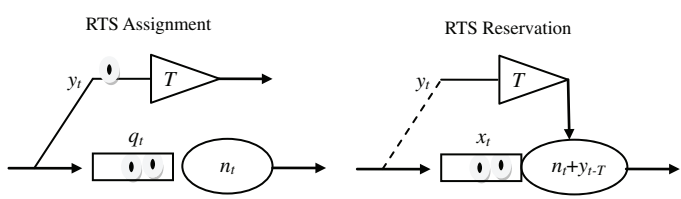


Figure 4.10. RTS assignment and RTS reservation [5].

where $Y_j(t',t) = y_{jt'} + \ldots + y_{jt}$, $Y_j(t - T^R + 1, t - 1)$ denotes the outstanding RTS assignment or reservation in day t. If $\mathbf{P}_j$ is a RTS assignment policy, q_{jt} corresponds to the **CTS queue length** and $x_{jt} = q_{jt} + \sum_{\tau=t-T+1}^{t} y_{j\tau}$. If $\mathbf{P}_j$ is a RTS reservation policy, x_{jt} is the queue length of waiting patients, and q_{jt} equals to x_{jt} minus the total number of outstanding RTS reservations.

In the following, four different implementation strategies are introduced:

$\mathbf{P}_0$ is the optimal **RTS assignment policy**, which is introduced in Geng *et al*. [2] to be a policy characterized by a control limit L_t and associated with each day with $L_t = L_{t+7}$. This policy keeps the CTS queue length q_{0t} at the end of each day t not exceeding L_t. As a result,

$$y_{0t} = \left(q_{0(t-1)} + a_t - n_t - L_t\right)^+ \tag{28}$$

$$q_{0t} = \min\left\{\left(q_{0(t-1)} + a_t - n_t\right)^+, L_t\right\} \tag{29}$$

With a RTS reservation policy $\mathbf{P}_j$ ($j = 1, 2, 3$), all patients wait in the same patient queue and are served by CTS and RTS in the FIFO order. At the beginning of period t, the length of the patient queue is $x_{j(t-1)}$, and the total number of available time slots is $n_t + y_{j(t-T^R)}$. Note that the number of RTS available in day t should be reserved $t - T^R$ days before. At day t, number y_{jt} of RTS is reserved according to three different implementation strategies.

$\mathbf{P}_1$ is similar to $\mathbf{P}_0$ and is called **RTS reservation with artificial queue**. $\mathbf{P}_1$ keeps track of an artificial CTS queue length q_{0t} as if $\mathbf{P}_0$ was used, and it determines y_{0t} with the artificial queue and relation (28). Let y_{1t} be equal to y_{0t}.

$\mathbf{P}_2$, called **RTS reservation with real queue**, is defined as follows:

$$y_{2t} = \left(x_{2t} - Y_2\left(t - T^R + 1, t - 1\right) - L_t\right)^+ \tag{30}$$

The number y_{2t} of RTS to reserve is determined by considering its effect on the patient queue. Assuming that the RTS reserved in day t is available in day $t + T^R$, the RTS reservation decision y_t only impacts the patient queue in day $t + T^R$. $\mathbf{P}_2$ tries to keep the expected queue length at time $t + T^R$ as close as possible to the threshold.

$\mathbf{P}_3$ is **RTS reservation with service ratio** α. Here, the service ratio at the end of a day is defined as the probability of having all the existing patients served. Each day t, the number of RTS to reserve is determined such that the service ratio at the end of day $t+T$ is at least α, i.e.,

$$P\left(x_{3(t+T-1)} + a_{t+T} \le n_{t+T} + y_{3t}\right) \ge \alpha$$

This policy requires the determination of the probability distribution of $x_{3(t+T-1)} + a_{t+T}$, which depends on the current queue length and all outstanding RTS reservations. In Geng *et al.* [5], this probability is determined by Monte Carlo simulation.

The performance of these four implementation strategies are compared analytically and numerically. Analytically, $\mathbf{P}_1$ and $\mathbf{P}_2$ both improve the RTS assignment policy $\mathbf{P}_0$ for the average waiting time, $\bar{W}_0 \ge \bar{W}_2 \ge \bar{W}_1$ and for the ratio of unused CTS, $\bar{U}_2 \le \bar{U}_1 = \bar{U}_0$. Numerically, all the analytical results are confirmed, and it is interesting to note that, although it is not proved, **P3** is the best strategy in terms of average waiting time, variance of waiting time, maximal waiting time, and unused CTS ratio in most parameter settings. For $\mathbf{P}_1$ and $\mathbf{P}_2$, the threshold policy could not only directly use the ones from $\mathbf{P}_0$, but also starts from the ones from $\mathbf{P}_0$ and improve by using local search. These improved control policies enhance the performance of $\mathbf{P}_1$ and $\mathbf{P}_2$.

4.6. Conclusions and Future Perspectives

High-priority patients, such as stroke patients, need quick diagnosis. However, significant delays are observed as many key examinations depend on expensive and heavily used imaging facilities such

as MRI scanners. The contract-based approach introduced in this chapter aims to reduce the waiting time of stroke patients for MRI examination without degrading the utilization of MRI scanner. Stochastic and dynamic programming-based approach is used to determine the contract decision and identify structural properties of patient assignment control policy and one- and two-day advance CTS cancellation control policy.

The results can be directly applied to design the contract and control policies for the department with high-priority patients for the critical facility. However, a lot of work still needs to be done if we want to develop a general approach. A direct extension is the consideration of multi-day advance cancellation of CTS and the implementation strategies with CTS advance cancellation. For the former problem, it is impossible to identify structural properties of the optimal control policies. Approximated dynamic programming model is possible to solve this problem. For the latter, the analytical comparison of different strategies is nearly impossible. Discrete-event simulation is a natural way to solve this problem. Another extension is to remove the assumption A4 and consider non-stationary arrival case. The form of the optimal contract is still an open issue. The management of multiple classes of patients and multiple imaging examinations is a natural but challenging research direction. The joint design of contract-based solutions of several departments for multiple examinations raises some fundamental questions, such as:

(i) How many time slots of a diagnostic facility to contract.
(ii) How to share these time slots among different departments.
(iii) How to make the real-time control to improve the utilization of imaging facilities.
(iv) How to consider the relationship of different examinations.

Results about the optimal control policies of this chapter can be extended to evaluate a contract solution. However, new approaches are needed to coordinate the contracts for different departments.

Acknowledgment

The authors acknowledge the support of the Natural Science Foundation of China (NSFC) under grants 71471113, and 71432006.

References

1. Geng, N (2010). Combinatorial optimization and Markov decision process for planning MRI examinations, l'École Nationale Supérieure des Mines de Saint-Étienne, PhD Thesis.
2. Geng, N, X Xie, V Augusto and Z Jiang (2011a). A Monte Carlo optimization and dynamic programming approach for managing MRI examinations of stroke patients. *IEEE Trans Autom Control*, 56(11), 2515–2529.
3. Geng, N, X Xie and Z Jiang (2011b). Capacity reservation and cancellation of critical resources. *IEEE Trans Autom Sci Eng*, 8(3), 470–481.
4. Geng, N and X Xie (2012). Optimizing contracted resource capacity with two advance cancelation modes. *Eur J Oper Res*, 221(3), 501–512.
5. Geng, N, X Xie and Z Jiang (2013). Implementation strategies of a contract-based MRI examination reservation process for stroke patients. *Eur J Oper Res*, 231(2), 371–380.
6. Vasanawala, SS and TS Desser (2005). Accommodation of requests for emergency US and CT: Applications of queueing theory to scheduling of urgent studies. *Radiology*, 235(1), 244–249.
7. Green, LV, S Savin and B Wang (2006). Managing patient service in a diagnostic medical facility. *Oper Res*, 54(1), 11–25.
8. Patrick, J and ML Puterman (2007). Improving resource utilization for diagnostic services through flexible inpatient scheduling: A method for improving resource utilization. *J Oper Res Soc*, 58(2), 235–245.
9. Patrick, J, ML Puterman and M Queyranne (2008). Dynamic multipriority patient scheduling for a diagnostic resource. *Oper Res*, 56(6), 1507–1525.
10. Kolish, R and S Sickinger (2008). Providing radiology health care services to stochastic demand of different customer classes. *OR Spectrum*, 30(2), 375–395.
11. Sickinger, S and R Kolisch (2008). A generalized bailey-welch rule and simple tabu search procedure for outpatient appointment scheduling.

In *Technical Report*. TUM Business School, Technishce Universitat Munchen.

12. Schütz, H-J and R Kolisch (2012). Approximate dynamic programming for capacity allocation in the service industry. *Eur J Oper Res*, 218(1), 239–250.
13. Schütz, H-J and R Kolisch (2013). Capacity allocation for demand of different customer-product-combinations with cancellations, no-shows, and overbooking when there is a sequential delivery of service. *Ann Oper Res*, 206, 401–423.
14. Chakraborty, S, K Muthuraman and M Lawley (2010). Sequential clinical scheduling with patient no-shows and general service time distributions. *IIE Trans*, 42(5), 354–366
15. Feldman, J, N Liu, H Topaloglu and S Ziya (2014). Appointment scheduling under patient preference and no-show behavior. *Oper Res*, 62(4), 794–811.
16. Laganga, LR and SR Lawrence (2012). Appointment overbooking in health care clinics to improve patient service and clinic performance. *Prod Oper Manag*, 21(5), 874–888
17. Zonderland, M and J Timmer (2012). Optimal allocation of MRI scan capacity among competing hospital departments. *Eur J Oper Res*, 19(3), 630–637.
18. Bertsekas, DP (1996). *Dynamic Programming and Optimal Control*, Vol. II. Belmont, MA: Athena Scientific.
19. Puterman, ML (1994). *Markov Decision Processes - Discrete Stochastic Dynamic Programming*. New York: John Wiley & Sons.
20. Koole, G (1998). Structural results for the control of queueing systems using event-based dynamic programming. *Queueing Sy*, 30(3–4), 323–339.

5. Simulation Modeling of Hospital Discharge Process

Zexian Zeng*, Xiaolei Xie†, Xiang Zhong‡, Barbara A. Liegel§, Sue Sanford-Ring§ and Jingshan Li¶

*Feinberg School of Medicine, Northwestern University, Chicago, IL 60611, USA
†Department of Industrial Engineering, Tsinghua University, Beijing, China, 100084
‡Department of Industrial and Systems Engineering, University of Florida, Gainesville, FL 32611, USA
§University of Wisconsin Hospital and Clinic, Madison, WI 53706, USA
¶Department of Industrial and Systems Engineering, University of Wisconsin, Madison, WI 53706, USA

Abstract

Introduction

Hospital discharge is an interdisciplinary process of critical importance and high complexity. Substantial efforts have been made to studying this process. However, less attention has been paid to using computer simulation models, which have been widely used in analyzing other health care units or delivery processes.

Objective

This study aims to quantitatively analyze the hospital discharge process and provide recommendations for potential improvement.

Methods

The computer simulation (or discrete event simulation) model was used to study the discharge process in the medical units at the University of Wisconsin Hospital and Clinics, USA. Using such a model, the impacts of discharge subprocesses under different scenarios were studied.

Results

The simulation results identified that the two main constraints of the discharge process were the waiting time for the physician's order and the waiting time before final discharge. A 26% reduction of the total discharge time can be achieved by reducing the wait time for the physician's order by half. A 9% reduction of the total discharge time can be achieved by cutting the wait time before the final discharge by half. In addition, a 3.59% reduction of the total discharge time can be achieved by slashing the pharmacist intervention rate by 60%–10%. Only 2.33% and 0.16% reduction of the total discharge time can be achieved by improving the efficiency of pharmacists and social workers/case managers by 50%, respectively.

Conclusions

The computer simulation model provided hospital administrations and discharge teams with insights to improve the hospital discharge process. Not only the bottlenecks of the discharge process were identified but also the areas that could be improved were quantitatively assessed. The possible areas of improvements include producing a shorter physician prescription processing time and a better coordination of events among discharge teams. Moreover, other factors such as high intervention rate of pharmacists and working efficiency of pharmacists, social worker, and case managers were not proved to be critical for the delay of the discharge process.

5.1. Introduction

In USA, a tremendously large population of patients are discharged from hospitals annually. It is estimated that 34.9 million patients were discharged from non-federal short-stay hospitals across the nation in 2006 [1]. The hospital discharge process is complex, with substantial variations and incredible challenges [2]. Due to its complexity, delays in the discharge process are common, which could impact the overall performance of hospitals [3]. Therefore, improving the quality and efficiency of the discharge process and other patient transitions across health care settings has become a national priority [4].

A typical discharge process involves multidisciplinary efforts from multiple care providers in the hospital, such as physicians (MDs), social workers (SWs), case managers (CMs), occupational therapists (OTs), physical therapists (PTs), pharmacists (RPHs), and nurses (RNs). It requires a wide range of clinical and organizational skills to address the needs of patients, families, aftercare facilities, and support systems. An efficient and high-quality discharge process is critical to reduce cost, improve resource utilization [2, 5], enhance performance of other departments (e.g., emergency department (ED)) [6], and limit the risk of adverse events after the patient leaves the hospital [7].

A substantial amount of effort has been made to studying the discharge process. From a hospital's perspective, a comprehensive review of current methods important for hospital discharge processes has been conducted in paper [8], which has identified challenges, including the continuity of inpatient-outpatient physician relationships, discrepancies in medication regimen, communication between physicians and their patients, and engaging patients in self-care. From the patient's perspective, paper [9] suggested that the assessment of the performance of the discharge process and the coordination between patients and care providers should be re-examined to ensure a successful transition experience.

Previous studies showed that more attention should be given to the participation of patients to better identify the needs of patients and facilitate the discharge process [10]. To improve the discharge

process, different interventions were studied to enhance discharge planning. It was shown that standardizing the discharge process through early discharge planning reduced the number of delayed discharges [11]. Moreover, the establishment of standardized medical criteria could increase discharge efficiency, specifically the length of hospital stay, without increasing re-admission rates [12]. To ensure optimal discharge planning, tools for planning, communication, education, and quality improvement were recommended in paper [13] for better practices. It was shown that suboptimal discharge planning could lead to delayed discharge [2], which could impact hospital operations management, for example, occupancy of ED beds [14, 15]. In addition, various disruptions might also result in discharge delays [16]. However, as shown in [17–20], it could be challenging to identify delay factors in the discharge process, such as medical or non-medical factors, internal or external reasons, psychological issues, evaluation errors, hospital capacity limitation, shortage of local facilities, and organizational assessment delays.

Despite these efforts, less research has been conducted to use advanced analytics to evaluate the process by considering all behaviors and factors, i.e., from a system's point of view. Although such approaches as discrete-event simulations (DES) have been widely used in analyzing other health care units or delivery processes (see reviews [21–25] and representative papers [26–34]), no such studies have been found to investigate the discharge process. As suggested in paper [35], more rigorous research is needed to discover how organizational factors, individual factors, and team factors affect the discharge process. DES provides an opportunity that can assess the efficiency of the existing system and investigate complex relationships among different procedures.

In this paper, we present a simulation study of the discharge process of the medical units at the University of Wisconsin Hospital and Clinics (UWHC) that are developed through extensive observations, historical data analyses, and discussions. Although the study was originated from UWHC, this modeling framework could be easily extended to the quantitative investigation of other discharge processes. Therefore, the significance of this work is to establish a

computer simulation framework to evaluate the current state of the discharge process, identify areas of variations and bottlenecks that impede the discharge process significantly, and provide strategies for potential improvement.

The remainder of the paper is divided in different sections. Section 2 describes the method, including the discharge process, the simulation model, data collection, model validation, and test design. Section 3 presents the results, and carries out what-if analyses to investigate the impact of parameter changes, such as pharmacist intervention rate and working time, social work and case manager working time, and potential waiting times. Further discussions are presented in Section 4, and conclusions are formulated in Section 5.

5.2. Methods

5.2.1. *The discharge process*

The discharge process is a coordinated multidisciplinary process, consisting of many components and variations. In UWHC and many other hospitals, the processes may vary significantly among different units. Even for the same unit, the process could vary with different physicians depending on factors such as their preferences, patient conditions, aftercare facilities, and options of transportation. To address such a complex process more effectively, we focus on the medical units at UWHC, which is more representative of the discharge process with less variations from case to case. Through extensive on-site observations, interviews, and discussions, we have gathered the following information about the discharge process.

Typically, the discharge process starts right after the admission of a patient. Once the patient is admitted in a medical facility, the social worker and the case manager begin to gather information related to the patient's health, insurance, referral, and contacts. As the patient approaches his/her discharge from the hospital, the final notes on the patient's medical condition, including the discharge summary, are compiled. Such a process should be almost completed before the patient is ready to be discharged. At UWHC, a morning-

round meeting usually starts at 9:00 a.m. It includes physicians, nurses, pharmacists, social workers, case managers, and therapists. In the meeting, physicians provide attendees a list of patients who are ready for discharge and inform them about the patients' conditions. Other staff will also report the progress of these patients. After the meeting, a discharge signal is triggered and the staff begins to cooperate, working toward the final discharge. In this study, we focus on the discharge process after the morning-round meeting, i.e., after the discharge decision is made.

As shown in Fig. 5.1, there are two primary parallel processes conducted by the social worker/case manager and the pharmacist. Since the social worker and the case manager share similar job functions in the discharge process, their workflows are grouped into one process in the figure. The third process shown in the figure is the transportation work for patients who need transportation assistance. This characterizes the period from discharge decision signal being triggered to the expected transportation being ready. Moreover, since RN education is always dependent on pharmacist education, it is included at the end of the pharmacist process. From our observations and interviews, in most cases, even though these three processes are concluded, there may be a delay before the final discharge due to the work of the occupational therapist and the physical therapist, as well as the patients' unawareness. The detailed workflows of social workers/case managers, pharmacists, and transportation work are described below.

5.2.1.1. *The SW/CM workflow*

Social worker and case manager share similar job functions in the discharge process and their work complement each other. The main responsibility of a social worker/case manager in the discharge process is to liaise with the patient and his or her family regarding discharge destination, aftercare facilities, insurance information, transportation arrangement, and to prepare the discharge packet. Such work can start long before the discharge decision is made. In most cases, the majority of the work is finished before the discharge day.

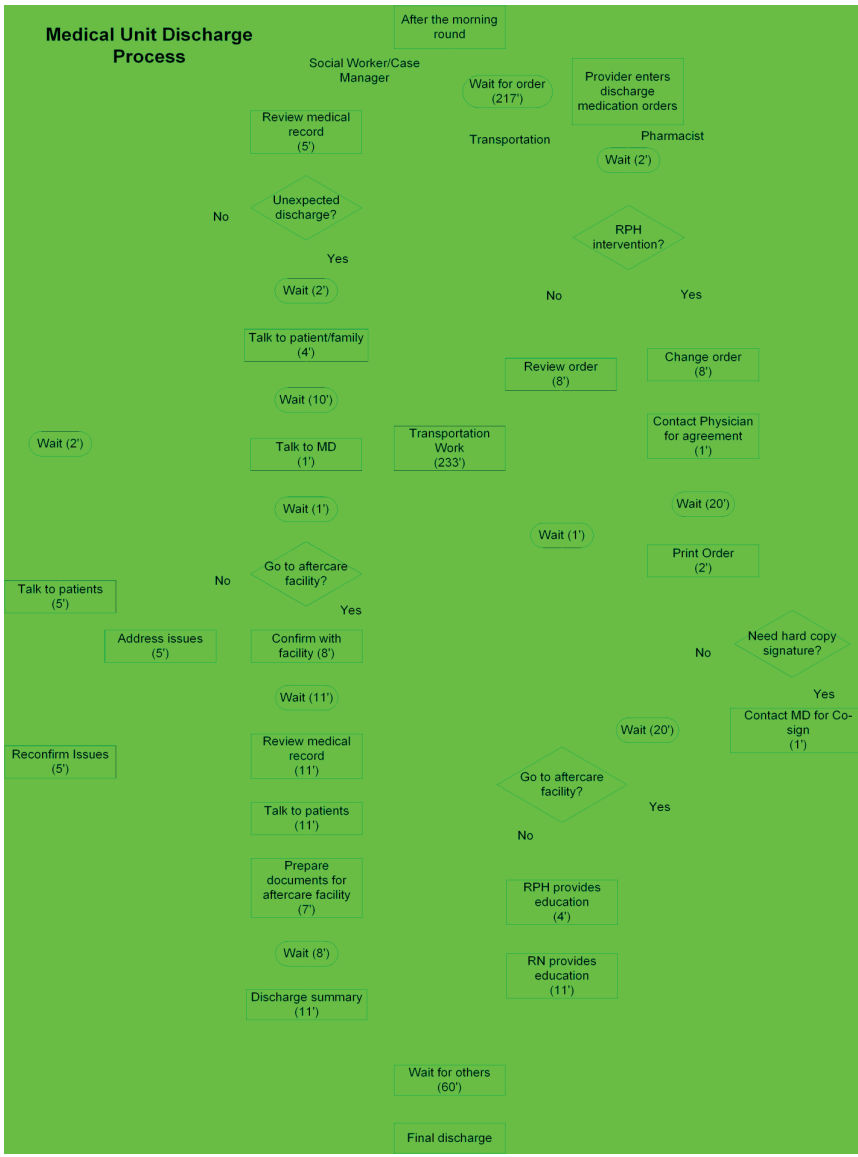


Figure 5.1. Hospital discharge process.

Based on our observations and interviews, it is estimated that only about 2% of the discharge decisions were made unexpectedly. Otherwise, the main work of a social worker/case manager on the discharge day is to confirm that all the information is correct and all the procedures are followed. Typically, the social worker/case manager runs a quick scan through all the referral information, addresses any issues that he/she finds, and then prepares a summary report. The left part of Fig. 5.1 illustrates the workflow of a social worker/case manager. The following is the breakdown of the steps in the process.

1) The social worker/case manager first reviews the medical record. If there are no unexpected changes, he/she will talk to the patient to reconfirm the status of things such as aftercare facility, insurance, and transportation. Afterwards, the social worker/case manager prepares the discharge packet.
2) If any change is required, the social worker/case manager talks to the patient and his/her family and also communicates with the physician.

If the patient is to be discharged to home, after addressing any arising issues, the discharge packet is handed to the patient. If the patient discharges to an aftercare facility, the social worker/case manager needs to confirm with the facility, talk to the patient, and prepare the discharge packet and necessary documents for the aftercare facility.

5.2.1.2. *The RPH workflow*

The main responsibility of a pharmacist in the discharge process is to clear any potential errors in the medical order and communicates with patients about the prescriptions. Specifically, he/she continues with the following procedures (shown on the right side of Fig. 5.1):

1) Once the medical order is prescribed after the morning-round meeting, the pharmacist will go through the medical order to

check for any error or issue. If no error or issue is found or raised, the pharmacist does not contact the physician.

2) There is about 60% chance that the pharmacist intervention is needed. In this case, the pharmacist makes the necessary changes and contacts the physician for agreement. If there is any specific medicine prescribed in the order, the pharmacist needs to ask the physician to co-sign.
3) If the patient's discharge destination is home, the pharmacist will print the order and bring it to the patient for education. If the patient is discharged to an aftercare facility, the pharmacist will only need to submit the order to the discharge packet.

5.2.1.3. *The transportation workflow*

The last process shown in Fig. 5.1 is the transportation work, which characterizes the period from discharge decision signal being triggered off to the expected transportation being ready. The expected transportation time and transportation methods are set up by the social worker, usually with an agreement from the patient and the physician. Some 50 observed discharge processes at UWHC indicated that only 14% of the patients needed this transportation work. Transportation can also become a bottleneck if the discharge-related work is finished, but the expected transportation is not ready yet.

5.2.2. *The simulation model*

Note that the discharge process is completed only when all the processes of the social worker/case manager, the pharmacist, and the transportation department are finished. Therefore, the actual discharge time could be much longer than the expected time of an individual process due to the maximum delay in each process. Thus, only studying one process (even if it is a critical path) is not sufficient, and developing a complete model to integrate all the processes is necessary.

In this study, a discrete event simulation model developed through a commercial software *SIMUL8* is used to emulate the discharge process. The model is constructed based on the workflow involved in

the case of a typical patient. The three parallel processes discussed above are included in the model. All parallel processes must be completed before the discharge of the patient. An illustration of the simulation model is shown in Fig. 5.2.

5.2.3. *Data collection*

The data and parameters used in the model were based on the combination of on-site observations, records from round meetings, structured interviews, and data extracted from the electronic health records (EHRs). In July 2012, at the medical units of UWHC, 50 on-site observations were conducted to estimate the processing time of most procedures. Round meetings were conducted to summarize the processing time of the procedures that were not collected and reach agreements from each party. Interviews were conducted to investigate different perspectives and potential modifications of the discharge process. In addition, the data extracted from the current EHR system of UWHC was used to calculate the routing probabilities of processes, such as the process of pharmacist intervention. A total of 2,934 discharge cases from the medical units were extracted for the analysis in May 2012.

The 50 on-site observations indicated that the total discharge time was 336 minutes on an average, with standard deviation of 35.6 minutes. Using the collected data, each procedure in the discharge process was modeled as a random processing duration given by a probability density function. As one single procedure can be characterized by multiple probabilistic distributions, the Stat-Fit function in the software was used to choose the best-fitted distribution. In addition, for procedures with insufficient data, extensive discussions were conducted to obtain expert opinions to characterize their durations.

The mean time of each procedure is illustrated in Fig. 5.1. The duration of the procedure "Wait for order" was modeled as Beta (12, 629, 28, 2.6); the duration of the procedure "Transportation work" was modeled as Beta (55, 451, 0.76, 1.1); and the duration of the procedure "Wait for others" was modeled as Lognormal (0, 3.6, 1.15) (note all above distribution units are in minutes). All other

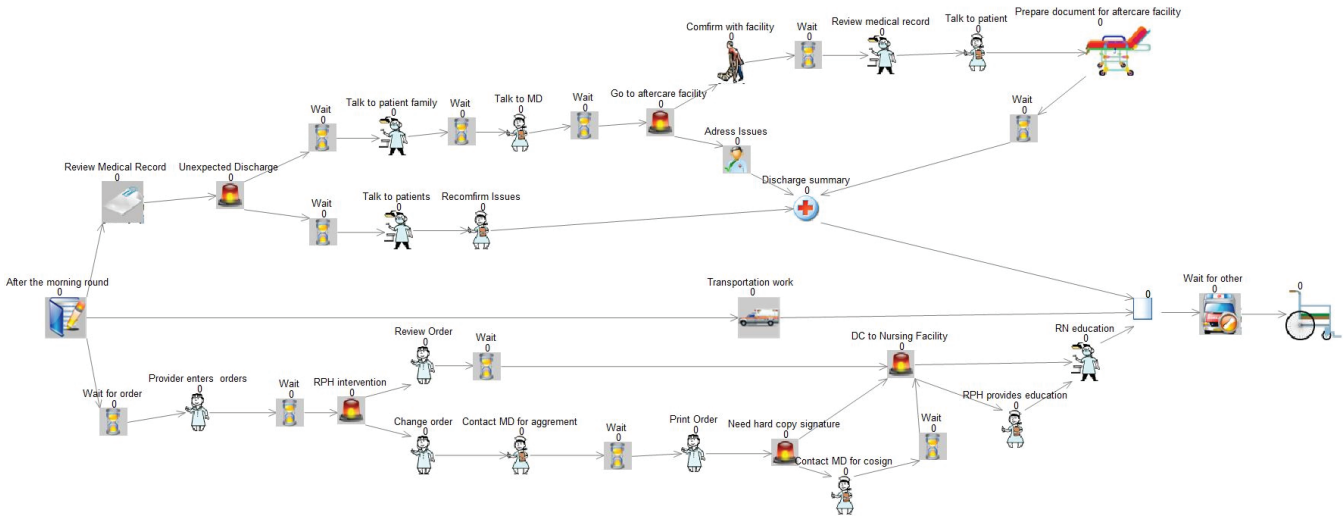


Figure 5.2. Illustration of the simulation model.

procedures were modeled as exponential distribution since these procedures were addressed by the combination of meetings, interviews, and observations.

5.2.4. *Model validation*

Extensive simulation experiments were carried out to imitate the discharge process using the developed simulation model. In all simulations, each experiment simulated 50 patients with 60 replications. The results indicated that the average discharge time under the current setting was 329 (±4.61) minutes. This result was compared with 336 (±35.6) minutes obtained from 50 observations at UWHC.

Let T_{sim} and T_{obs} represent the average discharge time obtained by simulation and observation, respectively, and δ denote the difference of the average discharge times. The difference was 2.21% using the following formula:

$$\delta = \frac{|T_{sim} - T_{obs}|}{T_{sim}} \cdot 100\%,$$

This result suggested that the simulation model had sufficient accuracy to estimate the discharge time. Therefore, such a model is validated and could be used for subsequent analysis.

5.2.5. *Test design*

To identify the bottleneck, i.e., the most impeding processes, and develop strategies to improve discharge efficiency and quality, the developed simulation model was used for "what-if" analysis under different parameter settings. Here, the bottleneck process refers to the process whose improvement will lead to the largest improvement of the overall discharge process. The pharmacist intervention rate, order processing time by pharmacist, case processing time by the social worker/case manager, the procedure of "wait for physician order", and the procedure of "wait for others" were studied.

5.3. Results

Different scenarios were created and simulated using the developed simulation model. The following results include the average values, half-widths of confidence interval (shown in parentheses), and p-values. All the time units are listed in minutes.

5.3.1. *RPH intervention rate*

Pharmacist Intervention is the process in which the pharmacist identifies the potential problems in the prescribed orders by the physician. As indicated by the collected data, such an intervention rate was 60%. The high intervention rate is a potential critical factor in the discharge process. In this experiment, the intervention rate was adjusted to 50%, 40%, 30%, 20%, and 10% in the simulated models. The corresponding time improvement and the p-values compared to the original discharge time from the simulation model are presented in Table 5.1. As one can see, the discharge time was decreased to 327, 325, 322, 320, and 317 minutes with the falling percentage of the intervention rate. The reduction of total discharge time is not practically significant when the intervention rate is above 30%.

5.3.2. *Reducing RPH working time*

In this study, whether the order processing time by the pharmacist is a critical factor for the system was investigated. Assuming the

Table 5.1. What-if analyses: Impact of pharmacist's intervention rate.

Intervention	50%	40%	30%	20%	10%
Discharge time (minutes)	327 (±4.53)	325 (±4.57)	322 (±4.56)	320 (±4.43)	317 (±4.43)
Improvement	0.61%	1.37%	2.15%	2.90%	3.59%
p-value	0.54	0.18	0.03	0.0041	0.0004

Table 5.2. What-if analyses: Impact of reducing pharmacist's working time.

Reduction to	90%	80%	70%	60%	50%
Discharge time (minutes)	328 (±4.54)	326 (±4.55)	324 (±4.58)	323 (±4.52)	321 (±4.53)
Improvement	0.42%	0.87%	1.42%	1.91%	2.33%
p-value	0.67	0.39	0.16	0.06	0.02

pharmacist to be efficient, the working time was cut to 90%, 80%, 70%, 60%, and 50% level of its original value. This reduced the discharge time to 328, 326, 324, 323, and 321 minutes, respectively. When the reduction was above 60% of the original value, the discharge time was not distinctively different (see Table 5.2). Therefore, the discharge time was not sensitive to changes in the pharmacist's working time.

5.3.3. *Reducing SW/CM working time*

As a potential factor in the discharge process, the working efficiency of social workers/case managers was also analyzed in this study. Their working time reduced to 90%, 80%, 70%, 60%, and 50% levels of the original time. As shown in Table 5.3, this led to a minimal change in the discharge time. Thus, a considerable decline in the working time of social workers/case managers will not affect the total discharge time significantly.

5.3.4. *Reducing the time of "wait for physician's order"*

The collected data suggested that, on an average, the physicians took about 207 minutes to issue discharge orders after morning-round meetings. By reducing this waiting time to 90%, 80%, 70%, 60%, and 50% level of its original value, the discharge time changed to 308, 288, 267, 247, and 228 minutes, respectively. As shown in

Table 5.3. What-if analyses: Impact of reducing social worker's/case manager's working time.

Reduction to	90%	80%	70%	60%	50%
Discharge time (minutes)	329 (±4.6)	329 (±4.63)	329 (±4.61)	329 (±4.62)	329 (±4.63)
Improvement	0.05%	0.11%	0.11%	0.14%	0.16%
p-value	0.96	0.92	0.92	0.89	0.87

Table 5.4. What-if analyses: Impact of reducing time of "waiting for physician's order".

Reduction to	90%	80%	70%	60%	50%
Discharge time (minutes)	308 (±4.11)	288 (±3.47)	267 (±3.25)	247 (±2.74)	228 (±2.52)
Improvement	6.37%	12.63%	19.01%	25.03%	30.68%
p-value	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001

Table 5.4, a decrease in such waiting time significantly reduced the discharge time.

5.3.5. *Reducing the time of "wait for others"*

In most cases, the patients are not discharged right way after the work of the pharmacist, social worker/case manager, and transportation is finished. The period from the end of the three processes to the actual discharge was characterized as 'wait for others' in the simulation model. Many factors contribute to this delay, such as OT education, PT education, patients' unawareness of discharge process, and delay of patients' families. The impact of reducing "wait for others" on the system performance was studied. Such a procedure time was reduced to 90%, 80%, 70%, 60%, and 50% levels of its original value. As shown in Table 5.5, the corresponding discharge time changed to 323, 317, 311, 305, and 299 minutes, respectively. Thus,

Table 5.5. What-if analyses: Impact of reducing time of "wait for others".

Reduction to	90%	80%	70%	60%	50%
Discharge time (minutes)	323 (±4.5)	317 (±4.40)	311 (±4.38)	305 (±4.22)	299 (±4.15)
Improvement	1.95%	3.73%	5.44%	7.39%	9.27%
p-value	0.05	0.002	<0.001	<0.001	<0.001

reducing the procedure time of "wait for others" could significantly improve the discharge duration.

5.4. Discussions

It is common to hear complaints that high intervention rate from pharmacists is a critical factor behind delayed discharge. This might lead to the pharmacist's intention to reduce the intervention rate despite the fact that pharmacist intervention was proved effective in reducing medication errors and adverse drug events [36]. However, we showed that by reducing the intervention rate from 60% to 10%, the total discharge time reduced by 3.59%. Thus, cutting pharmacist intervention will not delay the discharge process significantly. The pharmacist's intervention to eliminate medication error and adverse events is critically required.

In this study, the working efficiencies of social workers/case managers and pharmacists were also found to be non-critical to the discharge process. The discharge time was reduced only slightly when we doubled their working efficiencies. In other words, hiring more social workers/case managers and pharmacists will not improve the discharge process significantly.

From the above results, "wait for physician's order" is a critical bottleneck of the discharge process. A 50% reduction in its original value cut the discharge time significantly, by 30.68% from its original time value. In practice, physicians dominate the discharge process by making the final discharge decision. They are viewed as the captain of the ship in the discharge process. Thus, if the physician's order can

be prescribed at an earlier time, the whole discharge process can be shortened significantly. However, such a long waiting time is due to reasons such as waiting for lab results, patients' observation period, and high utilization of physicians. Thus, further investigations in reducing the waiting time for physician's order would help improve the efficiency of the discharge process significantly. More investigations in reducing lab turnaround time, freeing physicians' hands in the morning time, or possibly prescribing order before the day of discharge could lead to potential improvement in the discharge process.

Even if social workers/case managers and pharmacists finished their processes and the necessary transportation arrived on/before time, there was still a delay before the final discharge. This extra time was proved as a significant factor of the discharge delay because a 50% reduction of its original value could decrease the total discharge time by 9.27%. This critical waiting process could be addressed by better coordination among the discharge team members. In particular, the involvement of nurses in coordinating with physicians, therapists, lab, pharmacists, case managers, etc., can play a critical role to achieve timely discharge.

In addition, to make the discharge process more efficient, dedication of pharmacist to the discharge process in the morning could be helpful. Despite having many interventions during the discharge period, pharmacists were not a bottleneck in this study. If physicians could prescribe the discharge order earlier, then the discharge process might be more sensitive to the pharmacists' work due to their complex duties. Therefore, it is important that pharmacists continue focusing on the discharge process.

5.5. Conclusions

A delay in the hospital discharge process is a nationwide problem. Although many studies have been carried out, most of them emphasize a certain phase or aspect of the process. In the current literature, there is no quantitative analysis discovered studying the process from an overall or the system's point of view. Modeling the processes to predict the effects of various improvement strategies is

important. In this paper, a computer simulation study of the hospital discharge process in the University of Wisconsin Hospital and Clinics is presented, which can accurately emulate the discharge process and predict the impact of improvement efforts.

Using this model, we analyzed parameter changes in each procedure (social worker/case manager, pharmacist, and other possible delays) within the discharge process. The pharmacist intervention was not a sensitive factor of the system. The model also identified that both waiting time for order and waiting time for clearance at the end were the system bottlenecks. This result provided a direction of possible solutions to reduce the discharge time.

There are several limitations of this study. The actual discharge process could vary with different physicians' preferences, patient conditions, aftercare facilities, and options of transportation. In this study, generalizing the process into one standardized process might have failed to fully reflect the variations between patients. However, as the study was focused on the medical units only, which had less variations from case to case. Another limitation of the study is that the processing time of some procedures were obtained through discussions with professionals from each party. The estimated time might not have represented the actual procedure time accurately.

In future work, in addition to studying more units in the hospital, we plan to extend the model to include more factors that may affect the discharge process, such as lab testing time; scheduling and coordination among physicians, nurses, and pharmacist; as well as communication with aftercare facilities, primary care physicians, specialists and rehab clinics, patients, and their families. We would also investigate the specific processes for different patient groups who may have special needs and characteristics. Moreover, besides simulation models, developing an analytical model, such as using Markov chain and queueing theory, to characterize the discharge process is always needed and useful.

The successful development of this work will provide hospital professionals and managers a quantitative tool to improve the efficiency of the discharge process in health care delivery.

Acknowledgement

This work is supported in part by NSF Grants No. CMMI-1233807 and CMMI-1536987. The authors would like to thank M. Alderson, M. Brenny-Fitzpatrick, A. Burns, M. Demski, S. Fields, M. Heidi, D. Peterson, M. Thoma, P. Trapskin, and many others from the University of Wisconsin Hospital and Clinics for their support and help in this project.

Competing Interests

None.

References

1. DeFrances, CJ, CA Lucas, VC Buie and A Golosinskiy (2008). 2006 National Hospital Discharge Survey. *US Center for Disease Control Prevention National Health Statistics Report*, 5, pp. 1–20. Hyattsville: National Center for Health Statistics.
2. Srivastava, R, BL Stone, R Patel, M Swenson, A Davies, CG Maloney, PC Young and BC James (2009). Delays in discharge in a tertiary care pediatric hospital. *J Hosp Med*, 4(8), 481–485.
3. Hendy, P, JH Patel, T Kordbacheh, N Laskar and M Harbord (2012). In-depth analysis of delays to patient discharge: A metropolitan teaching hospital experience. *Clin Med,* 12(4), 320–323.
4. Coleman, EA and MV Williams (2007). Executing high-quality care transitions: A call to do it right. *J Hosp Med*, 2(5), 287–290.
5. Jack, BW, VK Chetty, D Anthony, JL Greenwald, GM Sanchez, AE Johnson, SR Forsythe, JK O'Donnell, MK Paasche-Orlow, C Manasseh and S Martin (2009). A reengineered hospital discharge program to decrease rehospitalization: A randomized trial. *Ann Intern Med*, 150(3), 178–187.
6. Powell, ES, RK Khare, AK Venkatesh, BD Van Roo, JG Adams and G Reinhardt (2012). The relationship between inpatient discharge timing and emergency department boarding. *J Emerg Med*, 42(2), 186–196.
7. Forster, AJ, HJ Murff, JF Peterson, TK Gandhi and DW Bates (2003). The incidence and severity of adverse events affecting patients after discharge from the hospital. *Ann Intern Med*, 138(3), 161–167.

8. Kripalani, S, AT Jackson, JL Schnipper and EA Coleman (2007). Promoting effective transitions of care at hospital discharge: A review of key issues for hospitalists. *J Hosp Med*, 2(5), 314–323.
9. Harrison, A and M Verhoef (2002). Understanding coordination of care from the consumer's perspective in a regional health system. *Health Serv Res*, 37(4), 1031–1054.
10. Almborg, A-H, K Ulander, A Thulin and S Berg (2009). Patients' perceptions of their participation in discharge planning after acute stroke. *J Clin Nurs*, 18(2), 199–209.
11. Ortiga, B, A Salazar, A Jovell, J Escarrabill, G Marca and X Corbella (2012). Standardizing admission and discharge processes to improve patient flow: A cross sectional study. *BMC Health Serv Res*, 12(1), 1–6.
12. White, CM, AM Statile, DL White, D Elkeeb, K Tucker, D Herzog, SD Warrick, DM Warrick, J Hausfeld, A Schondelmeyer and PJ Schoettker (2014). Using quality improvement to optimise paediatric discharge efficiency. *BMJ Qual Saf*, 23(5), 428–436.
13. Sims, DC, J Jacob, MM Mills, PA Fett and G Novak (2006). Evaluation and development of potentially better practices to improve the discharge process in the neonatal intensive care unit. *Pediatric*, 118(S2), S115–S123.
14. Wong, HJ, RC Wu, M Caesar, H Abrams and D Morra D (2010). Smoothing inpatient discharges decreases emergency department congestion: A system dynamics simulation model. *Emerg Med J*, 27(8), 593–598.
15. Majeed, MU, DT Williams, R Pollock, F Amir, M Liam, KS Foong and CJ Whitaker (2012). Delay in discharge and its impact on unnecessary hospital bed occupancy. *BMC Health Serv Res*, 12(1), 410.
16. Glasby, J, R Littlechild and K Pryce (2006). All dressed up but nowhere to go? Delayed hospital discharges and older people. *J Health Serv Res Policy*, 11(1), 52–58.
17. Armstrong, SH, NR Peden, S Nimmo and M Alcorn (2001). Appropriateness of bed usage for inpatients admitted as emergencies to internal medicine services. *Health Bull*, 59(6), 388–395.
18. Glasby, J and M Henwood (2007). Part of the problem or part of the solution? The role of care homes in tackling delayed hospital discharges. *Br J Social Work*, 37(2), 299–312.
19. Kydd, A (2008). The patient experience of being a delayed discharge. *J Nurs Manag*, 16(2), 121–126.

20. Soskolne, V, G Kaplan, I Ben-Shahar, V Stanger and G Auslander (2010). Social work discharge planning in acute care hospitals in Israel: Clients' evaluation of the discharge planning process and adequacy. *Res Soc Work Pract*, 20(4), 368–379.
21. Jacobson, SH, SN Hall and JR Seisher (2006). Discrete-event simulation of health care systems. *Patient Flow: Reducing delay in healthcare delivery*, RW Hall (ed.), pp. 211–252. Springer, Germany.
22. Brailsford, SC (2007). Advances and challenges in healthcare simulation modeling: Tutorial. *Proc. of the Winter Simulation Conference*,. pp. 1436–1448. Washington: IEEE Press Piscataway, NJ, USA.
23. Gunal, MM and M Pidd (2010). Discrete event simulation for performance modelling in health care: A review of the literature. *J Simulat*, 4(1), 42–51.
24. Wiler, JL, RT Griffey and T Olsen (2011). Review of modeling approaches for emergency department patient flow and crowding research. *Acad Emerg Med*, 18(12), 1371–1379.
25. Zhong, X, M Williams, J Li, SA Kraft and J Sleeth (2016). Primary care redesign: Review and a simulation study at a pediatric clinic. In *Healthcare Data Analytics: From Data to Knowledge to Healthcare Improvement*, H Yang, H. and E Lee, E. (eds.), John Wiley & Sons. pp. 339–426.
26. Hung, GR, SR Whitehouse, C O'Neill, AP Gray and N Kissoon (2007). Computer modeling of patient flow in a pediatric emergency department using discrete event simulation. *Pediatr Emerg Care*, 23(1), 5–10.
27. Rohleder, TR, DP Bischak and LB Baskin (2007). Modeling patient service centers with simulation and system dynamics. *Health Care Manag Sci*, 10(1), 1–12.
28. Hoot, NR, LJ LeBlanc, I Jones, SR Levin, C Zhou, DS Gadd and D Aronsky (2008). Forecasting emergency department crowding: A discrete event simulation. *Ann Emerg Med*, 52(2), 116–125.
29. Brenner, S, Z Zeng, Y Liu, J Wang, J Li and PK Howard (2010). Modeling and analysis of the emergency department at University of Kentucky Chandler Hospital using simulations. *J Emerg Nurs*, 36(4), 303–310.
30. Reynolds, J, Z Zeng, J Li and S-Y Chiang (2010). Design and analysis of a health care clinic for homeless people using simulations. *Int J Health Care Qual Assur*, 23(6), 607–620.
31. Lu, T, S Wang, J Li, P Lucas, M Anderson and K Ross (2012). A simulation study to improve performance in the preparation and delivery of

antineoplastic medications at a community hospital. *J Med Syst*, 36(5), 3083–3089.

32. Wang, J, J Li, K Tussey and K Ross (2012). Reducing length of stay in emergency department: A simulation study at a community hospital. *IEEE Trans Syst Man Cybernetics – Part A*, 42(6), 1314–1322.
33. Zeng, Z, X Ma, Y Hu, J Li and D Bryant (2012). A simulation study to improve quality of care in the emergency department of a community hospital. *J Emerg Nurs*, 38(4), 322–328.
34. Zhong, X, HK Lee, M Williams, SA Kraft, J Sleeth, R Welnick, L Hoschild and J Li (2016). Workload balancing — Staffing ratio analysis for primary care redesign. *Flexible Serv Manuf J*, doi:10.1007/s10696-016-9258-2.
35. Lin, F, W Chaboyer and M Wallis (2009). A literature review of organisational, individual and teamwork factors contributing to the ICU discharge process. *Aust Crit Care*, 22(1), 29–43.
36. Kaboli, PJ, AB Hoth, BJ McClimon and JL Schnipper (2006). Clinical pharmacists and inpatient medical care: A systematic review. *Arch Intern Med*, 166(9), 955–964.

6. Predictive Modeling of Care Demand and Transition

Xuxue Sun*, Zhouyang Lou†, Mingyang Li*,
Nan Kong‡ and Pratik J. Parikh§

*Department of Industrial and Management Systems Engineering, University of South Florida, USA
†School of Industrial Engineering, Purdue University, USA
‡Weldon School of Biomedical Engineering, Purdue University, USA
§Department of Biomedical, Industrial and Human Factor Engineering, Wright State University, USA

Abstract

With rising costs and increasing demands, care utilization is under scrutiny at hospitals and various other care organizations. Better prediction of care utilization for individual patient will help identify at-risk individuals. As a result, it facilitates the translation of effective interventions with precision, and in turns, improves population-level care management outcomes along care continuum. Although characterization and analysis of care utilization is hardly a new problem, advanced predictive modeling techniques have just begun to be utilized in various demand and transition modeling tasks. In this book chapter, we report two studies that focus on the 30-day hospital readmission and time-to-transition

from community to long-term care facility. We describe several predictive analytics methods for the resultant binary classification and survival analysis tasks, respectively. We also introduce key databases commonly used at present.

6.1. Background and Introduction

Due to the rising costs and increasing risks of aging-related diseases and disabilities in the population worldwide, many countries face critical challenges in improving patient outcomes along care continuum, together with innovating medical technologies. The National Institute on Aging estimates that, by 2050, people aged 65 years or older will double in number globally [1]. In USA, approximately 92% of older adults have at least one chronic disease, and 75% of them have two or more [2]. In terms of supporting healthy living, USA ranks last among many industrialized countries [3]. Despite recent advances in assistive technologies, efficient and cost-effective translation of these technologies is lacking in the country. A paradigm shift from acute care to preventive care is called upon to better integrate healthy living and proactive prevention into healthy people's everyday life in order to delay hospitalization. In USA, the top 1% of the spenders account for over 22% of the total health care expenditure and the top 5% account for roughly 30% of the expenditure [1]. In addition to skyrocketing expenses, the nation does not perform well in several other health outcome categories, such as efficiency, access, and equity [3]. There is a wide agreement that delivery system fragmentation is a root cause of many of these problems [4]. Here, fragmentation refers to the severe dearth of reliable communications; transfer of information; coordination of services; and consistency in goals, incentives, and regulations that exist among the different health service organizations (and, to a lesser extent, units within those organizations).

Under the current system, providers spend a few minutes with patients of chronic conditions during every 3–6 months. This provides, at best, "spot reports" regarding an individual's health and disease progress to the attending physician. Once the disease progresses to the point where the patient needs acute care, the

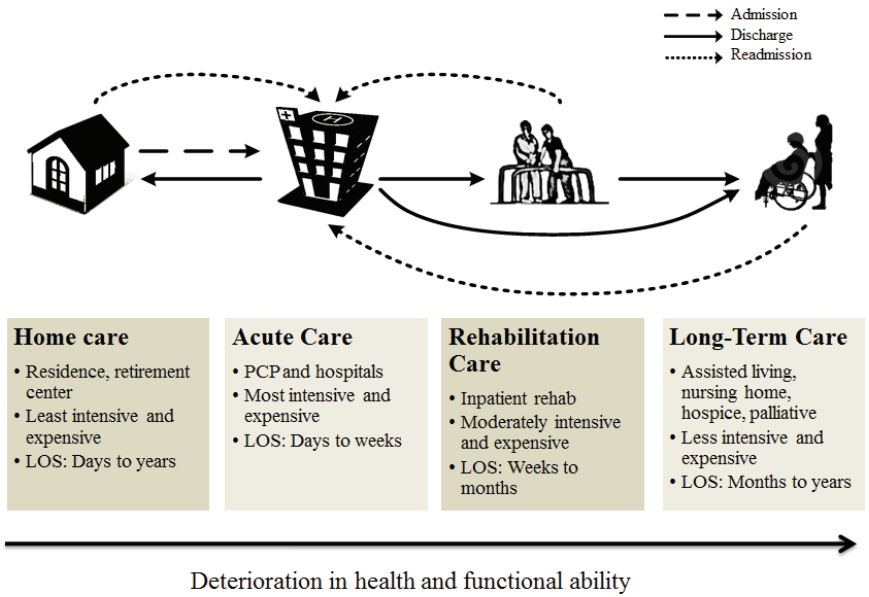


Figure 6.1. The US care continuum.

patient's encounter with an acute hospital starts, which may occur at various points, including emergency departments as well as inpatient wards for recovery from elective surgeries. After inpatient discharge, the patient may transition through diverse care facilities, including outpatient medical care, home-based care, and long-term care. See Fig. 6.1 for an illustration of the US care continuum.

At present, many major health policy organizations, including the American Medical Association (AMA) and the Center for Medicare and Medicaid Services and Joint Commission, strongly advocate careful management of medical conditions along the care continuum. However, these organizations also acknowledge effective coordination of care utilization and transition across departmental and organizational boundaries as a grand challenge for the current US healthcare system [5]. In recent years, operations research and management science communities have been actively involved in furthering the collaboration between data-enabled systems science and health services policy research.

In many predictive analytics research projects, research communities face several distinct methodological challenges. Effective coordination involves good medical decision-making, which is beneficial to every individual, as well as efficient and cost-effective to the overall population. Obviously, physiological heterogeneity among individuals makes such decisions difficult to make. Moreover, environmental and behavioral factors, such as compliance with medications and hospital-induced infections, and healthy lifestyle, can greatly influence the patients' conditions, thus, presenting additional challenges to effective decision-making. Despite that several large-scale projects have been conducted or are currently underway at various model systems of continuous care; the evidence on the interplays between care interventions and patient outcomes remains scarce. Thus, with such scarcity, it is difficult to expand the use of care interventions that have only been proven cost-effective on some small isolate cohort. Finally, given increasing pressure on care spending, it is impossible to offer the best intervention to each individual. Optimal care resource allocation, critical to the financial landscape of a sustainable healthcare system, must account for conflicting interests of various care facilities in the fragmented system.

In this chapter, we will focus on the first challenge, i.e., how to incorporate heterogeneity in predictive modeling of care demand and transition. Successful predictive modeling of the variables is a prerequisite to further success to operational excellence and policy refinement on transitions along the care continuum and the related resource planning decisions at various facilities.

In the remainder of this chapter, we present two sample research projects on the predictive modeling. Section 6.2 describes a binary classification study for predicting the 30-day hospital readmission, a hospital's key inpatient care quality indicator at present. Section 6.3 describes a Bayesian survival analysis study to characterize probabilistically each individual's time-to-transition from the community to long-term care and to quantify the influence of observed factors and latent heterogeneity due to unobserved/unknown factors. We draw conclusions and outline future research focuses in Section 6.4.

6.2. A Classification Study for 30-Day Hospital Readmission Prediction

6.2.1. *Summary of the study*

Unavoidable hospital readmissions raise healthcare costs and cause significant distress to the providers and the patients. It is, therefore, of great interest to health policy makers and administrators to predict which patients are at risk of being readmitted to the hospital. Like most of the existing studies, we relied on statewide administrative data retrospectively. We incorporated social and demographic determinants of health, and explored a comprehensive list of comorbidity variables.

In this work, we focused on improving classification performance, with more sophisticated statistical modeling techniques. We identified key determinants of readmission and developed conditional logistic regression models. That is, with one or several of the identified key determinants, we developed a distinct logistic regression model for each data stratum, derived by stratifying the original dataset with respect to the identified determinants. We further explored the effect of interacting variables in the logistic regression modeling.

Our comparative studies showed that developed conditional logistic regression models outperformed several standard classification models (e.g., straightforward logistic regression, step-wise logistic regression, random forest, and support vector machine). They are expected to offer insights into further development of prediction models in this area.

6.2.2. *Current landscape in practice*

In USA, it is common for patients to be readmitted to acute care hospitals after a short amount of time post hospital discharge [6–9]. Hospital readmissions incur unnecessary costs. It is estimated that preventable readmissions for Medicare patients alone cost $17 billion annually [10], which is equivalent to more than 10% of Medicare benefit payment for hospital inpatient services [11]. Hence, a reduction in the number of readmissions is critical

to the US public funding agencies, such as Medicare and Medicaid, whose spending has increased rapidly in recent years, with the rising number of aging population and high prevalence of chronic conditions. Meanwhile, hospital readmissions present significant but unnecessary burden to care utilization and can, thus, serve as an important indicator of poor health care quality and efficiency [12–16]. Hence, there is a clear impetus for hospitals to reduce readmissions. The US Center for Medicaid and Medicare Services (CMS) provides reputational pressure and financial incentives to hospitals to reduce preventable readmissions [17]. In 2009, the CMS began publicly reporting 30-day risk-standardized readmission rates for health failure, acute myocardial infarction, and pneumonia [18–21]. More recently, as part of Affordable Care Act (ACA), it started to cease the reimburse payment to hospitals for any 30-day readmission incidence of Medicare beneficiaries that is deemed to be preventable.

Many readmissions can be prevented with effective discharge management (e.g., [22–24]). Critical to the development of these programs is the understanding of influential factors causing readmissions. These factors include patients' diagnosis and severity of illness, patients' behavior such as adherence to discharge instructions, and the availability and quality of post-discharge care. While standalone observational studies have shown several management strategies to be effective in reducing preventable readmissions, these studies hardly provide much transferrable insight to other hospitals, especially when dealing with their own programmatic implementation issues. For detailed cost-benefit studies, hospitals must develop accurate readmission prediction models.

6.2.3. *State of the art in academic research*

In the academic research literature, most of the studies use descriptive, particularly discriminatory, analysis to decipher the influence of certain disease or disease class by one or few selected risk factors (such as age [25–28], sex [29], income [30], education background

[31], and insurance type [32], as well as comorbidity [29, 33]) on the readmission incidence. The rest of the studies are exploratory. A symmetric review prepared by the Veterans Health Administration [34] summarized 26 unique studies presented in English, developed before 2011 and found via searching MEDLINE, CINAHL, Cochrane Library, and EMBASE. Among these models, three models [35–37] were derived and tested based on large US population. Two other studies [38, 39] were derived and tested at multiple centers in a single state.

Additionally, in the systematic review [34], the authors commented that most of the up-to-date readmission risk prediction models had less satisfying discriminative ability (i.e., the c-statistics ranges from 0.55 to 0.8, with lower values in models purely based on administrative data). The following were the two main reasons for this deficiency:

i. Relatively poor quality of administrative data but high cost associated with collecting detailed clinical data at the inpatient stage (e.g., daily vitals) and social/behavior data at the post-discharge stage (e.g., whether to have informal care giver);
ii. limited success on applying alternative statistical machine learning methods other than standard logistic regression.

To our knowledge, these are most of the existing studies, if not exhaustively reviewed. Natale *et al.* [40] investigated a decision tree model and compared it with standard logistic regression models. Lee [41] compared three models: logistic regression, decision tree, and neural networks. Hosseinzadeh *et al.* [42] compared a decision tree classifier and a Naïve Bayes classifier. For other systematic reviews on the predictive modeling of readmission incidence, we refer to Desai *et al.* [43] and van Walraven *et al.* [44]. Lack of additional observations on potentially influential risk factors and accurate classifier development has led to less-than-satisfied performance on readmission risk prediction. Meanwhile, existing classifiers, treated at best as blackboxes, are not easy to be implemented in clinical practice.

6.2.4. *Data description*

The inpatient medical records we analyzed were acquired from the State Inpatient Database (SID) of California. SIDs are a powerful collection of data sets from participating providers throughout USA. These datasets contain data from almost 90% of all American hospital inpatient discharges. SID includes a set of patient data (e.g., patient's age, gender, race, and payer status), as well as information related to initial acute condition (e.g., ICD 9 codes) and inpatient care (e.g., discharge date, readmission date, and disposition location). These patient data provide the basis of specifying readmission outcomes and offer a large set of predictors to choose.

SIDs are a part of the Healthcare Cost and Utilization Project (HCUP). HCUP, sponsored by the Agency for Healthcare Research and Quality (AHRQ), is the largest collection of nationwide and state-specific longitudinal hospital care data in the USA. AHRQ/HCUP databases are derived from administrative data and contain encounter-level, both clinical and nonclinical information. These databases allow research on a wide range of health problems. For more information on HCUP and SID, please see the AHRQ webpage of HCUP at www.ahrq.gov/research/data/hcup/.

In this study, we extracted relevant SID records of Medicare beneficiaries. Much of the current healthcare debate is centered at how to provide public funding to purchase care services. Hence, the data from Medicare and Medicaid beneficiaries have been of great interest to health service researchers.

6.2.5. *Data modeling methodology*

Our study compared several classification methods to predict 30-day readmissions after hospitalization. They are standard logistic regression, random forest [45], support vector machines [46], and conditional logistic regression [47]. With only two labels (readmitted or not readmitted), the problem falls into the category of binary classification. In this study, we first constructed the dataset to be analyzed with necessary data extraction. Next, we identified influential

risk factors for dataset division and conditional regression modeling. Finally, we compared the developed models with alternative ones to justify the contribution of the study.

6.2.5.1. *Data preparation*

We examined the inpatient discharge records collected in year 2010 from California. The original data set contained 3,970,921 patient records. We selected the patient cohort by the following criteria: (1) heart failure (HF), which was the primary diagnosis, as identified by validated International Classification of Disease, Ninth Revision, diagnosis codes, i.e., ICD-9 code; (2) age 65 and older; (3) Medicare as primary payer; (4) primary residence in California; (5) discharged during January–November 2010; (6) not transferred to another hospital immediately after 1–2 days; and (7) discharged to home self-care, home health care, or nursing home care. In addition, we removed records with missing, errand information or very low frequency. For an illustration of the entire data extraction and cleaning procedure, please see Appendix B in Zhu *et al.* [48].

In summary, the relevant SID records we extracted were those associated with California Medicare HF patients discharged within the first 11 months of 2010. We considered only the prediction of first readmission incidence within 30 days. Three main reasons behind choosing HF patients: First, there seems to be more variation in 30-day readmission incidence among HF patients as opposed to patients from other major disease groups. Second, HF patients outnumbered those in other major disease groups, which could help ensure the model validity. Third, to most of the care organizations, reducing HF patient readmission was of high priority. With regards to criterion no. 6, we excluded transfers to another hospital with a short stay because this likely indicates that the studied hospital was unable to provide adequate care to the transfer-out patients. With regards to criterion no. 7, we did not consider several discharge options for their low occurrences. Similarly, we did not include records of Native Americans and patients who could not tell whether they were Hispanic or not. With the data extraction, the records of

22,410 patients remained, and 17,434 of them belonged to patients not readmitted within 30 days.

We selected a comprehensive set of independent variables (or features) for the predictive model development, which included such clinical variables as the numbers of chronic conditions and procedures, and such administrative variables as discharge location (e.g., routine, transfer, and home health care). The selected variables also included many commonly studied demographic and socioeconomic variables. In addition, 22 binary variables in total were included to indicate whether a patient had a particular comorbidity, e.g., acute kidney injury. A few other AHRQ comorbidity measures, e.g., CM_ULCER—an indicator of peptic ulcer disease, excluding bleeding, were not included since they were unable to create balanced dichotomy in terms of readmission.

Table 6.1 reports the characteristics of the cohort with respect to the selected features. Distributions of categorical variables are

Table 6.1. Characteristics of the cohort (n = 22,410).

	Readmission Within 30 Days	
Patient Characteristics	**No (n = 17,434 [77.8%])**	**Yes (n = 4,976 [22.2%])**
Age (years) (mean 80.6 ± 7.8)	80.6 (7.8)	80.7 (8.0)
Gender		
Men (45.3%)	7,836 (45.0%)	2,321 (46.6%)
Women (54.68%)	9,598 (55.1%)	2,655 (53.4%)
Race		
White (64.5%)	11,310 (64.9%)	3,153 (63.3%)
Black (8.2%)	1,366 (7.8%)	477 (9.6%)
Hispanic (17.2%)	2,994 (17.2%)	852 (17.1%)
Asian/Pacific Islander (8.3%)	1,459 (8.4%)	410 (8.2%)
Other (1.7%)	305 (1.8%)	84 (1.7%)

(*Continued*)

Table 6.1. (*Continued*)

Patient Characteristics	Readmission Within 30 Days: No (n = 17,434 [77.8%])	Readmission Within 30 Days: Yes (n = 4,976 [22.2%])
Resident Location		
Metropolitan (97.5%)	16,992 (97.5%)	4,863 (97.7%)
Micropolitan (1.6%)	277 (1.6%)	71 (1.4%)
Non-CBSA (0.9%)	165 (1.0%)	42 (0.8%)
Median Income		
Low (25.2%)	4,350 (25.0%)	1,299 (26.1%)
Medium (25.2%)	4,408 (25.3%)	1,227 (24.7%)
High (26.0%)	4,515 (25.9%)	1,308 (26.3%)
Very high (23.7%)	4,161 (23.9%)	1,142 (23.0%)
Admission Source		
Emergency Department (85.2%)	14,782 (84.8%)	4,301 (86.4%)
Another Hospital's ED (0.8%)	144 (0.8%)	41 (0.8%)
Other Health Facility (1.2%)	217 (1.2%)	47 (0.9%)
Routine (12.8%)	2,291 (13.1%)	587 (11.8%)
Disposition of Patient at Discharge		
Routine (57.3%)	10,457 (60.0%)	2,390 (48.0%)
Transfer to Other Facilities (20.0%)	2,959 (17.0%)	1,501 (30.2%)
Home Health Care (22.8%)	4,018 (23.1%)	1,085 (21.8%)
Number of Chronic Conditions (mean 8.3 ± 3.0)	8.2 (2.9)	8.7 (3.0)
Number of Procedures (mean 1.1 ± 1.9)	1.0 (1.8)	1.2 (2.2)
Weekend Admission		
Yes (22.5%)	3,891 (22.3%)	1,149 (23.1%)
No (77.5%)	13,543 (77.7%)	3,827 (76.9%)
Do Not Resuscitate		
Yes (14.3%)	2,525 (14.5%)	672 (13.5%)
No (85.7%)	14,909 (85.5%)	4,304 (86.5%)
Length of Stay (mean 4.6 ± 4.7)	4.6 (4.7)	4.7 (4.6)

expressed frequencies and continuous variables as mean (standard deviation). The listed characteristics, together with the comorbidity indicators, were the features for prediction.

In our study, high-level unbalance appeared in the extracted data. We, tested three common imbalance correction techniques: under-sampling, over-sampling, and different error cost [49]. Our study suggested that under-sampling was a more viable option that the other two. After the correction, we obtained a total of 9,952 cases, half with positive response and half with negative response.

6.2.5.2. *Ad-hoc conditional logistic regression modeling*

In our preliminary experiments, we employed the standard logistic regression, random forest (RF), and support vector machines (SVM), with the use of R packages glm and e1071. The experiment results did not imply much promise in readmission risk prediction. Moreover, the complexity of RF and SVM, which was more than decision tree and naïve Bayes, would not easily convert the "black-box" type decision-making procedure to meaningful intelligence in practice.

We speculated that the poor performance may be due to the profound heterogeneity in the patient population. Subsequently, we stratified the entire patient into several subgroups hoping that each of the subgroups was more homogenous and, thus, could enable the development of better classifiers. We employed the decision tree technique to identify 2–3 variables for the stratification. We, then, applied logistic regression on each population subgroup. Essentially, we combined the advantages of regression and decision tree with ad-hoc stratification variable selection.

After reviewing the first layer of the decision tree, we observed that the following three variables appeared most frequently: DISPUniform (i.e., disposition location after discharge), NPR (i.e., number of ICD-9-CM procedures), and NCHRONIC (i.e., number of chronic conditions). With the decision tree analysis, we also acquired the threshold value on each of the three variables. For DISPUniform, patients with values of 1 and 6 were in one stratum

and patients with values of 5 were in the other stratum; for NPR, patients with values less than or equal to 4 were in one stratum and patients with values more than 4 were in the other; and for NCHRONIC, patients with values less than or equal to 7 were in one stratum and patients with values more than 7 were in the other. We also noticed that the above three variables showed superior discriminatory ability in the logistic regression. We, thus, stratified the patient dataset based on each of the three variables identified above and applied logistic regression on each of the subsets.

Through further experiments, we noticed modest improvement on the prediction accuracy. We speculated that some level of heterogeneity still existed in most of the data subsets. We, thus, continued our exploratory stratification by using the combinations of the variables on the first two layers of the decision tree, instead of only the variable from the first layer. As a result, we obtained four mutually exclusive data strata (see Fig. 6.2 for the stratification).

Note that, in Fig. 6.2, the variables on the left and right branches of the second layer are typically not the same. Also, from the above exploration, we concluded that the stepwise variable selection would not lead to improved classification. We speculated that the less-than-satisfied performance arose from missing of higher-order modeling. Thus, within each of the four strata, we deployed logistic regression on additional variables, capturing the pairwise interactions between the original variables.

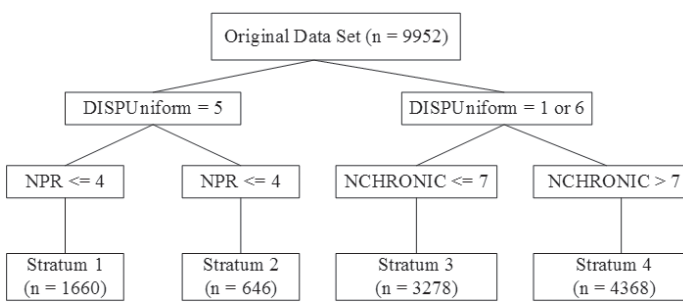


Figure 6.2. Overview of the data stratification.

In summary, we developed the following classification models: 1) standard logistic regression (whole dataset), 2) stepwise logistic regression (whole dataset), 3) random forest (whole dataset), 4) support vector machine (whole dataset), 5) conditional logistic regression (with two strata based on each of the top 3 variables), 6) conditional logistic regression (with four strata derived from a decision-tree based rule set), and 7) conditional logistic regression (employment of (6) with incorporation of additional pairwise interacting variables). For convenience, we call them LR, SLR, RF, SVM, CLR1, CLR2, and CLR3 in the remaining chapter. For RF, SVM, and CLR1, we also considered performing them on two different variable selection sets: all original variables and selected ones through conditional logistic regression. To compare the above classification models, we performed cross validation and assessed each model's classification accuracy. We describe the comparative study in the following section.

6.2.6. *Analysis results*

In the cross validation, we split the original data set into two subsets, i.e., 70% of the data in the training set and 30% in the test set. Our results suggested that

1. Conditional logistic regression made modest improvement in classification accuracy over more straightforward classification methods.
2. Among different ideas on conditional logistic regression, CLR2 made modest improvement over CLR1 and CLR3 further made slight improvement over CLR2.
3. It was beneficial to explore the use of decision tree modeling to guide the cohort stratification and it was possibly beneficial to investigate the inclusion of interacting variables as well.
4. The conditional logistic regression achieved improved sensitivity over the standard logistic regression. This improvement exceeded 10% with both CLR2 and CLR3, especially in certain strata. However, the improved sensitivity might be associated with inferior specificity.

Table 6.2. Classification model comparison.

Classification Model		Prediction Accuracy*
LR: standard logistic regression		0.547
SLR: stepwise logistic regression		0.539
RF: random forests	all original variable	0.577
	only variables selected via SLR	0.574
SVM: support vector machine		0.560
CLR1: conditional logistic regression with 3 influence prediction variables	DISPUniform	0.548**
	NCHRONIC	0.564**
	NPR	0.576**
CLR2: conditional logistic regression with 4 data strata based on the first two layers of the decision tree		0.608**
CLR3: CLR2 + consideration of interacting variables based on identified influence ones in CLR2		0.615**

*Prediction Accuracy = (True Positive + True Negative)/Total # of Subjects based on the test set.

**An optimal threshold was identified for each data stratum and the prediction accuracy is the combined measure over the multiple strata.

Table 6.2 reports the results of comparing different classifiers as above.

6.3. A Bayesian Modeling Study of Community Dwelling Duration Prior to Long-Term Care

6.3.1. *Summary of the study*

The time-to-transition from the community to the long-term care (LTC) system is an important measure to reflect the demand for LTC from community-dwelling elderly people. A variety of factors, from the sides of both care provider and care recipient, may influence such transition of elderly people. However, for many publically available healthcare data in studying care utilization, such as administrative claims data, detailed health information at the individual level is limited. Observing and identifying all factors that tend to influence

the time-to-transition become challenging, if not impossible. It is also important to quantify the latent heterogeneity, which is caused by the influence of missing and/or unobserved factors and varies among individuals.

In this work, we focused on developing Bayesian survival models to investigate the time-to-transition of elderly people from the community to the LTC systems. A probabilistic measure was first established to quantify the instantaneous transition rate of elderly individuals over time. Both influences of the observed factors and individuals' latent heterogeneity on transition rate were further quantified in a simultaneous manner. The Bayesian model formulation allowed jointly estimating latent heterogeneity of all individuals and providing rich quantification to the effects of the observed factors. New features were further extracted based on the available data to reduce the latent heterogeneity successfully. The proposed work provide a methodological framework to better investigate the transition rates among multiple healthcare settings.

6.3.2. *Current landscape in practice*

In the past few years, the transition among various health-care settings, such as acute care settings [50–52] and long-term care (LTC) settings [53–61], has been extensively studied. A better understanding of transition patterns of elderly people among various healthcare settings will help healthcare professionals and policymakers better identify complex care needs and facilitate better decision-making in care services, workforce management, and payment policies. With the prevalence of elderly people with disabilities due to rapid aging of the baby boomer generation, the excess LTC demand poses unprecedented challenges on capacity shortage and public financing of the current LTC systems. LTC demands among individuals are highly heterogeneous, partially because elderly people may be at risk or are suffering from various types of chronic diseases, injuries and impairments. An appropriate assessment of individual heterogeneity on LTC demands is critical to healthcare decision-making and healthcare policy deliberation. It can be

helpful to annihilate the waste in health-care delivery, to control the cost of care services, as well as to improve the quality of care [62–64]. Ineffective decisions of LTC transitions can lead to unexpected outcomes, such as inappropriate treatments, delays in diagnosis, severe adverse events, and increased costs [59]. In the process of rebalancing LTC resources in different LTC services settings, ranging from institutional settings (e.g., nursing homes) to non-institutional settings (e.g., assisted living facilities and in-home care), careful investigation of individual heterogeneity on LTC demand and its transition in various settings will enable health assurance and quality of care for elderly people. For instance, for an elderly person with minor cognitive disability, intensive care services will not only be cost ineffective but may also impair self-independence and privacy. On the other hand, for an elderly person without cognitive disability but moderate physical disability, inadequate care services will increase inconvenience of daily livings and may result in negative consequences, such as fall and injuries. Overall, better modeling and quantification of individual heterogeneity on LTC demand allow the policymakers and researchers to develop a viable option for elderly people with a more consumer-directed LTC system.

6.3.3. *State-of-the-art academic research*

A variety of observed factors influencing the demand for LTC from care receivers as well as care providers have been studied. From the side of care receivers, demographic characteristics such as age [65], gender [65, 66] and race [67, 68]; health condition, including both physical and mental health [69]; as well as economic conditions and financial support [70–72] of people in community-based facilities, have an unneglectable impact on the demand for LTC care. From the side of care providers, the demand for LTC is also affected by the capacity of LTC facilities [64], service price, and service management [58, 60, 69]. To characterize individual heterogeneity, many existing studies summarize descriptive statistics [73, 74] and/or perform hypothesis testing [58] on multiple groups of individuals with different

characteristics [60, 75]. Different statistical models, such as Poisson regression [55], discrete time hazard function [76], multi-state model [77], and proportional hazards model [78], have also been developed to study individual heterogeneity of LTC demands. As aforementioned, only partial information can be obtained from healthcare claims data while the unobserved and unavailable information still play a significant role in LTC demands. There is limited research to consider influence of unobserved factors [69, 79, 80]. However, to the best of our knowledge, no previous studies have been presented for individualized modeling of LTC demands in terms of time-to-transition [81] and jointly quantifying observed and unobserved individual heterogeneity.

6.3.4. *Data description*

To demonstrate the capability and effectiveness of our proposed method, a real case study is carried out, based on the Florida's Medicare and Medicaid claims data [82]. The available subset of data consists of healthcare service records of 217 elderly individuals and their individual characteristics, such as ethnicity and activities of daily living (ADL) scores.

Figure 6.3 shows time-to-transition observations of five individuals, which clearly demonstrate the existence of heterogeneity among individuals. To explain such heterogeneity, based on variable screening and selection techniques, several potentially relevant covariates are considered, including x_1 (ADL score), x_2 (age), x_3, (ethnicity), x_4 (marriage status), and x_5 (cancer indicator).

6.3.5. *Data modeling methodology*

To model the time-to-transition of the population of N elderly individuals, denote T_{ij} as the j^{th} time-to-transition of individual i, $i = 1, \ldots, N$ and $j = 1, \ldots, m_i$, where m_i is the total number of visits to the LTC facilities of individual i. The proposed time-to-transition model is given by $r_i(t) = r^{\text{b}}(t) \exp(\Gamma_i + \boldsymbol{\beta}^{\text{T}}\mathbf{x})$, where $r_i(t)$ is a time-variant probabilistic measure, which characterizes an individual's tendency of transiting from the community to a LTC facility. $r^{\text{b}}(t)$ is the population

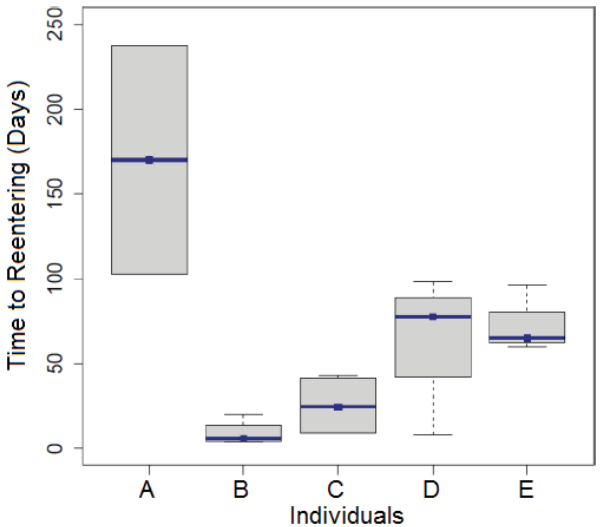


Figure 6.3. Heterogeneity of individual time-to-transition.

average tendency of transiting from the community to a LTC facility in the absence of the influence of **x**. In this paper, $r^{b}(t)$ is specified as Weibull hazard function $r^{b}(t) = \lambda k t^{k-1}$ due to its great flexibility and good interpretation, where λ is the rate parameter and k is the shape parameter. **x** is vector of covariates, which represent observed transition-related covariates, such as individual characteristics and health conditions. $\boldsymbol{\beta}$ is a vector of covariate coefficients that quantifies the effects of **x** on $r_i(t)$. Γ_i is a latent random variable that quantifies the individual latent heterogeneity. To perform model estimation, the conventional non-Bayesian estimation method, e.g., maximization likelihood estimation (MLE), tends to maximize the marginal likelihood function, where individual specific latent random variable Γ_i is not estimable. In addition, in non-Bayesian estimation methods, point estimate is often obtained (in the least square estimation) and confidence intervals in MLE are approximated based on the large sample size theory. To make exact inference, provide rich estimation summary, and realize joint estimation of {λ, k, $\boldsymbol{\beta}$} and Γ_i's, Bayesian estimation is considered and Markov Chain Monte Carlo (MCMC) sampling [83] is performed.

6.3.6. *Analysis results*

Bayesian method is performed to jointly quantify the influence of observed covariates as well as the latent heterogeneity. A rich information summary can be obtained to quantify the uncertainty of estimated parameters. The Bayesian estimation results with posterior mean/median and 95% credible interval are summarized in Table 6.3.

Based on 95% credible interval, β_1 is significant and has a positive effect on the transition from the community to LTC. It indicates that an individual who has larger ADL value is more likely to enter a LTC facility from the community, and thus, the time-to-transition will become shorter. β_3 is significant and has a negative effect on the transition. It implies that if an individual with white ethnicity will have a longer time to enter a LTC facility. Although other covariates are not significant, based on 95% credible interval, Bayesian estimation results can still provide rich information. For instance, 0.25 posterior quantile of β_2 is positive. It indicates that there is at least 75% assurance to assert that as age increases, an individual is more likely to enter a LTC facility. k is significantly larger than 1, implying that an individual will be more likely to enter a LTC facility as time increases. λ can be interpreted as the baseline average tendency of transitioning from the community to a LTC for all individuals in the absence of influence of Γ and $\mathbf{x}$.

Table 6.3. Bayesian estimation results of the proposed model.

Parameters	Mean	2.5%	25%	50%	75%	97.5%
β_1	0.39318	0.07694	0.28574	0.38978	0.49552	0.72514
β_2	0.01143	–0.00535	0.00576	0.01172	0.01720	0.02749
β_3	–0.40148	–0.70987	–0.49777	–0.40020	–0.30290	–0.11235
β_4	–0.19960	–0.66988	–0.34047	–0.19560	–0.05187	0.21512
β_5	0.00367	–0.58297	–0.18599	0.01418	0.19854	0.56175
Γ_A	–0.43220	–1.38593	–0.71779	–0.38740	–0.10206	0.28063
Γ_B	0.79342	–0.16895	0.33950	0.76425	1.20855	2.00327
λ	0.00532	0.00113	0.00324	0.00531	0.00906	0.02372
k	1.10187	0.93928	1.04483	1.10386	1.15861	1.26214

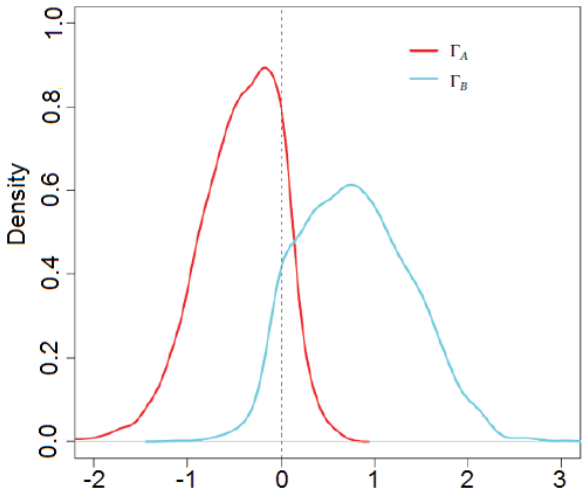


Figure 6.4. Posterior density plots of individual latent heterogeneity.

In addition to quantifying the observed covariates, Bayesian estimation also allows simultaneous quantification of individuals' latent heterogeneity. Take individuals A and B in Fig. 6.3 as an example. Their estimated posterior densities of individual latent heterogeneity, i.e., Γ_A and $\Gamma_{A,}$ are shown in Fig. 6.4. Γ_A is more concentrated on negative values, while Γ_B is more concentrated on positive values, which indicate that individual A is less likely to enter a LTC facility and, thus, has a longer time-to-transition, while individual B is more likely to enter a LTC facility and has a shorter time-to-transition. The estimation results are consistent with the real data records shown in Fig. 6.3. The individual latent heterogeneity of all individuals can be simultaneously obtained by the proposed method. Figure 6.5 shows estimation results of all Γ_i's. A positive value of Γ indicates that an individual will be more likely to enter a LTC facility and, thus, has a shorter time-to-transition and vice versa.

To further explain such latent heterogeneity, two new covariates, namely the total number of previous visits to LTC facilities and a previous hospital discharge indicator, are extracted. After including such newly extracted covariates, Fig. 6.6 shows individual latent heterogeneity plots of the updated model. Compared to Fig. 6.5,

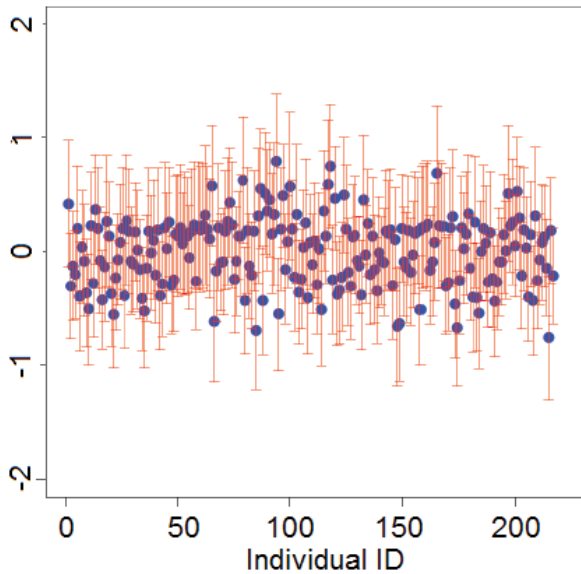


Figure 6.5. Estimated latent heterogeneity of individual time-to-transition.

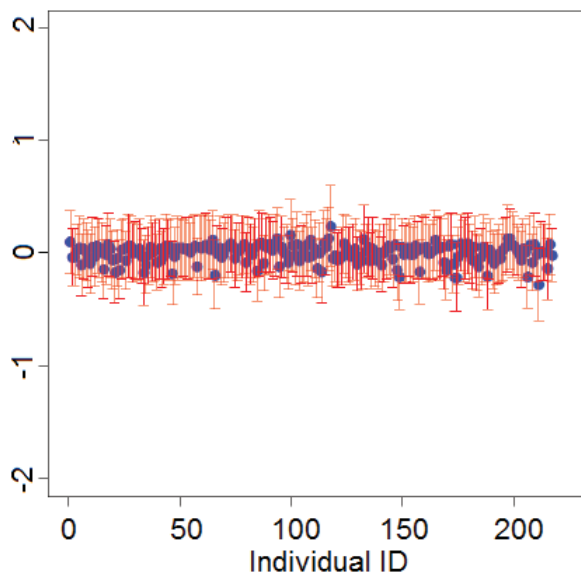


Figure 6.6. Estimated latent heterogeneity of individual time-to-transition after including additional extracted covariates.

more Γ_i's in Fig. 6.6 are approaching to 0, indicating that some of the individual latent heterogeneity can be further explained by these two extracted covariates. The extracted covariates are not directly available and observed. They are calculated and extracted based on the raw health claims data. Thus, they serve as "latent covariates" from observed data to influence individual's time-to-transition from community to LTC.

6.4. Conclusions and Future Work

In this chapter, we have presented two sample research projects. In the first project, we developed a binary classifier based on the conditional logistic regression model to predict 30-day hospital readmission incidence based on publicly available state-/nation-wide administrative data. To develop our model, we employed decision trees to identify influential risk factors and used several of them to stratify the dataset to achieve better homogeneity among the patient records. We conducted comparative studies to test several binary classifiers and showed improvement of our developed model over the existing models from the literature. A real case study based on California's Medicare heart failure patients' inpatient records was conducted to demonstrate the validity of the proposed method. In the second project, we proposed a Bayesian latent heterogeneity modeling and quantification approach for characterizing elderly individuals' time-to-transition from the community to LTC systems that reflects the LTC demand of community-dwelling elderly people. It allowed joint estimation and rich quantification of the influences of both observed covariates on the transition and individuals' latent heterogeneity. New covariates were further extracted from raw data as latent covariates to reduce the individuals' unexplained latent heterogeneity. A real case study based on Florida's Medicare and Medicaid claims data was conducted to demonstrate the validity of the proposed method.

To both projects, administrative claims data are reliable resources and provide important cross-sectional and longitudinal information on health care demand and transition. However, both projects are

limited by the in-depth richness of such data. For example, in the first project, additional risk factors can be identified from inpatient clinical and service information, as well as post-discharge care management information. In the second project, important transition-related factors can be extracted from more detailed health condition information to explain heterogeneity of the time-to-transition. Due to the data unavailability, the proposed model takes into account the influence of a variety of unobserved/missing/unknown factors by quantifying them as unobserved heterogeneity. Meanwhile, the number of health care visits for each individual is limited. It becomes challenging to estimate individual specific models based on a small sample size of data. In addition to the data availability issue, we well expect the fact that publicly available administrative data can be quite noisy — having coding errors and entry inconsistencies, which presents another limitation on the two studies.

In the future, it will be interesting to characterize and predict the quantity time-to-transition among different health-care settings (e.g., LTC and acute care). In addition, to develop better models, certain model assumptions (e.g., Weibull distribution on the degradation in the second project) can be relaxed to improve the modeling flexibility. Furthermore, care transition and utilization modeling will be further integrated with optimization models for better decision-making in care delivery. Finally, we will incorporate with more detailed information of patients into the future modeling, not only the detailed information from the clinical aspect but also from the patient choice aspect.

Acknowledgments

This work was supported in part by the National Science Foundation under Grant No. 1405357, and the University of South Florida Research and Innovation Internal Awards Program under Grant No. 0114783.

References

1. He, W, D Coodlind and P Kowal (2016). *An Aging World: 2015*. U.S. Census Bureau, International Population Reports, P95/16-1. U.S. Washington: Government Publishing Office.

2. National Council on Aging. (2016). Healthy aging facts. https://www.ncoa.org/resources/fact-sheet-healthy-aging/. [Accessed November 28, 2016].
3. Davis, K, C Shoen, SC Schoenbaum, MM Doty, AL Holmgren, JL Kriss and KK Shea (2010). *Mirror, Mirror on the Wall: How the Performance of the U.S. Health Care System Compares Internationally, 2010 Update*. New York: The Commonwealth Fund.
4. Elhauge, E (2010). *The Fragmentation of U.S. Health Care*. Oxford: Oxford University Press.
5. The National Transitions of Care Coordination (NTOCC) (2011). Improving transitions of care. http://www.ntocc.org. [Accessed December 20, 2016].
6. Lindenauer, PK, SM Bernheim, JN Grady, Z Lin, Y Wang, Y Wang, AR Merrill, LF Han, MT Rapp, EE Drye, SL Normand and HM Krumholz (2010). The performance of US hospitals as reflected in risk-standardized 30-day mortality and readmission rates for Medicare beneficiaries with pneumonia. *J Hosp Med*, 5(6), E12–E18.
7. Bernheim, SM, JN Grady, Z Lin Y Wang, Y Wang, SV Savage, KR Bhat, JS Ross, MM Desai, AR Merrill, LF Han, MT Rapp, EE Drye, SL Normand and HM Krumholz (2010). National patterns of risk-standardized mortality and readmission for acute myocardial infarction and heart failure: Update on publicly reported outcomes measures based on the 2010 release. *Circ Cardiovasc Qual Outcomes*, 3(5), 459–467.
8. Ross, JS, J Chen, Z Lin, H Bueno, JP Curtis, PS Keenan, SL Normand, G Schreiner, JA Spertus, MT Vidán, Y Wang, Y Wang and HM Krumholz (2010). Recent national trends in readmission rates after heart failure hospitalization. *Circ Heart Fail*, 3(1), 97–103.
9. Centers for Medicare & Medicaid Services (2011). *Medicare Hospital Quality Chartbook 2011: Performance Report on Readmission Measures for Acute Myocardial Infarction, Heart Failure, and Pneumonia*. Washington: Centers for Medicare & Medicaid Services.
10. Jencks, SF, MV Williams and EA Coleman (2009). Rehospitalizations among patients in the Medicare fee-for-service program. *N Engl J Med*, 360, 1418–1428.
11. Centers for Medicare and Medicaid Services. The Medicare and Medicaid Statistical Supplement. 2013 Edition. http://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/MedicareMedicaidStatSupp/2013.html. [Accessed July 23, 2014].

12. Ashton. CM and NP Wray (1996). A conceptual framework for the study of early readmission as an indicator of quality of care. *Soc Sci Med*, 43(11): 1533–1541.
13. Ashton, CM, DH Kuykendall, ML Johnson, NP Wray and L Wu (1995). The association between the quality of inpatient care and early readmission. *Ann Intern Med*, 122(6), 415–421.
14. Ashton, CM, DJ Del Junco, J Souchek, NP Wray and CL Mansyur (1997). The association between the quality of inpatient care and early readmission: A meta-analysis of the evidence. *Med Care*, 35(10), 1044–1059.
15. Medicare Payment Advisory Commission (2007). *Promoting Greater Efficiency in Medicare, Report to the Congress*. Washington: MedPAC. http: //www.medpac.gov/documents/Jun07_EntireReport.pdf [12 September 2012].
16. Centers for Medicare & Medicaid Services Hospital Pay-for-Performance Workgroup (2007). US Department of Health and Human Services Medicare hospital value-based purchasing plan development, issues paper, 1st public listening session. Washington: Centers for Medicare &Medicaid Services. http: //www.cms.gov/Medicare/Medicare-Fee-for-Service-Payment/AcuteInpatientPPS/downloads/hospital_VBP_plan_issues_paper.pdf [Accessed 12 September 2012].
17. Kocher, RP and EY Adashi (2011). Hospital readmissions and the Affordable Care Act: Paying for coordinated quality care. *JAMA*, 306(16), 1794–1795.
18. National Quality Forum (2008). *National Voluntary Consensus Standards for Hospital Care 2007: Performance Measures — A Consensus Report*. Washington: National Quality Forum.
19. Keenan, PS, SL Normand, Z Lin, EE Drye, KR Bhat, JS Ross, JD Schuur, BD Stauffer, SM Bernheim, AJ Epstein, Y Wang, J Herrin, J Chen, JJ Federer, JA Mattera, Y Wang and HM Krumholz (2008). An administrative claims measure suitable for profiling hospital performance on the basis of 30-day all-cause readmission rates among patients with heart failure. *Circ Cardiovasc Qual Outcomes*, 1(1), 29–37.
20. Krumholz, HM, Z Lin, EE Drye, MM Desai, LF Han, MT Rapp, JA Mattera and SL Normand (2011). An administrative claims measure suitable for profiling hospital performance based on 30-day all-cause readmission rates among patients with acute myocardial infarction. *Circ Cardiovasc Qual Outcomes*, 4(2), 243–252.

21. Lindenauer, PK, SL Normand, EE Drye, Z Lin, K Goodrich, MM Desai, DW Bratzler, WJ O'Donnell, ML Metersky and HM Krumholz (2011). Development, validation, and results of a measure of 30-day readmission following hospitalization for pneumonia. *J Hosp Med*, 6(3), 142–150.
22. Jack, B, VK Chetty, D Anthony, JL Greenwald, GM Sanchez, AE Johnson, SR Forsythe, JK O'Donnell, MK Paasche-Orlow, C Manasseh, S Martin and L Culpepper (2009). A reengineered hospital discharge program to decrease rehospitalization: A randomized trial. *Ann Intern Med*, 150(3), 178–187.
23. Phillips, C, SM Wright, DE Kern, RM Singa, S Shepperd and HR Rubin (2004). Comprehensive discharge planning with post-discharge support for older patients with congestive heart failure: A meta-analysis. *JAMA*, 291(11), 1358–1367.
24. Crossen-Sills, J, I Toomey, and M Doherty (2006). Strategies to reduce unplanned hospitalizations of home healthcare patients: A step-by-step approach. *Home Healthc Nurse*, 24(6), 368–376.
25. Silverstein, MD, H Qin, SQ Mercer, J Fong and Z Haydar (2008). Risk factors for 30-day hospital readmission in patients ≥65 years of age. *Proc (Bayl Univ Med Cent)*, 21(4), 363–372.
26. Reed, RL, RA Pearlman and DM Buchner (1991). Risk factors for early unplanned hospital readmission in the elderly. *J Gen Intern Med*, 6(3), 223–228.
27. Corrigan, JM and JB Martin (1992). Identification of factors associated with hospital readmission and development of a predictive model. *Health Serv Res*, 27(1), 81–101.
28. Marcantonio ER, S McKean, M Goldfinger, S Kleefield, M Yurkofsky and TA Brennan (1999). Factors associated with unplanned hospital readmission among patients 65 years of age and older in a Medicare managed care plan. *Am J Med*, 107(1), 13–17.
29. Chu, LW and CK Pei (1999). Risk factors for early emergency hospital readmission in elderly medical patients. *Gerontology*, 45(4), 220–226.
30. Jasti, H, EM Mortensen, DS Obrosky, WN Kapoo, and MJ Fine (2008). Causes and risk factors for rehospitalization of patients hospitalized with community acquired pneumonia. *Clin Infect Dis*, 46(4), 550–556.
31. Smith, DM, A Giobbie-Hurder, M Weinberger, EZ Oddone, WG Henderson, DA Asch, CM Ashton, JR Feussner, P Ginier, JM Huey,

DM Hynes, L Loo and CE Mengel (2000). Predicting non-elective hospital readmissions: a multi-site study. Department of Veterans Affairs Cooperative Study Group on Primary Care and Readmissions. *J Clin Epidemiol*, 53(11), 1113–1118.

32. Runball-Smith, J, P Hider and P Graham (2009). The readmission rate as an indicator of the quality of elective surgical inpatient care for the elderly in New Zealand. *N Z Med J*, 122 (1289), 32–39.
33. Thakar, CV, PJ Parikh, and Y Liu (2012). Acute kidney injury (AKI) and risk of readmissions in patients with heart failure. *Am J Cardiol*, 109(10), 1482–1486.
34. Kansagra, D (2011). *Risk prediction models for hospital readmission: A systematic review*. Evidence-based Synthesis Program. Department of Veterans Affairs Health Services Research & Development Service. Washington DC.
35. Krumholz, HM, SL Normand, P Keenan, Z Lin, EE Drye, KR Bhat, Y Wang, J Ross, J Schuur, B Stauffer, S Bernheim, A Epstein, J Herrin, J Federer, J Mattera, Y Wang, G Mulvey and GC Schreiner (2008). Hospital 30-day heart failure readmission measure: Methodology. Report prepared for Centers for Medicare & Medicaid Services.
36. Krumholz, HM, SL Normand, PS Keenan, MM Desai, Z Lin, EE Drye, JP Curtis, KR Bhat and GC Schreiner (2008). Hospital 30-day acute myocardial infarction readmission measure: Methodology. A report prepared for the Centers for Medicare & Medicaid Services.
37. Krumholz, HM, SL Normand, PS Keenan, MM Desai, Z Lin, EE Drye, KR Bhat and GC Schreiner (2008). Hospital 30-day pneumonia readmission risk measure: Methodology. A report prepared for the Centers for Medicare & Medicaid Services.
38. Philbin, EF and TG DiSalvo (1999). Prediction of hospital readmission for heart failure: Development of a simple risk score based on administrative data. *J Am Coll Cardiol*, 33(6), 1560–1566.
39. Thomas, JW (1996). Does risk-adjusted readmission rate provide valid information on hospital quality? *Inquiry*, 33(3), 258–270.
40. Natale, J, S Wang and J Taylor (2013). A decision tree model for predicting heart failure patient readmissions. 10 pages. In *Proc. of the 2013 Industrial and Systems Engineering Research Conference*. A Krishnamurthy and WKV Chan (eds.), May 18–22. San Juan.
41. Lee, EW (2012). Selecting the best prediction model for readmission. *J Prev Med Public Health*, 45, 259–266.
42. Hosseinzadeh, A, M Izadi, A Verma, D Precup and D Buckeridge (2013). Assessing the predictability of hospital readmission using

machine learning. pp. 1532–1538. In *Proc. of the Twenty-Fifth Innovative Applications of Artificial Intelligence Conference*. H Munoz-Avila and D Stracuzzi (eds.), July 14–18. Bellevue, Palo Alto: AAAI Press.

43. Desai MM, Stauffer BD, Feringa H and GC Schreiner (2009). Statistical models and patient predictors of readmission for acute myocardial infarction a systematic review. *Circ Cardiovasc Qual Outcomes*, 2(5), 500–507.
44. van Walraven, C, IA Dhalla, C Bell, E Etchells, IG Stiell, K Zarnke, PC Austin and AJ Forster (2010). Derivation and validation of an index to predict early death or unplanned readmission after discharge from hospital to the community. *Can Med Assoc J*, 182(6), 551–557.
45. Breiman, L (2001). Random forests. *Mach Learn*, 45(1), 5–32.
46. Cortes, C and V Vapnik (1995). Support-vector networks. *Mach Learn*, 20(3), 273–297.
47. Breslow, NE, NE Day, KT Halvorsen, RL Prentice and C Sabai. Estimation of multiple relative risk functions in matched case-control studies. *Am J Epidemiol*, 108(4), 299–307.
48. Zhu, K, Z Lou, J Zhou, N Ballester, N Kong and PJ Parikh (2015). Predicting 30-day hospital readmissions with publicly available administrative database. *Method Inform Med*, 54(6), 560–567.
49. Rahman, MM and DN Davis (2010). Addressing the class imbalance problem in medical datasets. *Int J Mach Learn Comput*, 3(2), 224–228.
50. Wysocki, A, RL Kane, B Dowd, E Golberstein, T Lum and T Shippee (2014). Hospitalization of elderly Medicaid long-term care users who transition from nursing homes. *J Am Geriatr Soc*, 62(1), 71–78.
51. Crotty, M, CH Whitehead, R Wundke, LC Giles, D Ben-Tovim and PA Phillips (2005). Transitional care facility for elderly people in hospital awaiting a long term care bed: Randomised controlled trial. *BMJ*, 331(7525), 1110.
52. Renom-Guiteras, A, L Uhrenfeldt, G Meyer and E Mann (2014). Assessment tools for determining appropriateness of admission to acute care of persons transferred from long-term care facilities: A systematic review. *BMC Geriatr*, 14, 80.
53. Bohl, A, J Schurrer, W Lim and CV Irvin (2014). The changing medical and long-term care expenditures of people who transition from institutional care to home-and community-based services. https://www.medicaid.gov/medicaid/ltss/downloads/mfp-field-reports-2015.pdf. [Accessed 15 December 2016].

54. Arling, G, KA Abrahamson, V Cooke, RL Kane and T Lewis (2011). Facility and market factors affecting transitions from nursing home to community. *Med Care*, 49(9), 790–796.
55. Murtaugh, CM and A Litke (2002). Transitions through post-acute and long-term care settings: Patterns of use and outcomes for a national cohort of elders. *Med Care*, 40(3), 227–236.
56. Strang, VR, PM Koop, S Dupuis-Blanchard, M Nordstrom and B Thompson (2006). Family caregivers and transition to long-term care. *Clin Nurs Res*, 15(1), 27–45.
57. Naylor, MD, ET Kurtzman and MV Pauly (2009). Transitions of elders between long-term care and hospitals. *Policy, Polit Nurs Pract*, 10(3), 187–194.
58. Li, IC, SL Fann and HT Kuo (2011). Predictors of the utilization of long-term care (LTC) services among residents in community-based LTC facilities in Taiwan. *Arch Gerontol Geriat*, 53(3), 303–308.
59. Chhabra, PT, GB Rattinger, SK Dutcher, ME Hare, KL Parsons and IH Zuckerman (2012). Medication reconciliation during the transition to and from long-term care settings: A systematic review. *Res Social Adm Pharm*, 8(1), 60–75.
60. Robison, J, N Shugrue, M Porter, RH Fortinsky and LA Curry. Transition from home care to nursing home: Unmet needs in a home- and community-based program for older adults. *J Aging Soc Policy*, 24(3), 251–270.
61. Eika, M, GA Espnes, O Söderhamn and S Hvalvik (2014). Experiences faced by next of kin during their older family members' transition into long-term care in a Norwegian nursing home. *J Clin Nurs*, 23(15–16), 2186–2195.
62. Berwick, DM and AD Hackbarth (2012). Eliminating waste in US health care. *JAMA*, 307(14), 1513–1516.
63. Berwick, DM, TW Nolan and J Whittington (2008). The triple aim: Care, health, and cost. *Health Affair*, 27(3), 759–769.
64. Coleman, EQ and RA Berenson (2004). Lost in transition: Challenges and opportunities for improving the quality of transitional care. *Ann Intern Med*, 141(7), 533–536.
65. Murtaugh, CM and A Litke (2002). Transitions through post-acute and long-term care settings: Patterns of use and outcomes for a national cohort of elders. *Med Care*, 40(3), 227–236.
66. Murtaugh, CM, P Kemper and BC Spillman (1990). The risk of nursing home use in later life. *Med Care*, 28(10), 952–962.

67. Headen, AE Jr (1992). Time costs and informal social support as determinants of differences between black and white families in the provision of long-term care. *Inquiry*, 29(4), 440–450.
68. White-Means, SI (2000). Racial patterns in disabled elderly persons' use of medical services. *J Gerontol B Psychol Sci Soc Sci*, 55(2), S76–S89.
69. Borsch-Supan, A, V Hajivassiliou, LJ Kotlikoff and JN Morris (1992). Health, children, and elderly living arrangements: A multiperiod-multinomial probit model with unobserved heterogeneity and autocorrelated errors. In *Topics in the Economics of Aging, the National Bureau of Economic Research Book Series — the Economics of Aging*, DA Wise (ed.), pp. 79–108. Chicago: University of Chicago Press.
70. Reschovsky, JD (1998). The roles of Medicaid and economic factors in the demand for nursing home care. *Health Serv Res*, 33(4 Pt 1), 787.
71. Ng, T, C Harrington and M Kitchener (2010). Medicare and Medicaid in long-term care. *Health Affair*, 29(1), 22–28.
72. Reschovsky, JD (1996). Demand for and access to institutional long-term care: The role of Medicaid in nursing home markets. *Inquiry*, 33(1), 15–29.
73. Brown, JR, and A Finkelstein (2007). Supply or demand: Why is the market for long-term care insurance so small? *J Public Econ*, 91(10), 1967–1991.
74. Bauer, EJ (1996). Transitions from home to nursing home in a capitated long-term care program: The role of individual support systems. *Health Serv Res*, 31(3), 309.
75. Wittenberg, R, A Comas-Herrera, D King, J Malley, L Pickard and R Darton (2006). Future demand for long-term care, 2002 to 2041: Projections of demand for long-term care for older people in England. Tech Report. Kent: University of Kent. http://www.pssru.ac.uk/pdf/dp2330.pdf. [Accessed 15 December 2016].
76. Greene, VL and JI Ondrich (1990). Risk factors for nursing home admissions and exits: A discrete-time hazard function approach. *J Gerontol*, 45(6), S250–S258.
77. Rickayzen, BD and DE Walsh (2002). A multi-state model of disability for the United Kingdom: Implications for future need for long-term care for the elderly. *BAJ*, 8(2), 341–393.
78. Zuckerman, IH, P Langenberg, M Baumgarten, D Orwig, PJ Byrns, L Simoni-Wastila and J Magaziner (2006). Inappropriate drug use and risk of transition to nursing homes among community-dwelling older adults. *Med Care*, 44(8), 722.

79. Charles, KK and P Sevak (2005). Can family caregiving substitute for nursing home care? *J Health Econ*, 24(6), 1174–1190.
80. Aneshensel, CS, LI Pearlin, L Levy-Storms and RH Schuler (2000). The transition from home to nursing home mortality among people with dementia. *J Gerontol B Psychol Sci Soc Sci*, 55(3), S152–S162.
81. Chafe, R, P Coyte and NA Sears (2010). Improving the management of waiting lists for long term care. *Healthc Manage Forum*, 23(2), 58–62.
82. Meng, H, D Dobbs, S Wang and K Hyer (2013). Hospice use and public expenditures at the end of life in assisted living residents in a Florida Medicaid waiver program. *J Am Geriatr Soc*, 61(10), 1777–1781.
83. Gamerman, D and HF Lopes (2006). *Markov chain Monte Carlo: Stochastic simulation for Bayesian inference*, 2nd Ed. Boca Raton: Chapman & Hall/CRC.

7. A Multi-agent-based Simulation Model to Analyze Patients' Hospital Selection in Hierarchical Healthcare Systems

Jianpei Wen and Jie Song

Department of Industrial Engineering and Management, College of Engineering, Peking University, Beijing, China

Abstract

Hierarchical healthcare systems, which consist of general hospitals (GHs) and community healthcare centers (CHCs), have been gradually established in urban China to improve the accessibility of healthcare services. In this paper, a multi-agent simulation model is proposed to quantitatively analyze the impact of different factors on a patient's choice of healthcare facility. Results show that improving the quality and reducing CHC-related costs can encourage more patients to select CHC. Enhancing the quality of CHCs is an effective measure to relieve the congestion of GHs and

balance the load of GHs and CHCs. This information can help government decision-makers improve the patient flow distribution in urban China.

7.1. Introduction

With the increasing demand for healthcare services, along with insufficient public resources, governments have been prompted to explore various methods for medical resource allocation. For example, governments have established hierarchical healthcare systems, which consist of GHs and CHCs. In general, GHs provide better medical resources than CHCs do, but CHCs are more convenient for patients than GHs with regard to services. Such systems initially aim to implement an approach separating minor and severe diseases, i.e. minor diseases should be treated in community clinics, whereas severe diseases should be cured in general hospitals. CHCs, as primary care providers, can improve access to health services, enhancing familiarity with patients and reducing wasteful expenditures due to inappropriate specialist care; [1]. In CHCs, patients who suffer from minor diseases can be treated, thus, addressing long waiting in crowded GHs.

Medical resources of GHs should not be utilized for minor diseases, which can be effectively treated in CHCs, because such resources are allocated for the treatment of severe diseases. Access to health services in CHCs can be improved as minor diseases can be treated with reduced healthcare expenses and relatively shortened travel and waiting time.

Patient flow becomes unbalanced because patients overly rely on GHs, and they lack trust in the diagnostic ability of CHCs. According to the World Health Organization (WHO), 80% of the common diseases can be treated in CHCs. In China, however, patients with minor illnesses prefer to be treated in GHs. Consequently, GHs become congested and CHCs are underused. This study aims to investigate the effects of different factors on patients' behavior with regard to hospital choice and to encourage more patients with minor diseases to seek medical attention in CHCs by adjusting relevant factors with certain incentives.

In our study, hospital selection was defined as a process. In this process, patients select a nearby healthcare facility for healthcare services under the influence of static and dynamic factors. We focus on patients' hospital selection rather than on the hospitals where patients are eventually admitted. We consider the patients who are only suffered from common and frequently encountered diseases that can be treated in either CHCs or GHs.

Patients' hospital selection has been extensively investigated. Hospital management requires information regarding different factors influencing patients' hospital selection to encourage them to choose a particular hospital. Likewise, government policy-makers need such information to smoothen the patient flow distribution in healthcare delivery systems. Such factors have also been widely explored (e.g., [2] and [3]) and can be grouped into hospital attributes and patient characteristics. Hospital attributes mainly include medical services (e.g., quality of nurses and the availability of modern medical equipment), accessibility (e.g., waiting time), administrative services (hospital near residence), reputation, environment (e.g., hospital cleanliness), accessories (e.g., available parking area), and expenditures (e.g., outpatient cost). Patient characteristics generally include gender, education, income, marital status, occupation, and age. The effects of hospital attributes on a patient's preference for a certain hospital have been studied. Patient characteristics are incorporated by forming groups and estimating separate equations. Porell and Adams [4] conducted a survey in the previous study on hospital selection. Brown and Theoharides, [5] used a nested logit model to analyze the determinants of health-seeking behavior. Sivey [6] applied latent-class multinomial logit models to examine the influence of travel time and waiting time on the choice of hospitals for cataract operations. Different factors affecting the hospital choice behavior of patients were examined in these papers. However, the influence of the interactions among patients' choices has been disregarded. A group's behavior is distinct from the superposition of an individual's behavior. The non-cooperative relationship among patients when they seek healthcare services is also overlooked. We must consider not only the patients' preferences but also the interaction between the patients.

In this chapter, an agent-based simulation model is used to investigate the hospital selection process and the behavior of certain patient groups. In contrast to traditional simulation methodologies, such as discrete-event simulation, agent-based simulation focuses on modeling individuals, interactions between them, and interactions with physical or influential external factors. Agent-based models are applied to address the problems in healthcare operations management for several years in various focus areas, such as healthcare delivery, healthcare economics and policy, and epidemiology [7]. In our study, the agent-based model is used to consider the patients' individual preferences and the influence of other patients. The agent-based simulation model can support quantitative studies to determine the effects of different factors on resource allocation and hospital selection. These patients are regarded as agents with distinct sets of behaviors and characteristics, regardless of the effects of each type on the system. Their choice is typically a function of their preference and current system status. We have used utility function to model the patients' preferences, and we have assumed that each patient with a distinct preference selects a hospital that satisfies him or her in terms of treatment. Patients are myopic; as such, they select hospitals that completely satisfy them in terms of their preference and hospitals' current system states. In this process, they can continuously change their choices until they can no longer find a more suitable hospital relative to their previous choices.

The patients' decisions are guided by utility functions that describe their preferences based on different factors. Thus, determining a set of factors and using an appropriate structure for the utility function to describe their influences are the key parts of this study. The factors examined in this study are divided into static and dynamic factors. To incorporate these factors in the utility function, we have introduced a multi-attribute utility function. A multi-attribute utility function is a major analytical tool associated with the field of decision analysis [8]. It can explicitly identify the patients' trade-off among different factors. Patients select hospitals according to their utility function to maximize their benefits. We divided the

patients into several categories because they have different preferences. Each category was characterized in terms of utility function. The agent-based simulation aims to obtain the information on the patients' choices to determine the patient distribution in different healthcare facilities. We can also identify how different factors influence this distribution through the inputs of various levels of these factors.

In Section 2, a mathematical model is described to explain the objectives and constraints of the problem. In Section 3, the implementation of the agent-based simulation model and the sensitive analysis are discussed. In Section 4, results, conclusions, and future research problems are presented.

7.2. Model Description

A region composed of a set of streets $C = \{1, 2, ...\bar{C}\}$ is considered. A set of healthcare facilities (consisted of CHC and GH) $H = \{1, 2, ...\bar{H}\}$ and a set of patients $N = \{1, 2, ...\bar{N}\}$ are distributed in these streets. Before making their decisions, these patients obtain a set of information on healthcare facilities in set H. According to such information and their preferences, the patients in set N select a hospital in set H to get treated. We use the set $S = \{s_1, s_2, ...s_{\bar{N}}\}$ to express the result of patients' choices. The choice S_n of patient $n \in N$ can be any of the hospital in set H. This set of choices results S, which is designated as the outcome of the process. Each patient has a set of preference over these outcomes. We assume that each patient's preference over S can be represented by von Neumann-Morgenstern utility function [9], which means all patients make decisions that maximize their expected utility payoff. At the end of a certain choice, the patient $n \in N$ will obtain a utility $u_n(s_n, s_{-n})$, where $s_{-n} = \{s_1, s_2, s_{n-1}, s_{n+1}, s_{\bar{N}}\}$, which is the choice of all patients except patient n. In our model, the utility that each patient receives depends not only on the hospital selected by the patients according to their own preferences but also on the other patients' choices. Before constructing we the utility function, we will study the factors that influence patients' choice.

7.2.1. *Patients' preference*

We classify the factors into static factors and dynamic factors, as shown in Table 7.1. The static factors are determined before the patients select a hospital. These factors include the patients' attributes (patients' age, education, and income) and the attributes of healthcare facilities (distance, service capacity, price, and quality). The information about these factors is explicit in most cases. By contrast, the dynamic factors are not realized until the selection of a hospital by the patients. The dynamic factors include the waiting time and rejection probability. These factors dynamically change while the selection process by the patients is ongoing, and they are realized until all the patients select their hospitals. Patients cannot obtain the information on these factors in the hospital choice process. To make a better decision, the patients must conjecture this information according to public information and their assumptions on other patients' preference. To include all the three types of factors, we assumed the time when the patients select a hospital where they first obtain a certain utility, which is determined by static factors. The disutility increases as the patients start to select the same hospitals. This part of utility is designated as cost. We first constructed the utility determined by static factors. We assumed that every hospital $h \in H$ has available public information tuple $I_h = (P_h, Q_h, D_h)$, but that the patients obtain different utilities because of their own preferences.

Table 7.1. Attributes that impact the patients' choice.

Static factors	Facilities' attributes	Price (P) Quality (Q) Distance (D) Service capacity (c)
	Patients' attributes	Income, education, age, etc.
Dynamic factors	Factors that change along with the patients' choices, such as hospital's waiting time.	

However, a unique utility function for each patient is unnecessary and intractable. In this chapter, we have focused on how outpatient cost, service quality, and distance affect the decision of patients with different characteristics (income, education, age, and work) with regard to hospital choice. Three types of patients are considered in this research: *price-driven (PD)*, *quality-driven (QD),* and *distance-driven (DD)*. The patients' attributes were considered in this classification. The *PD* patients are mostly concerned about the outpatient cost. The *QD* patients are mostly concerned about the service quality. Similarly, the *DD* patients are the patients whose hospital choice decisions are mainly influenced by the distance between their homes and the hospital. The set of patient's type described above is denoted as Θ. We use $\theta_n \in \Theta$ to denote the type of patient n. We assumed that each patient is uncertain about the other patients' type. However, the probability distribution over patients' types $B(\theta_n)$ is identical and independent, and this information is known to all patients suffered from common disease in this region.

To consider all the static factors, we have constructed a multi-attribute utility function to describe the preferences of the patients. First, we have studied the relationship between these factors and constructed a single-attribute utility function for different factors. Then, we have used certain forms of theoretically valid multi-attribute utility functions to determine how the performance on each factor aspects the overall performance. We have constructed different utility functions for different types of patients.

In this chapter, we have used the exponential utility function to describe the patients' preference for each factor in the set of hospital attributes. The exponential utility function is flexible enough to model a wide variety of preferences. It not only captures the risk-seeking utility but also describes the risk-avoidance preference. Meanwhile, it is tractable to collect the data and estimate these parameters. We use x to denote the outcome of the game. Let $V(x)$ denote the utility associated with x. The exponential utility function can be determined by the following formula:

$$V(x) = \nu - \delta e^{\varpi x} \tag{1}$$

where ϖ is the patients' assessed risk tolerance and υ and δ are scaling constants for the factor.

We have assumed that each patient's preference is the same when only one of the hospital's attributes is considered, but different types of patients have different key factors affecting their choices when all of the hospital's attributes are considered. We have used the additive models to combine the three factors in the set of hospital attributes. The additive models are generally quite robust, and they typically provide a good approximation of the preferences that do not satisfy the additive independence [8]. We have used the weight of the factors in the additive model to discriminate the types of patients. Analytic hierarchy process (AHP) is used to obtain these sets of weight. $u(s_n,\theta_n)$ is used to denote the utility obtained when the patient who is classified as type θ_n selects a hospital S_n, which is determined by static factors. It is expressed as follows:

$$u(s_n,\theta_n) = w_{\theta_n,1}V(P) + w_{\theta_n,2}V(R) + w_{\theta_n,3}V(D), \tag{2}$$

$$w_{\theta_n,1} + w_{\theta_n,2} + w_{\theta_n,3} = 1. \tag{3}$$

- $V(P)$, $V(R)$, $V(D) \in [0,1]$ is the single-attribute utility function to outpatient cost, service quality, and distance, respectively.
- $w_{\theta_n,i}, i = 1,2,3$ is the weight of each factor, which is determined by the patients' type θ_n. The influence of a patient's attributes is indicated by this parameter.

Definition 1 *Patient's Type Characterization: As we have described above, it is intuitive that each type of patient can be defined according to the rank of the weight in each factor:*

1. *PD patients:* $w_{\theta,1} \geq w_{\theta,2}, w_{\theta,1} \geq w_{\theta,3}$
2. *QD patients:* $w_{\theta,2} \geq w_{\theta,1}, w_{\theta,2} \geq w_{\theta,3}$
3. *DD patients:* $w_{\theta,3} \geq w_{\theta,1}, w_{\theta,3} \geq w_{\theta,2}$

Definition 1 figures out the characteristic of each type of patients analytically. In the Definition 1, the rank of the weights shows the most important factor for each type of patients when they select a hospital. The *PD* patients are mostly concerned about the outpatient

cost, and thus, the influence of the outpatient cost is larger than the distance and service quality. The *QD* patients are mostly concerned about the service quality, such that when they are treated in a hospital with better service quality, they are likely to select this hospital regardless of the outpatient cost and distance. Similarly, the *DD* patients are mostly concerned about the distance.

In the above paragraph, we have determined how to describe the patients' preferences. However, more patient preference data is required to estimate the parameters in the patients' utility function. The large data revolution in healthcare is now under way. The sources of the large data in healthcare contain activity (claims) and cost data, clinical data, pharmaceutical research and development data, and patient behavior and sentiment data [7]. We can obtain information on the diseases that can be treated in CHCs through the clinical data. The activity (claims) and the cost data contain the outpatient cost information. Furthermore, patient behavior and sentiment data facilitate the classification of patients and the accurate estimation of the parameters in the patients' utility function.

To estimate the parameter of the patients' utility function on each factor, we have analyzed the data in aspect of each factor. We have obtained the minimum and the maximum value of each factor. We have also determined the point of each factor that represents the probability of the patients' selection of hospital, with the value of this factor being less than or equal to 0.5. With these three points, we can determine the patients' utility function on each factor.

In addition, considering the patient behavior and sentiment data, we can obtain a better classification of patients through cluster analysis, using the attributes of the patient. After the classification, we can use the data to estimate the patients' weight of each factor in the additive multi-attribute utility function by Bayesian preference elicitation method [10].

7.2.2. *Patients' decision model*

Intuitively, too many patients selecting the same hospital deteriorates the accessibility of the healthcare service, which leads to utility loss for the patients. Hence, we have assumed that if the number of

patients selecting the hospital N_{s_n} surpass this hospital's service capability, these patients get a cost A. The cost is rising as the increasing number of patients are selecting the same hospital. Then, patients finally obtain the net utility u_n^{net}, which is the difference between the utility $u(s_n, \theta_n)$ and the cost A, as shown by the following equation,

$$u_n^{net} = u(s_n, \theta_n) - A(s_n, s_{-n}) \tag{4}$$

We have assumed that the capacity c_h of hospitals $h \in H$ is known to all patients. We defined cost A as a function of the load L_{s_n} of the hospital selected by patient $n \in N$. We defined the load as follows,

$$L_{s_n} = N_{s_n} / c_{s_n}$$

We then formulated the cost as follows,

$$A(s_n, s_{-n}) = \max(0, 1 - 1 / L_{s_n}) u_n(s_n, \theta_n)$$

The total utility that a certain hospital can control is c_{s_n}. When the number of patients that selected this hospital is less than c_{s_n}, they obtain utility $u(s_n, \theta_n)$, which is determined by their preferences and choices. When the number of patients selecting this hospital is more than c_{s_n}, the patients must pay a cost to make up for the utility shared by the excess patients. To obtain better utility, every patient, then, not only selects a better hospital but also considers the number of patients who have the same preferred hospital.

With the identification of the choice set and the net utility of each patient, we have modeled each patient as an individual agent, belonging to one of the three agent types. Furthermore, we have found that each agent changes their decisions to maximize their net utility according to the current system status. Thus, we can derive the steady-state distribution of number of the patients accepted by each hospital.

7.3. Case Study

In the above section, we have defined the agents and their behavior. In this section, we have set up an agent-based simulation model according to the data collected from the current policy in real-time using commercial software, and validated the model. Lastly, we have provided a comprehensive analysis based on the model to obtain insight.

7.3.1. *Input parameter*

We selected a region in *Beijing*, *China*, which consisted of 14 hospitals and seven streets, as well as five GHs and nine CHCs. We collected the data on hospitals' attributes from the "*Annual Chinese health statistics report*" and from the official statistics report of each hospital. The data included the service capacity, mean outpatient cost (RMB), and service quality. We calculated the number of patients in each street using the street's population and the *Two-week Hospital Visit Rate*, which was obtained from "*Sixth Population Census Of Beijing*" and "*China Social Statistical Yearbook 2013*". A total of 3,992 patients competed to choose their preferred hospitals, while the total service capacity was 3,527. Based on the geographical location, the distance between a street and a hospital was collected using *GoogleMaps*. The attributes of each hospital are listed in Table 7.2.

To obtain the exponential utility function defined in Eq. (1), we surveyed a group of patients about three parameters: the least preferred outcome a, the most preferred outcome b, and the median preferred outcome x_0 (the patient will get utility 0.5 with this outcome), corresponding to three different factors. We then used the data to calculate the parameter Ψ, defined in Eq. (5). Finally, the exponential utility function was determined with Eq. (6). The results are shown in Table 7.3.

$$V(x) = \frac{1 - \exp(-(x - a) / \Psi)}{1 - \exp(-(b - a) / \Psi)} \tag{5}$$

Table 7.2. The attributes of hospitals included in the case study.

Hospital	Outpatient Cost (RMB)	Service Capacity	Quality
GH1	242.1	952	9
GH2	242.1	162	9
GH3	242.1	480	9
GH4	157.4	110	6
GH5	157.4	200	6
CHC1	84.6	168	1
CHC2	84.6	182	6
CHC3	84.6	71	1
CHC4	84.6	214	3
CHC5	84.6	74	3
CHC6	84.6	253	6
CHC7	84.6	114	6
CHC8	84.6	267	4
CHC9	84.6	280	3

Table 7.3. Single attribute utility function.

Attribute	(b,1)	(a,0)	$(x_0, u(x_0))$	Utility Function
Price	(70,1)	(260,0)	(140,0.5)	$0.1237\,(\exp(1.2956 - 0.005p) - 1$
Quality	(9,1)	(1,0)	(6,0.5)	$0.5431\,(\exp(0.1305q - 0.1305) - 1$
Distance	(0.18,1)	(11.5,0)	(3.2,0.5)	$0.1237\,(\exp(2.2419 - 0.1949d) - 1$

$$\nu = \frac{1}{1 - \exp(-(b-a)/\Psi)}$$
$$\delta = \frac{\exp(a/\Psi)}{1 - \exp(-(b-a)/\Psi)} \tag{6}$$
$$\varpi = -\frac{1}{\Psi}$$

According to the patient's type characterization and the data on the patient attributes we collected, we divided the patients into three

Table 7.4. Multi-attribute utility function.

Types	The Weight of Factors (w_P, w_Q, w_D)	The Proportion of Different Type of Patients
PD	(0.85, 0.14, 0.01)	0.3
QD	(0.1, 0.8, 0.1)	0.5
DD	(0.1, 0.2, 0.7)	0.2

groups: *PD* patients, *QD* patients, and *DD* patients. Because we defined the different multi-attributes utility function for different types of patients, we surveyed three groups of patients regarding the weight of the three factors with AHP. Meanwhile, we obtained the proportion of patient types. The results are shown in Table 7.4.

7.3.2. *Simulation analysis*

Commercial simulation software AnyLogic was used for the study. We set up an agent-based simulation model to represent the decentralized decision-making process of the patients. Figure 7.1 provides the interface of our simulation. The simulation logic can be summarized as follows:

1. Initialization: At the beginning of the simulation, the model randomly assigned the patients' type according to the distribution of the patients' type. The model subsequently distributed the patients with different types to each street.
2. Decision-making: Every minute, the patients selected the most satisfactory hospital according to their utility function and the real-time information of hospital. Every 50 minutes, we recorded the distribution of the number of patients accepted by each hospital as a sample.
3. Output: We used this model to obtain the steady-state distribution of the number of patients accepted by each hospital.

The length of the simulation is 90,700 minutes (model time), and we obtained 1,814 samples. The "warm-up" period of 700 minutes was observed to be sufficient to reduce the issues of the initial

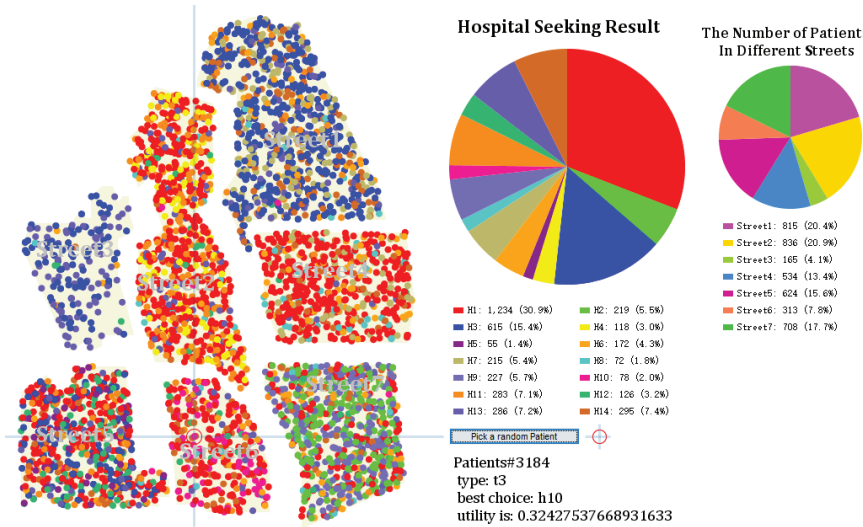


Figure 7.1. The simulation environment.

conditions. We used the batch mean method [11] to obtain the estimation of the number of the patients who selected each healthcare facility. We, initially, grouped these samples into batches of 400 and calculated the corresponding batch means. We, then, calculated the lag-1 auto-correlation of the batch means. A lag-1 auto-correlation of the batch mean was close to 0, indicating that the batch means were nearly independent. This result appeared mainly because of the 50-minute sampling we conducted. We, then, regrouped these samples into batches with 40 individuals each, and the new corresponding batch means were computed. We used the later batch means to estimate their mean. Table 7.5 summarizes the results of the simulation analysis.

Here, we defined an overcrowded ratio OR_h, $h \in H$ as a measure of the congestion of the hospital h as follows: $OR_h = (N_h - SC_h)/SC_h$, where N_h is the number of patients selecting hospital M, and SC_h is the service capacity of hospital h.

First, we analyzed the selected result of each type of the patients. The result showed that 94.84% of the *PD* patients selected CHC, and 69.36% of the *PD* patients selected the CHC with the best

Table 7.5. Summary of results.

Hospital	The Selection Result of Each Type of Patients			Overcrowded Ratio
	PD	*QD*	*DD*	
GH1	0.03908	0.45691	0.18136	0.15924
GH2	0.00476	0.17485	0.04008	1.38704
GH3	0	0.27079	0.20741	0.47104
GH4	0.00526	0.00927	0.06112	–0.33091
GH5	0.00251	0.00902	0.06237	–0.64600
CHC1	0.02680	0	0.05962	–0.52560
CHC2	0.33742	0.01954	0.01879	1.51703
CHC3	0.01754	0	0.00301	–0.67042
CHC4	0.07690	0	0.11698	–0.13318
CHC5	0.04484	0	0.04684	0.23108
CHC6	0.21969	0.04785	0.12625	0.81581
CHC7	0.13652	0.01177	0.07565	1.17018
CHC8	0.04584	0	0	–0.79438
CHC9	0.04284	0	0.00050	–0.81536

service quality (CHC2, CHC6, CHC7, their service quality is "6", which is the best service quality among the CHCs). The *PD* patients, thus, had preference for CHCs, especially when the CHCs exhibited high service quality. Some 92.08% of the *QD* patients selected GHs. This result indicated that the *QD* patients preferred GH. AS much as 79.06% of the *DD* patients selected the top three nearest hospitals, and 55.24% of them selected GHs. The *DD* patients, thus, preferred the nearest hospital. More patients selected GHs because CHCs were not significantly closer to the patients, relative to the GHs (the average distance between the patients and the GHs was 5.287 *km*, and that between the patients and CHCs was 5.157 *km*, and we considered the center of the street as the origin point of all the patients in that street when we collected the data for distance), and that the service quality, which was the second most important

factor in *DD* patients, significantly affected the result. We, then, analyzed the overcrowded ratio of each hospital. The result showed that congestion mainly occurred in GHs and CHCs with high service quality, such as GH1, GH2, GH3, CHC2, CHC5, CHC6, and CHC7. This finding appeared because of the service quality being the second most important factor in *PD* and *DD* patients. The results are consistent with our assumption on the characteristic of each patient type, thus validating our model.

To investigate the impact of the proportion listed in Table 7.4 on the results, we compared the result shown in Table 7.5 with the result when the proportion of different type of patients was "PD:0.4, QD:0.4, DD:0.2". The result is displayed in Fig. 7.2. From this figure, we find that hospitals with positive (negative) overcrowded ratio when the proportion is "PD:0.3, QD:0.5, DD:0.2" remain positive (negative) overcrowded ratio when we change the proportion to "PD:0.4, QD:0.4, DD:0.2". It implies that even if we slightly adjust the proportion, the fact that the hospital with relative higher quality is crowded, while the hospital with lower quality is idle is unchanged. In Fig. 7.2, when we increase the proportion of the PD patients and decrease the proportion of QD patients simultaneously, the overcrowded ratio of CHC2 and CHC7 increase significantly, while, the overcrowded ratio of GH3 decrease significantly. It is

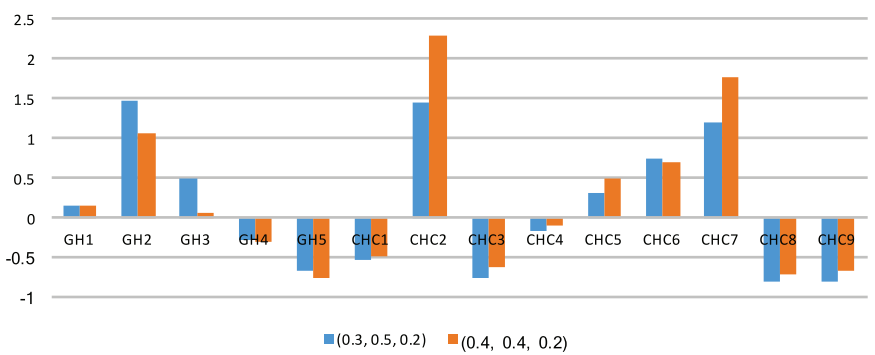


Figure 7.2. The change of overcrowded ratio for each hospital with the change in the proportion of different type of patients.

mainly because more PD patients change their choice from GH to CHC with relatively higher quality.

7.3.3. *Two incentive policies*

As we addressed in this chapter, the patients' decisions was not determined by a single factor. As indicated in the above case study, the outpatient cost and service quality acted as the most important factors affecting the patients' hospital choice behavior. This section presents the the continued case study to improve the system by changing the outpatient cost and the lowest service quality of CHCs.

7.3.3.1. *Reducing the outpatient cost of CHCs*

Price is an important factor that affects the patients' hospital choice behavior. The government frequently introduces policies to regulate the patients' hospital choice behavior by adjusting the price. However, the result is frequently distinct from the situation due to individual or single factors. In this section, we have provided an analysis on the prices of CHCs and have also studied how the price affects the patients' hospital choice. We have varied the CHCs' price from RMB 75 to RMB 125 by five a step and studied the change pattern on the percentage of patients selecting a particular hospital. The results are shown in Fig. 7.3. The change in the price did not significantly change the number of patients who selected GH1-GH4 and CHC2-CHC9. However, an increase in the price significantly increased the number of people who selected GH5. In CHC1, this performance decreased significantly. This finding was mainly because 65.4% of the total capacity of GH5 was not used when the price was \$84.6. When the price of the low-quality CHC increased, a large number of patients who initially selected CHCs eventually preferred GH5.

However, this result was counter-intuitive. We expected that part of patients would change their choice from GHs to CHC due to an reduction in the price of CHCs. When we only reduced the CHCs' price, we might encourage some patients to select CHCs, resulting in

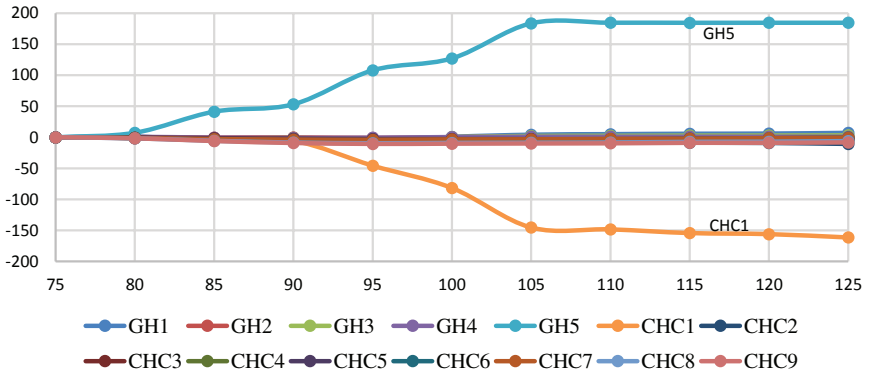


Figure 7.3. The change in the number of patients due to an increase in the CHCs' price.

the GHs with low quality getting low utilization, while the hospitals with the best quality remained crowded.

7.3.3.2. *Improving the lowest quality of the CHC*

As solely adjusting the CHCs' price cannot encourage more patients to select CHCs and balance the utilization of CHCs and GHs simultaneously, we attempted to improve the lowest quality of CHCs. We varied the lowest quality of CHCs from 1 to 3 by one a step and studied the change pattern of the number of people selecting a particular hospital. The results are shown in Fig. 7.4.

Since the change of the lowest quality was not significantly different, this change did not cause a significant difference in the overall layout of the entire urban healthcare service system. However, we could still get the trend of this change. The result showed that this measure did not change the percentage of people who selected CHC2-CHC9 and GH4. However, it relieved the congestion of GH1-GH3 and improved the utilization of GH5 and CHC1 and CHC3. The reason was that CHC1 and CHC2, which had the lowest quality, were located around the GH1-GH3. Thus, CHC1 and CHC2 attracted some patient whose original strategy was to select GH1-GH3. When CHC1 and CHC2 were congested, some patients

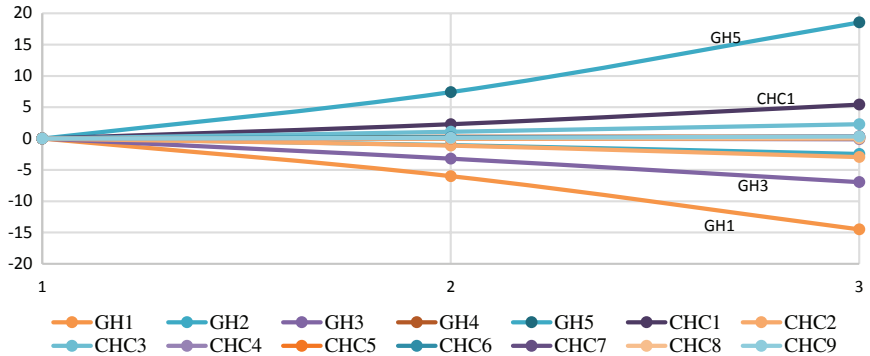


Figure 7.4. Change in the number of patients who selected each hospital along with the improvement of the lowest quality.

selected GH5, which had some idle resource. These results indicate that minimal improvement in the CHCs' lowest quality, without changing the overall layout of the entire urban healthcare service system, can relieve the congestion of GH1-GH3 and can balance the load of GH and CHC.

7.4. Conclusion

In this study, an agent-based simulation model was proposed to analyze the obtained data and examine the effects of different factors on resource allocation in hospital selection. We considered different preferences of various patients and the relevant dynamic factors. Using some public data and referring to the current situation in the urban healthcare service system in Beijing, China, we validated our simulation model. We, then, performed a sensitive analysis on the expenditures and quality of CHCs to obtain interesting insights.

The improvement of the quality of CHCs and the reduction of CHC expenditures can encourage more patients to select CHCs. However, reducing the CHCs expenditures influences patients who prefer GHs with the lowest rank. Hospitals with the highest rank remain crowded. Furthermore, GHs with low rank are poorly utilized. The improvement of the low-quality CHCs can help decongest

GHs and balance the load of GHs and CHCs. Therefore, alleviating the gap of diagnosis and treatment quality between CHCs and GHs is the key to solving this problem.

Considering the limited data, we classified the patients into three groups on the basis of the preference for GHs or CHCs. The precise classification of patients on the basis of large data can provide further insights into hospital selection. The proposed agent-based approach can be applied not only to hospital selection but also to other processes regarding decentralized or localized decisions that address resource allocation problems.

Acknowledgement

This research was partially supported by the National Science Foundation of China (Grant No. 71301003) and the Specialized Research Fund for the Doctoral Program of Higher Education (Grant No. 20130001120007).

References

1. Balasubramanian, H, A Muriel, A Ozen, L Wang, X Gao and J Hippchen (2013). Capacity allocation and flexibility in primary care. In *Handbook of Healthcare Operations Management*, BT Denton (Ed.), pp. 205–228. New York: Springer.
2. Draper, M, P Cohen and H Buchan (2001). Seeking consumer views: what use are results of hospital patient satisfaction surveys? *Int J Qual Health Care*, 13(6), 463–468.
3. Cooper-Patrick, L, NR Powe, MW Jenckes, JJ Gonzales, DM Levine and DE Ford (1997). Identification of patient attitudes and preferences regarding treatment of depression. *J Gen Intern Med*, 12(7), 431–438.
4. Porell, FW and EK Adams (1995). Hospital choice models: A review and assessment of their utility for policy impact analysis. *Med Care Res Rev*, 52(2), 158–195.
5. Brown, PH and C Theoharides (2009). Health-seeking behavior and hospital choice in china's new cooperative medical system. *Health Econ*, 18(S2), S47–S64.

6. Sivey, P (2012). The effect of waiting time and distance on hospital choice for English cataract patients. *Health Econ*, 21(4), 444–456.
7. Barnes, S, B Golden and S Price (2013). Applications of agent-based modeling and simulation to healthcare operations management. In *Handbook of Healthcare Operations Management*, BT Denton (Ed.), pp. 45–74. New York: Springer.
8. Butler, J, DJ Morrice and PW Mullarkey (2001). A multiple attribute utility theory approach to ranking and selection. *Manage Sci*, 47(6), 800–816.
9. Roger, BM (1991). *Game Theory: Analysis of Conflict*. Cambridge: Harvard university press.
10. Guo, S and S Sanner (2010). Real-time multiattribute Bayesian preference elicitation with pairwise comparison queries. In *International Conference on Artificial Intelligence and Statistics*, YW Teh and M Titterington (Eds.), pp. 289–296. Sardinia, Italy.
11. Kelton, WD and AM Law (2000). *Simulation Modeling and Analysis*. Boston: McGraw Hill.

8. Forecasting Recipient Outcomes of Deceased Donor Livers

Rachel M. Townsley*, Maria E. Mayorga*, A. Sidney Barritt IV† and Eric Orman‡

*Department of Industrial and Systems Engineering, North Carolina State University, Raleigh, NC, USA
†Department of Medicine, Indiana University School of Medicine, Indianapolis, IN, USA
‡Department of Medicine, University of North Carolina, Chapel Hill, NC, USA

Abstract

Liver transplantation has been the standard treatment of end stage liver disease for over three decades. While demand for liver transplants has increased over the years, the number of transplants performed have decreased or stagnated over the last decade. Health trends in the general population could play a role in the growing gap between the supply and demand of livers for transplantation. Obesity, diabetes, and an aging population are the cause of declining donor liver quality as well as the cause of growing transplant waitlists. We use United Network for Organ Sharing (UNOS) data to develop statistical and simulation models to evaluate post-transplant outcomes of liver allocation in the USA in light

of these trends. In particular, we predict the characteristics of the donor population, create a population dynamics model of the recipient population, match donors and recipients, and predict survival outcomes after transplantation based on an existing survival analysis model. We find that despite dynamic trends in both donor and recipient populations, overall survival outcomes will remain stable over the next ten years. However, the trend is not the same for all diseases groups, with some experiencing an increased risk and others a decreased risk, adding to disparities in outcomes between disease groups.

8.1. Introduction

Approximately 35 million Americans are impacted by varying levels of liver and biliary diseases. Hepatitis, cirrhosis, and liver cancer are among the most serious causes of liver disease, which is the 12th leading cause of death in USA. It is estimated that chronic liver disease and cirrhosis are the cause of 36,000 deaths and 100,000 hospitalizations annually [1, 2].

Since the 1980s, liver transplantation has been the standard treatment for end-stage liver disease (ESLD). While living donation is possible, the majority of liver transplants performed in USA use the liver of a deceased donor. The waitlists and allocation processes are managed by the United Network for Organ Sharing (UNOS) and are prioritized by the Model for End-Stage Liver Disease (MELD) score, a measure of disease severity based on readily available patient data. Currently, there are more than 15,000 individuals on the liver transplant waitlist. In 2015, approximately 6,000 liver transplants were performed, and an estimated 1,400 individuals died waiting for a transplant [3].

Although the number of donor livers available for transplantation is growing with the population, the utilization of these livers has been decreasing and is projected to continue to decline during the next decades. The utilization rates are declining largely due to changing demographics and health trends in the donor population, specifically, the increasing rates of obesity and diabetes, an aging population, as well as increased rates of organ donation after cardiac

death (contrasted with brain death). In addition to liver utilization rates, these health trends also significantly impact the demographics of transplant recipients.

In this study, we investigate the trends at play in liver transplantation and evaluate the effects of these trends on the transplant recipient population and resulting predictors of survival, namely the survival probability and the D-MELD score.

8.2. Existing Work and Motivation

In recent years, liver transplantation has been a rich area of research in the field of medical decision-making. In the following literature summary, we provide an overview of studies that investigate and propose varying strategies to improve or optimize outcomes from both an overall systems perspective and an individual patient perspective. Our analysis builds on the existing studies about donor liver availability and utilization, which we have discussed in more detail.

8.2.1. *Liver transplantation*

Several studies approach the problem from a systems level perspective, such as the design of allocation policies [4, 5] and redistricting of regions [6–9]. These papers focus on improving outcomes via improved resource allocation. The problem is complex due to the trade-off between equity and efficiency, which is inherent in the distribution of such a tightly constrained, scarce, and highly valued life-saving resource. Donor liver availability is stochastic both in timing and geographical distribution. This stochastic nature of the availability, coupled with the deterioration of organ quality associated with travel distances, makes districting and allocation rules critical to efficiency and equity in outcomes.

Other studies related to liver transplantation approach the problem from an individual patient perspective. These studies focus on several decisions of the patient or the surgeon, such as the decision to accept or reject a cadaveric liver that becomes available [10, 11], the decision of timing for living donor liver transplants [12], or the decision of choosing among cadaveric and living donor livers [13].

8.2.2. *Donor liver availability and utilization*

Additionally, recently published studies, [14–16], forecast a decline in donated liver utilization due to population health and demographic changes.

Parikh *et al.* analyze Organ Procurement and Transplantation Network (OPTN) data from 2000 to 2012, calculating the number of donor livers over time in subgroups stratified by age, race, and body mass index (BMI). They then apply general population demographic and health trends predicted from census data and national nutritional surveys to make projections of donor liver availability and utilization, concluding that population growth will outstrip donor population growth in the coming decade.

Similarly, Toro-Diaz *et al.* and Orman *et al.* seek to forecast donor liver availability in the coming years but utilize a different approach. Comparing health and demographic measures, they find that donor trends are not well represented by general population trends seen in census data and nationwide health surveys like the National Health and Nutrition Examination Survey (NHANES) and the Behavioral Risk Factor Surveillance System (BRFSS).

Using historical clinical UNOS data, Toro-Diaz *et al.* develop multiple statistical models to predict relevant donor characteristics, including gender, age group, race, BMI (obese or not), alcohol use, diabetes status, cause of death (stroke or not), mechanism of death (cardiac or brain), and biological marker values (bilirubin and alanine transaminase (ALT) levels). Based on these donor characteristics, they use logistic regression to compute the probability of liver utilization. These statistical models are implemented in a discrete event simulation model, which generates the donor characteristics and viability of each liver in a donor population. They conclude that cadaveric liver utilization will decline significantly in the coming years.

8.2.3. *Transplant recipient outcomes*

Numerous studies in the clinical literature have attempted to quantify or characterize survival outcomes based on the characteristics of transplant liver donors and recipients [17–19].

Ioannou develops and validates a proportional hazards survival model based on the characteristics of a donor-recipient pair [17]. The model is based on UNOS data from 1993 to 2003 and predicts post transplant survival based on three donor characteristics (age, gender, and race/ethnicity), transplant recipient characteristics (age, BMI, MELD score, UNOS priority status, gender, race/ethnicity, diabetes status, cause of liver disease, and serum albumin level), as well as cold ischemia time (CIT), the amount of time the donor organ is chilled between harvest and transplant. Ioannou develops a model specific to recipients with Hepatitis C and another model that predicts survival for all other recipients. In general, livers from female, minority, and older donors are associated with increased post transplant risk. Recipients who were older than 50, male, or African American were also associated with increased post transplant risks.

Rana *et al.* develop the Survival Outcomes Following Liver Transplantation (SOFT) score with the goal of balancing the waitlist mortality captured by MELD score with predicted 3-month post transplant outcomes based on donor, recipient, and operational characteristics [18]. The analysis uses data from 2002 to 2006 to create a logistic regression model, predicting recipient survival for 3 months after transplant. The authors use the odds ratios from the regression model to develop a point system to use as a score measure of survival risk for the recipient-donor pair.

Halldorson *et al.* propose that the D-MELD, the product of the donors' age and the recipients' MELD score, provides a simple but effective predictor of post-transplant mortality [19]. The authors conclude that D-MELD scores greater than 1,600 predict poor post transplant outcomes and recommend this as a cut-off in liver allocation decisions.

8.2.4. *Methodology overview*

In this study, we predict characteristics of donors by replicating and building on the methodologies used in Toro-Diaz, *et al.* [14], develop methods to predict recipient characteristics, and evaluate and forecast survival outcomes at the macro scale by implementing these

with Ioannou's survival model [17] in a discrete event simulation. That is, we have modified an existing population dynamics model to predict the donor population. Then, we create a population dynamics model of the recipient population. Lastly, donors and recipients are matched, and an existing survival analysis model is used to predict survival outcomes after transplantation.

8.3. Statistical Models

The statistical models developed to predict donor and recipient characteristics are based on 2004–2014 data from the UNOS Standard Transplant Analysis and Research (STAR) file provided by the Organ Procurement and Transplantation Network (OPTN) which contains de-identified patient level data. The statistical analysis was conducted using SAS 9.4.

8.3.1. *Recipient characteristics*

According to the UNOS STAR data, 68,822 deceased liver transplants were performed during 2004–2014. From these, we excluded Status 1 patients with acute liver failure, split liver transplants, pediatric patients, and adult patients who received very young pediatric livers (less than 10 years old). These cases were excluded as their characteristics were very distinct from the characteristics of the general recipient population, leaving 56,296 transplants. Additional exclusions were made due to incomplete records, resulting in 55,489 data points, which were used to create statistical models of recipient characteristics.

Significant recipient characteristics identified in Ioannou's survival model are age, race, gender, BMI, diabetes status, disease type, albumin levels, and laboratory MELD score. Table 8.1 provides an overview of the statistical methods used to predict each variable, as well as the independent variables that were used to predict each dependent variable. The input variables for each response variable were based on both clinical significance and statistical significance, with an effort to minimize the number to variables necessary to predict each response. Historical data demonstrated that the proportion of males and females in the recipient population was fairly consistent

Table 8.1. Summary of statistical methods and variables used for modeling recipient characteristics.

Dependent Variable	Independent Variable	Statistical Model
Gender	—	Constant Discrete Distribution
Race Group	Gender, Year	Linear Regression
Age Group	Gender, Year	Multinomial Logistic Regression
BMI Category	Gender, Race, Year	Multinomial Logistic Regression
Diabetes	Gender, Race, BMI	Logistic Regression
Disease	Gender, Race, Age, BMI, Diabetes, Year	Multinomial Logistic Regression
Albumin Level	Age, Disease, BMI	Ordered Logistic Regression
MELD	Disease, Year	Empirical Distribution with Linear Parameter Trends

over time. Therefore, an empirical constant discrete distribution was used (32.3% female and 67.7% male).

Race categories were defined as non-Hispanic white, non-Hispanic black, Hispanic, and other. Predictions of race were stratified by gender, and the proportion of the population falling into each race group was based on a linear regression of historical trends by year. Age groups were defined as 18–42, 43–49, 50–56, 57–63, and over 63. A multinomial logistic regression was used to predict age group, based on gender and year.

BMI categories were defined as 18–25, 25–30, 30–35, 35–40, and 40–55. This was modeled using a multinomial logistic regression, with independent variables gender, race, and year. Diabetes status was a binary recipient attribute, which was modeled using logistic regression based on the recipient's gender, race, and BMI.

Disease types were categorized into seven groups: alcoholic cirrhosis, cryptogenic cirrhosis, primary biliary cirrhosis, Hepatitis B, Hepatitis C, hepatocellular carcinoma, and "other". Independent variables for predicting recipient disease type were gender, race, age, BMI, diabetes status, and year.

Albumins are blood plasma proteins produced in the liver, and levels are grouped into five categories: less than 2.1, 2.1–2.4,

2.4–2.8, 2.8–3.3, and greater than 3.3. Recipient albumin levels were predicted based on an ordered logistic regression model, using age, disease type, and BMI as independent variables.

MELD score is a measure of expected patient survival without transplant and is used to determine waitlist priority. The "laboratory MELD" was is calculated based on creatinine, bilirrubin, and international normalized ratio (INR) values as shown in equation (1). Exception points were, then, added to account for specific circumstances (disease type and other health measures), and this summed score determined the order of the waitlist. The value of interest here was the computed laboratory MELD score, not the waitlist MELD score.

$$\begin{aligned} LabMELD = {} & (0.957) * \ln(creatinine) + \\ & (0.378) * \ln(bilirrubin) + \\ & (1.12) * \ln(INR) \end{aligned} \tag{1}$$

According to experts disease type and year are the most important predictors of the recipient MELD score. In order to incorporate these independent variables, we stratified the MELD score data by disease type and year, and performed a best fit analysis of the resulting distributions using the SAS Univariate procedure. Figure 8.1 shows the distribution of MELD scores for alcoholic cirrhosis transplant recipients in 2004.

The Weibull distribution was chosen as providing a good fit for the strata. For each disease, a linear annual trend was fitted to the each of the Weibull parameters. Using this approach, the distribution of MELD scores for each disease type can be predicted for future years. The Weibull distribution parameters for MELD scores of alcoholic cirrhosis recipients are shown in Fig. 8.2, along with the linear trend fitted to each parameter.

8.3.2. *Donor characteristics*

Donor characteristics were modeled using the same methodologies described in Toro-Diaz *et al.* [14] while utilizing the most recent

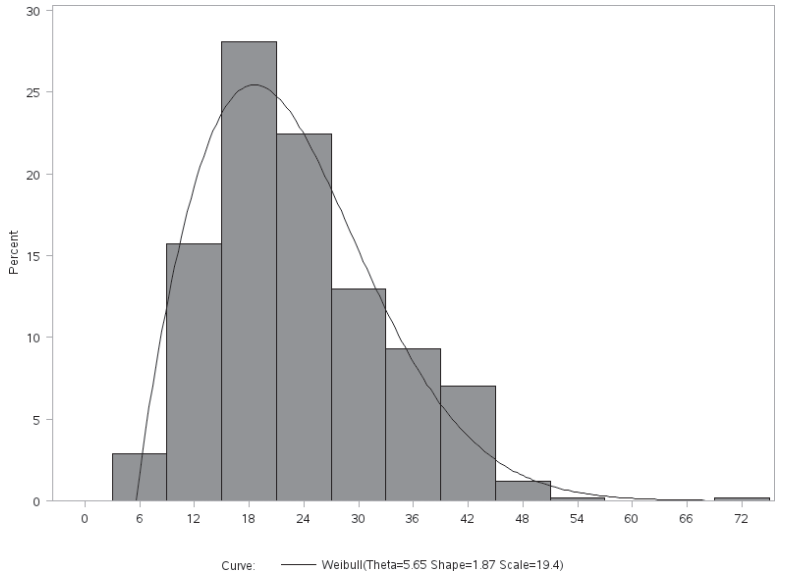


Figure 8.1. Histogram and distribution fit of MELD scores for alcoholic cirrhosis transplant recipients in 2004.

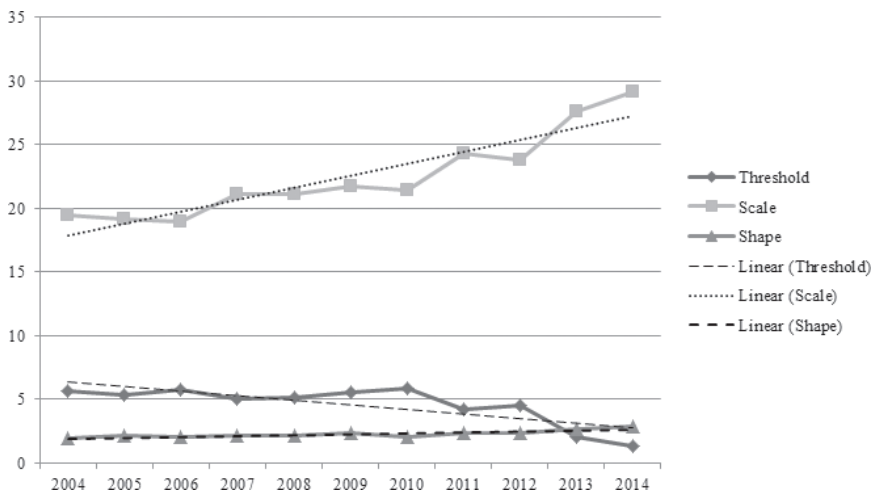


Figure 8.2. Weibull parameters for MELD scores of alcoholic cirrhosis recipients and predictive linear trends used to model these parameters over time.

STAR data available (2004–2014). The analysis of the historical transplant data shows that approximately 10% of adult transplant recipients receive pediatric livers. On the basis of this, we created a set of new models to characterize pediatric donor livers, which were allocated to adult recipients using similar methodologies. We used step-wise parameter selection for all pediatric statistical models. For some statistical models of pediatric donor liver characteristics, fewer independent variables were significant when compared to adult donor liver models (e.g., gender was not a significant predictor of obesity for pediatric donors while it was significant for adult donors). Additionally, we found that stratifying the data by gender yielded a better-fit statistical model for the cause of death. A summary of the statistical methods and variable relationships is shown in Table 8.2.

8.3.3. *Cold ischemia time*

The parameter of Cold Ischemia Time (CIT) is the number of hours the organ is chilled between procurement and transplantation, and is typically a function of the travel time required. In Ioannou's model CIT was categorized as less than 6.4, 6.4–8.8, 8.8–11.3, 11.3–14.3, and 14.3- 60 hours. The historical proportion of transplants with CITs in each category can be seen in Fig. 8.7. The figure indicates clear trends in decreasing CIT each year. In the 2004–2014 transplant data, approximately 4% of transplants had unknown CIT. We replaced these with the median CIT for each year and used multinomial regression to predict the CIT category for each donor-recipient pair using year as the indedpendent variable.

8.4. Simulation Model

The statistical models described in Section 8.3 were implemented in a simulation model to generate donated livers and recipients to forecast population level survival outcomes of liver transplant recipients. We implemented the model in Arena 14.7 (Rockwell Automation).

Table 8.2. Summary of statistical methods for modeling donor characteristics.

Dependent Variable	Independent Variable: Adult	Independent Variable: Pediatric	Statistical Model
Gender	—	—	Constant Discrete Distribution
Age	*Gender, Year*	—	Linear Regression
Race	*Gender, Year*	Gender, Year	Linear Regression
Obesity	*Gender, Age, Race, Year*	Race, Year	Logistic Regression
Alcohol Use	*Gender, Age, Race*	—	Logistic Regression
Diabetes	*Gender, Age, Race, Year, Obesity*	Year, Gender, Obesity	Logistic Regression
Cause of Death (Stroke or not)	*Gender, Age, Race, Year*	Gender, Race	Logistic Regression
Bilirrubin	—	—	Constant Discrete Distribution
DCD	*Year*	—	Linear Regression/ Constant Discrete Distribution
ALT	*Gender, Age, Race, Year*	Year	Ordered Logistic Regression
Utilization	*Gender, Age, Race, Obesity, Alcohol Use, Diabetes Cause of Death, DCD, ALT, Bilirrubin, Year*	Gender, Race, Obesity, Diabetes, Stroke, DCD, ALT, Bilirrubin, Year	Logistic Regression

8.4.1. *Model description*

Simulation used the statistical models discussed in Section 8.3.2 to probabilistically assign the attributes of each donated liver and then determine whether that liver was viable for transplant utilization.

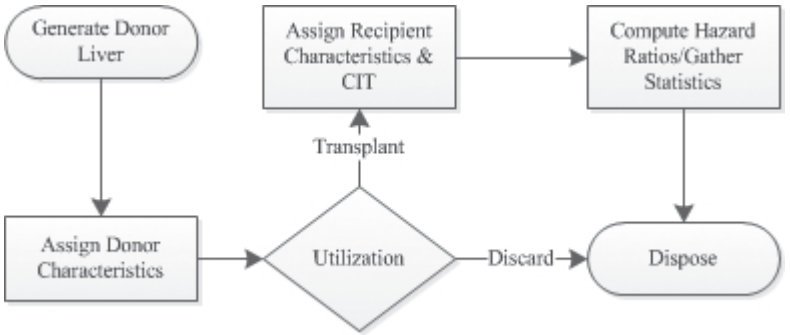


Figure 8.3. Simulation model structure showing interaction between statistical models.

For livers that were utilized, the attributes of the recipient and CIT were probabilistically generated based on the statistical models discussed in Section 8.3.1. Thus, each donated and utilized liver "pulls" a recipient. There are no matching characteristics modelled (such as blood type) that will preclude a donated liver from finding a recipient. While this model can certainly be refined to include this level of detail (as discussed in Section 8.6.2), this assumption is not too limiting as currently the number of patients on the list far outnumber the number of donations, such that it is very rare for a viable liver to be discarded. For each of these donor-recipient matches, a hazard rate according to Ioannou's survival model was calculated and recorded for each transplant, and the average overall transplant hazard rate was also recorded. Figure 8.3 provides a high-level overview of the simulation model structure.

8.4.2. *Model validation*

In order to validate our model, we compared the historical prevalence of all donor and recipient characteristics to the prevalence of the characteristics in the simulated donor and recipient populations. In the case of recipient MELD score, we compared the percentiles of historical MELD scores for each year and disease stratum to the distribution of the simulated MELD scores. For example, Table 8.3

Table 8.3. Recipient MELD score percentiles and historical versus simulated values.

Pctl.	Alc.C		HepB		HepC		PBC		Crypt.C		HCC		Other	
5th	10	*10*	*7*	*7*	8	*9*	10	*10*	*9*	*11*	*6*	*6*	8	*8*
10th	12	*12*	*9*	*7*	10	*10*	12	*12*	11	*12*	*7*	*7*	10	*10*
25th	16	*16*	12	*11*	14	*14*	15	*16*	15	*16*	8	*9*	14	*14*
50th	21	*21*	18	*18*	19	*18*	20	*19*	19	*20*	11	*11*	19	*19*
25th	27	*29*	*29*	*29*	*26*	*26*	*26*	*25*	*25*	*27*	*15*	*14*	*26*	*26*
90th	34	*37*	40	*38*	*33*	*35*	32	*35*	32	*36*	23	*19*	33	*34*
95th	39	*40*	40	*40*	*37*	*39*	*35*	*40*	*37*	*39*	28	*26*	39	*39*

shows the simulated versus historical MELD score percentiles for each recipient disease category in 2004. The simulated and historical distributions matched closely and passed the Kolmogorov-Smirnoff test at a 95% confidence level. For all other recipient and donor characteristics, the simulated prevalence was within +/–2% of the historical prevalence. Table 8.4 shows the average absolute difference of simulated versus historical prevalence for various recipient characteristics. Figure 8.4 shows the historical and simulated prevalence of key donor characteristics.

8.5. Results

The simulation was run for 15 replications for each year from 2004 to 2024, with 10,000 donated livers in each run. The number of replications was chosen to achieve a desired 95% confidence interval half-width of 2% or less about the estimates of donor liver utilization for all years.

8.5.1. *Proportional hazard model*

We evaluated survival outcomes based on calculated hazard ratios, as outlined in Ioannous survival model. For each individual, we calculated a risk score, X, by adding the adjusted regression coefficients

Table 8.4. Recipient characteristics, average absolute deviation between historical and simulated values.

Recipient Characteristic		Average (Simulated-Historical)
Recipient Age	18 to 43	0.67%
	43 to 50	0.70%
	50 to 57	1.11%
	57 to 63	1.56%
	over 63	1.40%
Recipient Race	White	0.73%
	Black	0.35%
	Hispanic	0.38%
	Other	0.30%
Recipient Diabetic	Diabetic	0.74%
Recipient BMI	1: 15 to <25	0.61%
	2: 25 to <30	0.66%
	3: 30 to <35	0.77%
	4: 35 to <40	0.60%
	5: 40 to <55	0.35%
Recipient Disease	Alcoholic cirrhosis	1.16%
	HepB	0.30%
	HepC	0.63%
	Primary Biliary cirrhosis	0.25%
	cryptogenic cirrhosis	0.42%
	hepatocellular carcinoma	1.28%
	other	0.50%

shown in Table 8.2 of [17]. These coefficients are associated with donor and recipient characteristics. Then, the probability of survival at time t is given below, where HazardRatio(X) = $e^{(X)}$.

$$S(t, X) = S_0(t)^{\text{HazardRatio}(X)} \tag{2}$$

The simulation results revealed no significant trends in overall average transplant hazard ratios. However, stratifying the simulated

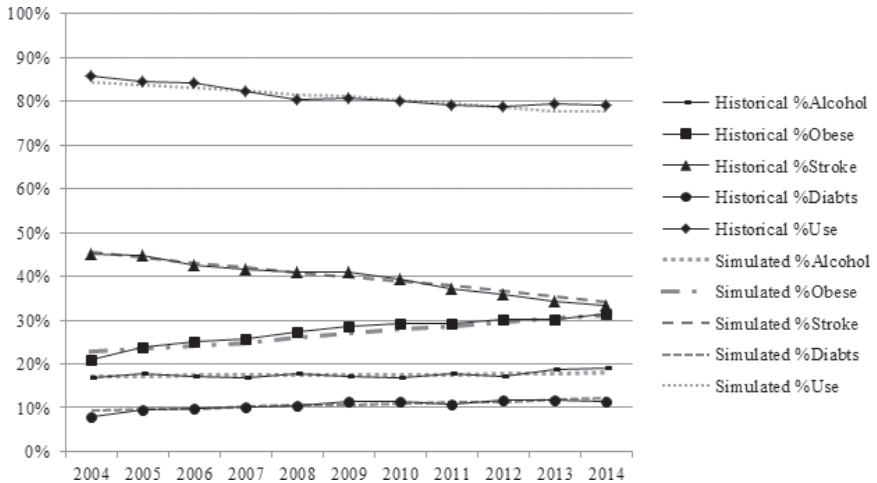


Figure 8.4. Historical versus simulated prevalence of donor characteristics.

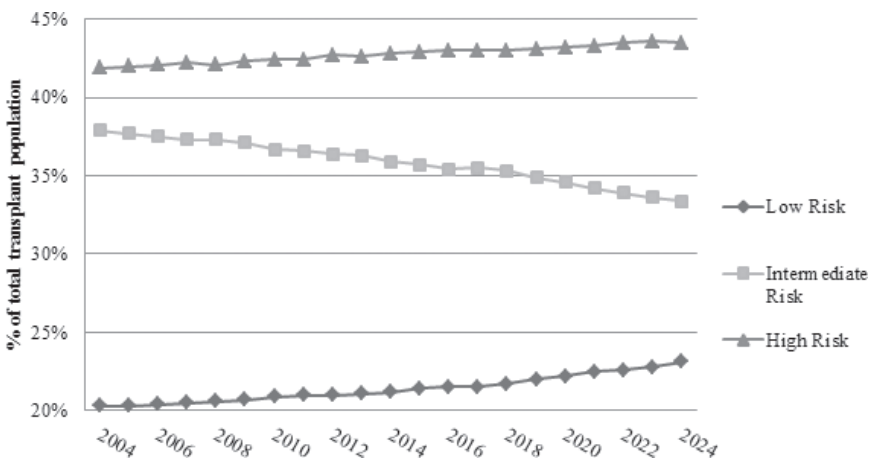


Figure 8.5. Trends in risk group shown as proportion of population in each risk group.

transplants into risk groups low, intermediate, and high (with hazard ratios less than 1, 1-1.5, and greater than 1.5, respectively) revealed gradual trends in the distribution of risk among transplant recipients, as seen in Fig. 8.5. The simulation results forecast growth in the pro-

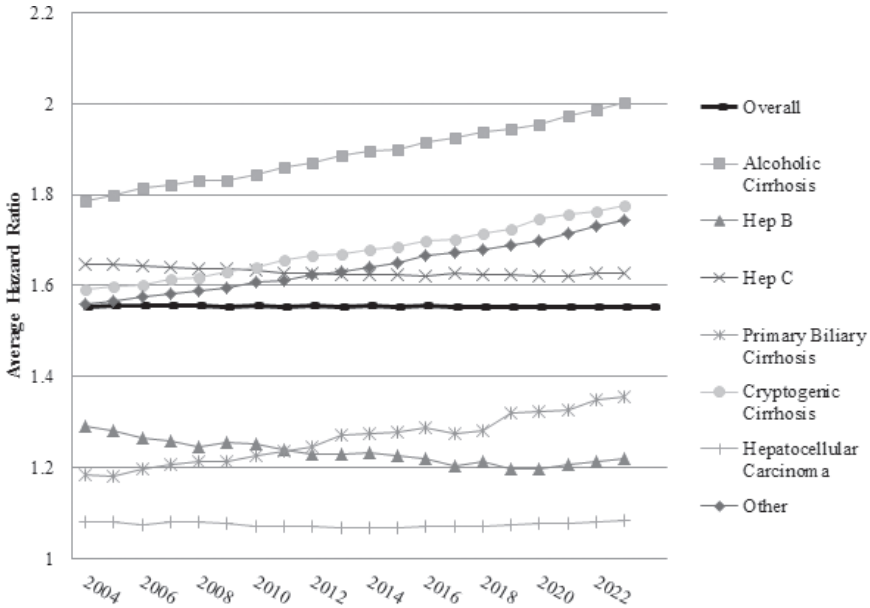


Figure 8.6. Trends in average forecasted hazard ratios over all transplants.

portion of both high and low risk transplants (0.14% and 0.08% per year on average, respectively) and a decreasing proportion of transplants that were categorized as intermediate risk (–0.22% per year on average), indicating a gradual polarization of risk in the transplant recipient population. A possible explanation for this polarization can be seen by looking at forecasted hazard ratios stratified by disease, as shown in Fig. 8.6. The increased risk is likely due to the recipient population (as the donors will likely get uniformly worse). The polarization effect could be driven by exception points (which were points added to lab MELD if the patient had cancer, for example) which moved patients up on the waitlist. As there were so many patients on the list, the median MELD score at transplant was increasing over time. Thus, patients with some diseases, such as cirrhosis, will have to get very sick to receive an offer while another cohort of patients with cancer (HCC) can have a low lab MELD and receive a transplant. The very sick patients are likely to have high risk, while the HCC patients will have low risk. Changes in donor and recipient populations were associated with both harmful and

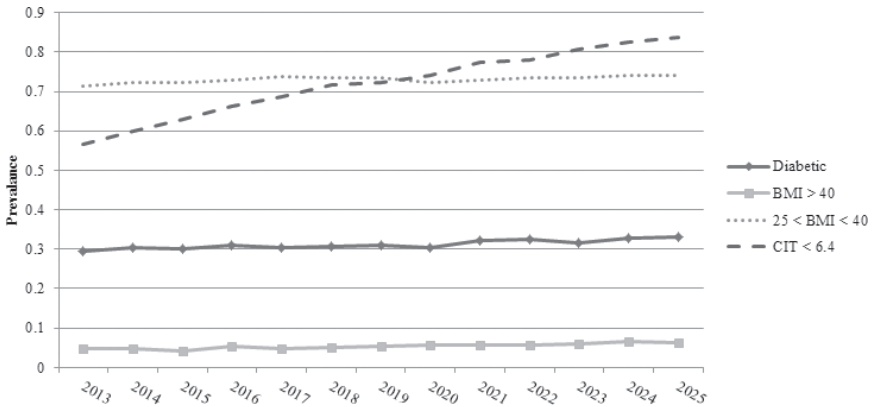


Figure 8.7. Trends in recipient characteristic prevalence, dashed lines are associated with decreased risk, while solid lines are associated with increased risk.

protective effects in Ioannou's survival model, creating canceling effects. This contributed to the stable overall average hazard ratios. For example, growing prevalence of diabetes and obesity in the recipient population (though these have been small) are associated with much increased risk; in contrast, the trend in decreasing CIT is associated with decreased risk. Figure 8.7 shows some examples of recipient characteristics that have competing effects on risk.

Although these trends effectively cancel each other out in the scope of this analysis, there may be a limit to how long these trends counteract each other. For example, the prevalence of diabetes may continue to increase, but it is unlikely that CIT will decrease much further due to operational constraints.

8.5.2. *D-MELD*

In addition to using the simulation model to forecast hazard ratios according to Ioannou's survival model, we also evaluated average D-MELD, the product of the donor age and recipient MELD score. D-MELD is a simple but effective predictor of post-transplant graft survival as described in Halldorson *et al.* [19]. As shown in Fig. 8.8, there are clearer trends in overall average D-MELD with an average increase of 7 points per year.

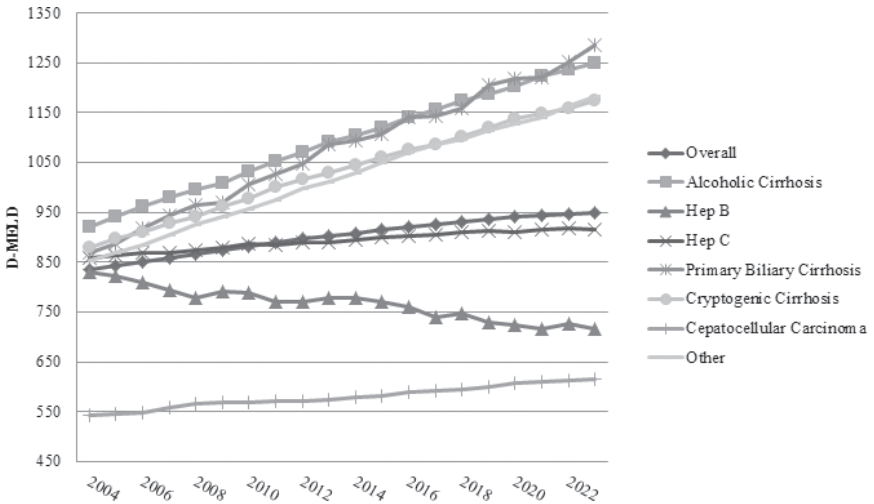


Figure 8.8. Trends in average forecasted D-MELD over all transplants.

8.6. Summary and Discussion

Although the simulation forecasts significant trends in D-MELD for transplants taking place in the next 10 years, the more detailed measure of Ioannou's hazard ratios do not demonstrate any changes in overall survival outcomes. However, in both the evaluated risk measures more drastic trends were seen when isolating each disease type, with some trends increasing while others are decreasing. These, in effect, cancel each other out such that average risk over all transplants does not reflect the underlying dynamics of survival risk in the transplant recipient population.

8.6.1. *Limitations*

The primary limitation of this study is the structure of random donor-recipient matching. This structure overlooks the careful allocation strategies that are implemented in matching donated organs to recipients, including biological and operational constraints such as blood type compatibility and regional allocation rules.

However, we believe that our modeling approach is capable of capturing population-level trends although estimates may be consistently conservative. Additionally, the dynamics of transplant waitlists are not in the scope of this analysis, although it may provide more granular insights.

Another limitation is that the survival model used to predict outcomes may benefit from incorporating updated medical realities such as Hepatitis C treatments, as well as more recent data.

8.6.2. *Future work*

Opportunities for future work include incorporating more detailed dynamics of donor-recipient matching, as well as waitlist population dynamics. These may provide additional insights about transplant recipient outcomes for more specific contexts.

The development of an updated survival model may better capture current dynamics of graft survival and risk. Lastly, the methodologies outlined in this analysis can be used as a basis for characterizing transplant survival in different scenarios, providing a framework to evaluate changes in allocation rules or medical technologies.

Acknowledgments

This work was supported, in part, by the National Science Foundation, under award CMMI-141833, PI-Mayorga.

References

1. CDC (2010). Number of first-listed diagnoses for discharges from short-stay hospitals. http://www.cdc.gov/nchs/data/nhds/10Detaileddiagnosesprocedures/2010det10_numberfirstdiagnoses.pdf. Last accessed August 31, 2016.
2. CDC (2013). Number of first-listed diagnoses for discharges from short-stay hospitals. http://www.cdc.gov/nchs/data/nvsr/nvsr64/nvsr64_02.pdf. Last accessed August 31, 2016.

3. UNOS (2016). Data reports. https://optn.transplant.hrsa.gov/data/view-data-reports/build-advanced/. Last accessed August 31, 2016.
4. Bertsimas, D, VF Farias and N Trichakis (2013). Fairness, efficiency, and exibility in organ allocation for kidney transplantation. *Oper Res*, 61, 1, 73–87.
5. Akan, M, O Alagoz, B Ata, FS Erenay and A Said (2012). A broader view of designing the liver allocation system. *Oper Res*, 60, 4, 757–770.
6. Kong, N, AJ Schaefer, B Hunsaker and MS Roberts (2010). Maximizing the efficiency of the US liver allocation system through region design. *Manage Sci*, 56, 12, 2111–2122.
7. Stahl, JE, N Kong, SM Shechter, AJ Schaefer and MS Roberts (2005). A methodological framework for optimally reorganizing liver transplant regions. *Med Decis Making*, 25, 1, 35–46.
8. Gentry, SE, AB Massie, SW Cheek, KL Lentine, E Chow, CE Wickliffe, N Dzebashvili, PR Salvalaggio, MA Schnitzler, DA Axelrod and DL Segev (2013). Addressing geographic disparities in liver transplantation through redistricting. *Am J Transplant*, 13, 8, 2052–2058.
9. Freeman, RB, AM Harper and EB Edwards (2002). Redrawing organ distribution boundaries: Results of a computer-simulated analysis for liver transplantation. *Liver Transplant*, 8, 8, 659–666.
10. Howard, DH (2002). Why do transplant surgeons turn down organs? A model of the accept/reject decision. *J Health Econ*, 21, 6, 957–969.
11. Alagoz, O, LM Maillart, AJ Schaefer and MS Roberts (2007). Determining the acceptance of cadaveric livers using an implicit model of the waiting list. *Oper Res*, 55, 1, 24–36.
12. Alagoz, O, LM Maillart, AJ Schaefer and MS Roberts (2004). The optimal timing of living-donor liver transplantation. *Manage Sci*, 50, 10, 1420–1430.
13. Alagoz, O, LM Maillart, AJ Schaefer and MS Roberts (2007). Choosing among living-donor and cadaveric livers. *Manage Sci*, 53, 11, 1702–1715.
14. Toro-Diaz, H, ME Mayorga, AS Barritt, ES Orman and SB Wheeler (2014). Predicting liver transplant capacity using discrete event simulation. *Med Decis Making*, 23, 9, 784–796.
15. Orman, ES, ME Mayorga, SB Wheeler, RM Townsley, HH Toro-Diaz, PH Hayashi and A Sidney Barritt (2015). Declining liver graft quality threatens the future of liver transplantation in the United States. *Liver Transplant*, 21, 8, 1040–1050.

16. Parikh, ND, D Hutton, W Marrero, K Sanghani, Y Xu and M Lavieri (2015). Projections in donor organs available for liver transplantation in the United States: 2014-2025. *Liver Transplant*, 21, 6, 855–863.
17. Ioannou, GN (2006). Development and validation of a model predicting graft survival after liver transplantation. *Liver Transplant*, 12, 11, 1594–1606.
18. Rana, A, M Hardy , K Halazun, D Woodland, L Ratner, B Samstein, J Guarrera, R Brown Jr and J Emond (2008). Survival outcomes following liver transplantation (soft) score: A novel method to predict patient survival following liver transplantation. *Am J Transplant*, 8, 12, 2537–2546.
19. Halldorson, J, R Bakthavatsalam, O Fix, J Reyes and J Perkins (2009). D-meld, a simple predictor of post liver transplant mortality for optimization of donor/recipient matching. *Am J Transplant*, 9, 2, 318–326.

9. Internet of Hearts — Large-Scale Stochastic Network Modeling and Analysis of Cardiac Electrical Signals

Chen Kan and Hui Yang

Complex Systems Monitoring, Modeling and Analysis Laboratory,
The Pennsylvania State University,
310 Leonhard Building,
University Park, PA 16802, USA

Abstract

Rapid advancement of mobile sensing and Internet-of-Things (IoT) technology provides an unprecedented opportunity to realize smart and connected health. However, large-scale IoT systems lead to big data. Realizing the full potential of big data depends, to a great extent, on the development of new human-centered computing methodologies for real-time health monitoring, on-the-fly disease diagnosis, and timely delivery of life-saving treatments. Thus far, very little has been done to develop advanced IoT technologies for smart monitoring and control of heart health. This chapter presents

a new IoT technology of Mobile and E-Network Smart Health (MESH) specific to the heart, also called the Internet of Hearts (IoH), to advance the cardiac mHealth with IoT sensing, stochastic modeling and network analytics. The MESH technology will enable and assist (1) the acquisition of electrocardiogram (ECG) signals pertinent to space-time cardiac dynamics anytime, anywhere; (2) real-time management and compact representation of multilead ECG signals; (3) big data analytics in large-scale IoT contexts. In particular, we first developed a spatiotemporal approach to visualize the real-time motion of 3D VCG cardiac vectors. Then, an optimal model-based representation algorithm was developed to facilitate the compression of ECG signals and the extraction of features pertinent to disease-altered signal patterns. Further, we developed stochastic network models for real-time patient-centered monitoring, modeling, and analysis of stochastic variations between heartbeats from an individual and among human subjects. The MESH technology shows a great potential in providing an indispensable and enabling tool for realizing smart heart health and wellbeing for the population worldwide.

9.1. Introduction

Cardiac diseases are the leading cause of death in the world. About 30% of global deaths (17.3 million) are due to cardiac diseases. According to the report from World Health Organization (WHO), this number will increase to 23 million by 2030. In United States, heart diseases are responsible for one in every four deaths, amounting to an annual loss of $448.5 billion [1]. Cardiac diseases claim more lives each year than the next four leading causes of death combined — cancer, chronic lower respiratory diseases, accidents, and diabetes mellitus. As opposed to chronic ones, most of the cardiac diseases are acute and can occur at any time in daily life [2]. For example, a heart attack is caused by the blockage in coronary arteries, which results in insufficient blood and oxygen supply to cardiac muscles. When a heart attack occurs, every minute counts. Patients who experience acute heart attacks are required to receive the treatment within 90 minutes after the onset of the symptom. A delay

could result in permanent heart muscle damage and increased risk of death. However, if the sign of heart attack is detected early, life-saving medications and treatments can be delivered to avoid hospitalization and even reduce the mortality rate. Therefore, the optimal management and treatment of cardiac diseases hinge on the identification of cardiac disorders in the early stage and the delivery of timely medical interventions.

In the past decade, mobile health (mHealth) has gained increasing attention from the health-care research community. Advances in sensing technology and the rapid expansion of mobile networks have made remotely monitoring of patient's condition and provision of timely feedback possible and affordable. mHealth technologies, therefore, offer a great opportunity to improve diagnosis, treatment, and adherence; increase access to health services, and lower health-care costs. The applications of cardiac mHealth have increased in recent years. Wireless sensors are readily available to measure single-lead electrocardiogram (ECG). Patients can forward recorded ECG signals to physicians and receive feedbacks remotely. However, the existing mHealth technologies are limited in their ability to analyze complex patterns of ECG signals for the identification of cardiac diseases. This is mainly because the spatiotemporal cardiac electrical activity manifests stochastic behaviors. It poses significant challenges on the existing mHealth systems, which implement simple algorithms to recognize disease patterns. It is well known that ECG signals are initiated at the sinoatrial (SA) node, then conducted in both atria, relayed through the atrioventricular (AV) node to further propagate through the bundle of His and Purkinje fibers toward ventricular depolarization and repolarization [3, 4]. Such electrical conduction, nevertheless, is a stochastic process and can be influenced by various types of uncertainties. For example, the excitation of SA node may be too slow or too fast, may pause, or fail to exit the SA region. To investigate the underlying mechanisms, researchers developed multiscale recurrence models [5–8] that revealed nonlinear stochastic dynamics in vectorcardiogram (VCG) signals. Furthermore, the process of orchestrated depolarization and repolarization of cardiac muscle cells are controlled by the orchestrated function of

individual ion channels in the cell membrane and are, thereby, coupled with real-world uncertainties [9, 10]. Notably, cardiac electromechanical function is closely related to cyclic changes in the differences between intracellular and extracellular concentration of ions. The potential difference increases as multiple ions travel across the cell membrane through ion channels. Ions flow through these channels and, thus, change the action potential across the cell membrane [11, 12]. The rate at which ionic channels open and close is in a stochastic manner and is based largely on the potential difference across the membrane.

The stochastic behavior of the cardiac electrical activity consists of two aspects: within-a-patient and between-patient stochastic dynamics. On the one hand, cardiac electrical activity within a patient demonstrates temporal dynamics. As shown in Fig. 9.1a, a 10-second ECG signal is generated from continuous monitoring. It may be noted that the amplitude of the 4th cycle of the ECG signal is smaller than the first three, so as the 8th cycle. Furthermore, the 6th cycle shows a significant S wave and an elevated T wave. Moreover, apparent variability can be identified even among those cycles that look similar, for example, cycle #1, #2, #3, #9, and #10. The stochastic behavior of cardiac activity for an individual patient is critical to the identification of arrhythmic events. Taking consideration of historical variabilities in cardiac activity is conducive to the delivery of personalized treatment planning. On the other hand, the cardiac activity is different between patients. As shown in Fig. 9.1b, 2-second ECG signals of six patients demonstrate big variability. For example, the heart rate is apparently different among these patients. Patient P1, P3, and P6 have only two ECG cycles, but the others have 2.5–3 cycles within 2 seconds. Also, the morphology of these ECG signals shows significant dissimilarities. Patient P3 shows inverted T waves (i.e., T wave is pointing downward instead of upward). P3 has an abnormal wave before the onsite of Q wave, and the R peak of P6 is notched. Notably, between-patient stochastic behaviors are closely pertinent to the disease-altered cardiac patterns. The detection and differentiation of cardiac diseases hinge on the effective characterization of both within-a-patient and between-patient stochastic behaviors.

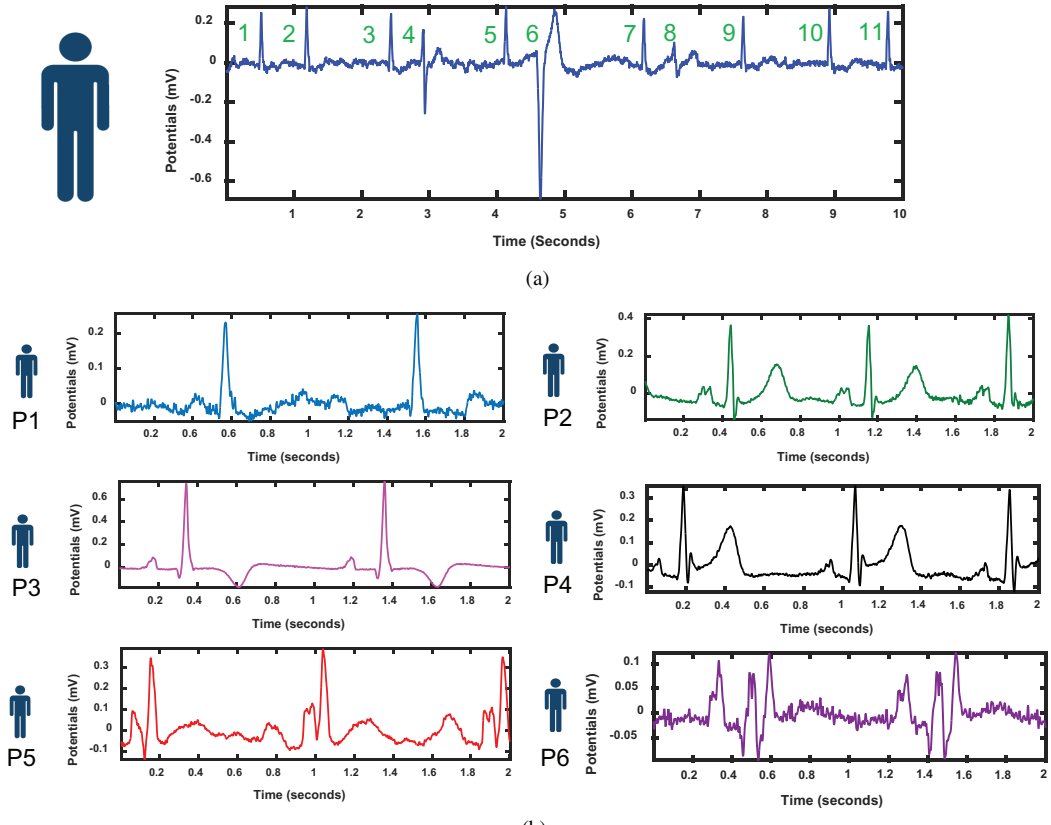


Figure 9.1. (a) Within-a-patient and (b) between-patient stochastic behaviors of cardiac electrical activity.

In the present investigation, we developed a new technology of Mobile and E-Network Smart Health (MESH) to advance the cardiac mHealth with stochastic modeling and network analytics [13]. The MESH technology is developed in the world's most widely used iOS mobile operating system, which is compatible with iPhone, iPad, and iPod Touch devices. In addition, it supplies in-situ information processing capabilities and enables physicians to access the patients' ECG signals in real time, remotely interact with the patients, and rapidly respond to life-threatening cardiac disorders. The MESH system is composed of three components: real-time visualization of three-dimensional (3D) VCG trajectory and feature detection, optimal model-based representation of ECG signals, and stochastic network modeling and online diagnosis.

The remainder of this chapter is organized as follows: Section 9.2 presents the background of ECG sensing and signal patterns; Section 9.3 throws light on the present analytical modules for large-scale ECG sensing systems; Section 9.4 provides the design of the MESH system, including the wearable sensor, MESH database, and smart phone applications; Section 9.5 presents marketing research, and Section 9.6 concludes this chapter.

9.2. Background

The human heart is essentially an autonomous electro-mechanical blood pump that operates near-periodically to maintain vital living organs. The heart consists of four compartments: right and left atria and right and left ventricles. This autonomous pump circulates blood in the body and constantly produces a sequence of electrical activities within every heartbeat. It is well known that an electrical activity begins in a specified pacemaker region, called the SA node, to excite the atrial muscle contraction. Then, the activity spreads through the upper chambers of the heart (the atria) and reaches the AV node. The AV node propagates the stimulus through bundle of His and Purkinje fibers toward the ventricles [3, 4]. The ordered stimulation, starting from the SA node, leads to the orchestrated

contraction of the heart, thereby, pumping the blood throughout the body.

The ECG system, designed by Augustus Waller in 1889 and further improved by Willem Einthoven in 1901, has been used for over 100 years for the monitoring of cardiac electrical activity and clinical diagnosis of cardiovascular disorders [14]. One lead ECG captures one-dimensional (1D) temporal view of a space-time cardiac electrical activity. Multi-lead ECG systems provide multi-directional views of such space-time dynamics [15]. A normal ECG tracing is often segmented into P wave, QRS complex, and T wave (see Fig. 9.2a [16]). Atrial depolarization (and systole) is represented by the P wave, ventricular depolarization (and systole) is represented by the QRS complex, and ventricular repolarization (and diastole) is represented by the T wave [17, 18]. It may be noted that ECG signals contain a wealth of dynamic information pertinent to cardiac operations, which is indispensable for cardiac care — from monitoring and diagnosis to treatment planning to smart health management. Existing time-domain algorithms were developed to quantify the characteristics of ECG wave deflections (i.e., P, QRS, and T waves) for the identification of cardiac diseases. Examples of ECG features

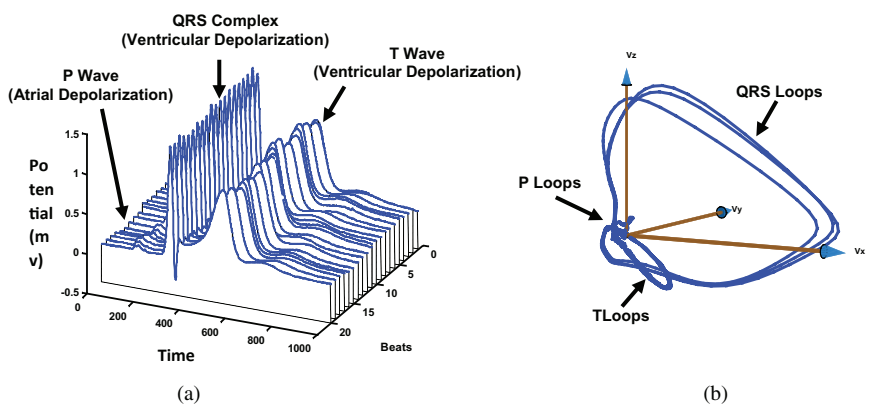


Figure 9.2. Two types of cardiac signals: (a) 2D ECG cycles and (b) 3D VCG loops.

include PR interval, RR interval, ST elevation/depression, QT interval, and R amplitude.

However, time-domain projections of space-time cardiac electrical activity will diminish important spatial information of cardiac pathological behaviors. As such, medical decisions that are made can be significantly influenced by such an information loss [3]. Therefore, 3-lead vector cardiogram (VCG) is designed to provide multidirectional views of space-time electrical activity. VCG observes the heart potentials as a cardiac vector in three orthogonal components instead of the scalar amplitude (ECG curve) [19]. In VCGs, the mutually orthogonal bipolar measurements are taken by placing parallel electrodes on the opposite sides of the torso. As shown in Fig. 9.2b, VCG signals contain P loops, QRS loops, and T loops, which correspond to P wave, QRS complex, and T wave in the ECG, respectively. Dower *et al.* and our previous studies [20–22] have demonstrated that 3-lead VCG can be linearly transformed to 12-lead ECG by multiplying a generalized transformation matrix. Thus, the information in 12-lead ECG is redundant and the 3-lead VCG surmounts not only the information loss in 1-lead ECG but also the redundant information in 12-lead ECG.

In clinical practice, the 12-lead ECG is widely used because physicians are trained and are accustomed to using them. It has, thus, proven its value, time-tested, and considered as the gold standard. It is generally difficult for physicians to interpret disease patterns via the high-dimensional VCGs. However, VCGs capture important space-time information of cardiac electrical activity, which is not contained in ECG signals. The methodologies developed in our previous research were proved to be efficient and effective for identifying disease patterns in VCG signals. Those algorithms have fueled increasing interests in VCG signals. However, they have not been applied to clinical practice due to lack of user-friendly software. Therefore, there is a need to develop software that implements those advanced algorithms. MESH incorporates novel pattern recognition algorithms that will serve as a tool to enable and assist physicians in characterizing VCG patterns and identifying early signs of cardiac disorders. The MESH system not only enables access to patients'

data anywhere and anytime but also extracts valuable diagnostic information from the signals to help physicians in the decision-making process. MESH is designed to enable physicians and nurses to access and visualize the patients' ECG signals in real time, as well as timely analysis of patient's data and rapidly respond to life-threatening cardiac disorders.

9.3. Analytical Modules

As shown in Fig. 9.3, the proposed MESH system consists of three analytical modules. We first develop a spatiotemporal representation approach to visualize the real-time dynamics of 3D VCG trajectories. This enables physicians and nurses to easily interpret the high-dimensional VCG patterns and extract space-time characteristics. Second, an optimal model-based representation algorithm is developed to facilitate the compression of cardiac signals and extraction of features pertinent to the disease-altered cardiac activity. Third, a stochastic network model is designed for real-time patient-centered

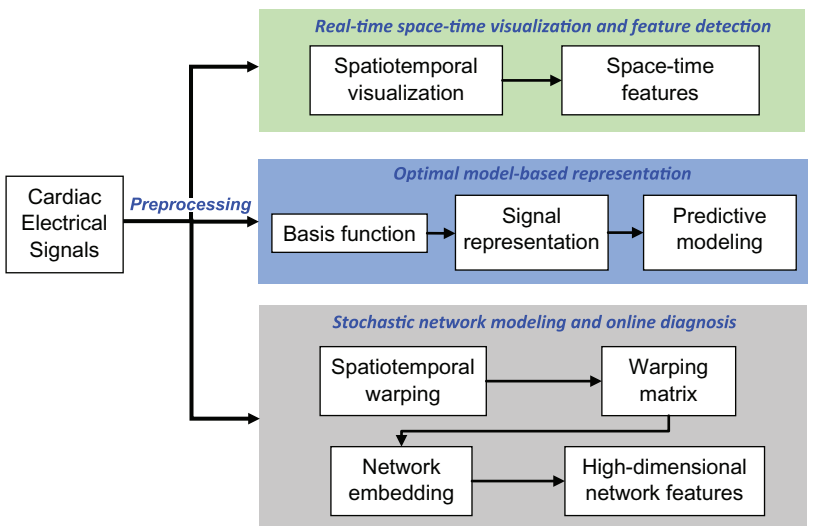


Figure 9.3. The overall structure of the proposed MESH system.

monitoring of cardiac variations. The developed spatiotemporal warping algorithm characterizes the cardiac variations in a warping matrix, which is further embedded into a high-dimensional network to facilitate classification and prediction of patients' cardiac conditions.

9.3.1. *Real-time spatiotemporal visualization and feature extraction*

ECG signals are recorded on body surface to track the continuous dynamic details of cardiac functioning. Such valuable real-time information is usually unavailable in static and discrete clinical laboratory tests, for example, computer imaging, chest x-ray, and blood enzyme test. Even if routine laboratory examinations are performed multiple times per day, discontinuity often fails to prevent the lethal consequences of acute cardiac disorders. The awareness about the importance of real-time cardiac monitoring for the early identification pathological patterns is increasing as it tracks cardiac dynamic behaviors, as opposed to static screenshots.

However, lead ECG signals only capture one perspective temporal view of the space-time excitation and propagation of cardiac electrical activities. Multiple lead ECG systems, for example, 12-lead ECG and 3-lead VCG, are designed to capture the multi-directional view of space-time cardiac electrical activities [23]. Time-domain visualization is the traditional routine for representing cardiac electric signals. It is the major function of most of the existing cardiac mHealth systems. The medical doctors are used to the time-domain identification of cardiac disease patterns. Therefore, this module is preserved in MESH. The characteristic points of cardiac signals, for example, locations of R peak and the end of T wave, are automatically detected by implementing the wavelet-based algorithm developed in our previous research.

However, cardiac electrical dynamics are initiated and propagated spatiotemporally. The projection of spatiotemporal activity into 1D time domain diminishes important spatial information underlying cardiac electrical activities. In MESH, a novel dynamic

spatiotemporal visualization of VCG signals is implemented [23]. In the Frank XYZ lead system, VCG is represented as three orthogonal scalar measurements with respect to time, which is given as:

$$\begin{cases} v_x = f(t) \\ v_y = g(t). \\ v_z = h(t) \end{cases}$$

The dynamic VCG signal representation embeds the cardiac vector, composed of three scalar measurements, in real time. As shown in Fig. 9.4, three scalar x, y, and z components are plotted in the top and the simultaneous 3D movement of cardiac vectors in the bottom.

The top plot displays VCG signals in three-vector components as a function of time, and the bottom part shows the real-time cardiac vector movement in the 3D space. Head (green) gives the current position of cardiac vector. Body (red) indicates the direction and rotation of cardiac vector movements [23].

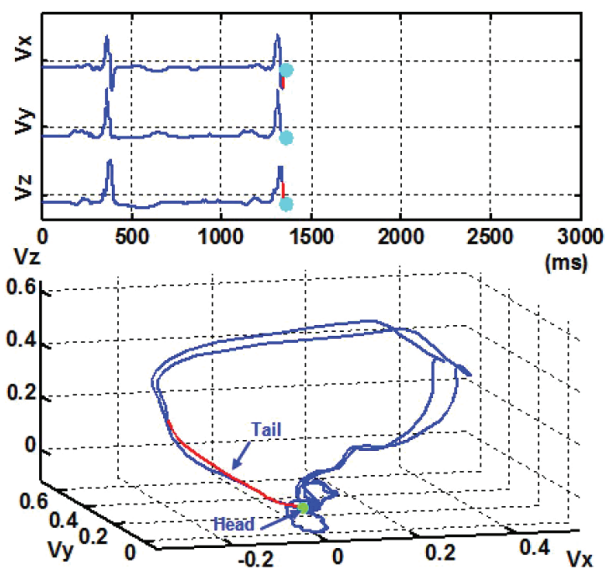


Figure 9.4. Real-time spatiotemporal VCG representation.

This real-time spatiotemporal VCG representation makes it easier to integrate with prior knowledge and experiences of time-based ECG. As shown in Fig. 9.4, this representation consists of three components, namely, head (green), body (red), and tail (blue). Head gives the current position of the cardiac vector. Body records a short history of the cardiac vector movements, which clearly indicates where the current vector is from. It avoids the confusion regarding the group of heart activity to which the current cardiac vector belongs as they usually intersect at the isoelectric points. The tail provides full history pertinent to the complete topological shape of VCG state space. By following the cardiac vector movement with respect to time, the P, QRS, and T waves will be easily located in the VCG state space [23].

The real-time visualization of spatiotemporal ECG signals is an enabling tool that can be used in clinical practices of cardiac care. This approach incorporates additional dynamical properties of cardiac vector movements (such as curvature, velocity, octant, and phase angle) with the color coding scheme, which can be used for the interpretation of high-dimensional cardiac vectors by physicians or nurses. Our prior research [23] showed that the proposed dynamic VCG surmounts some drawbacks of time-domain representation and provides critical spatial, as well as temporal information of the heart dynamics. The cardiovascular pathological patterns are found to be effectively captured by this new 3D dynamic representation approach. The presence of both spatial and temporal characteristics in dynamic representation improves the automatic assessment of cardiovascular diseases with the use of VCG signals.

9.3.2. *Optimal model-based representation*

The proposed MESH system enables long-term continuous cardiac monitoring. However, continuous sensing in days, months, and even years generates enormous amount of data, which contains a rich amount of information pertinent to the evolving dynamics of process operations. As such, it provides physicians with a spatially and temporally data-rich environment in the process of medical decision-making.

Big data poses significant challenges for human experts (e.g., physicians, nurses, and quality technicians) to accurately and precisely examine all the generated high-dimensional sensor signals for fault diagnosis and quality inspection. Moreover, the proliferation of sensing data also provides an unprecedented opportunity to develop sensor-based methodologies for realizing the full potential of multidimensional sensing capabilities toward real-time process monitoring and disease diagnosis.

In MESH, a new model-driven parametric monitoring strategy [16, 24] is developed for the detection of dynamic fault patterns in high-dimensional functional profiles that are non-linear and non-stationary. Specifically, a sparse basis function model is developed to represent high-dimensional functional profiles, which minimizes the number of basis functions involved but maintains sufficient explanatory power. As such, large amount of data is reduced to a parsimonious set of model parameters (i.e., weight, shifting, and scaling factors in basis functions) while preserving the signal information.

The 3D VCG is represented as the superposition of M multiscale basis functions:

$$\vec{v}(t,w) = \vec{w}_0 + \sum_{j=1}^{M} \vec{w}_j \vec{\varphi}_j \left((t - \mu_j) / \sigma_j \right) + \varepsilon,$$

where $\varphi(t)$ is the general basis function form, which is not limited to Gaussian function, μ_j is the shifting factor, and σ_j is the scaling factor. The objective is to minimize the representation error, that is, $\text{argmin}\left[\left\| \vec{\boldsymbol{v}}(t) - \vec{\boldsymbol{w}}_0 - \sum_{j=1}^{M} \vec{\boldsymbol{w}}_j \vec{\boldsymbol{\varphi}}_j (t)^2 \right\|, \{\boldsymbol{w}, M, \boldsymbol{\varphi}(t)\} \right]$, between VCG signals and basis function models. In a matrix form, the basis function model is rewritten as $\boldsymbol{V} = \boldsymbol{W}^T \boldsymbol{\varphi}$, where $\boldsymbol{W}$ is the weight matrix and $\boldsymbol{\varphi}$ is the basis function matrix.

An iterative procedure, i.e., a matching pursuit algorithm [25], was developed to search the suboptimal solution based on characteristic wave patterns in the VCG/ECG signals. The VCG matching pursuit method is started from an initial approximation $S^{(0)} = 0$, residual $R^{(0)} = \vec{v}(t)$, and dictionary $D = \{\varphi_j(t),\ j = 1, 2, \ldots, N\}$.

The first step identifies the basis function in the dictionary that best correlates with the residual, that is, finding $\gamma 0$ such that $\left|\langle R^{(0)}, \varphi^{(\gamma 0)}\rangle\right| = \max\left|\langle R^{(0)}, \varphi^{(\gamma)}\rangle\right|$, $\gamma \in N$ and $\varphi^{(\gamma 0)} \in D$. Then, the current approximation will be $s^{(1)} = s^{(0)} + \langle R^{(0)}, \varphi^{(\gamma 0)}\rangle \varphi^{(\gamma 0)}$, and the residual is defined as $R^{(1)} = R^{(0)} - \langle R^{(0)}, \varphi^{(\gamma 0)}\rangle \varphi^{(\gamma 0)}$. If the orthogonal wavelet bases are used, it may be noted that $\varphi^{(\gamma 0)}$ is orthogonal to $R^{(1)}$ because:

$$\begin{aligned}\langle \varphi^{(\gamma 0)}, R^{(1)}\rangle &= \langle \varphi^{(\gamma 0)}, R^{(0)} - \langle R^{(0)}, \varphi^{(\gamma 0)}\rangle \varphi^{(\gamma 0)}\rangle \\ &= \langle \varphi^{(\gamma 0)}, R^{(0)}\rangle - \langle \varphi^{(\gamma 0)}, \langle R^{(0)}, \varphi^{(\gamma 0)}\rangle \varphi^{(\gamma 0)}\rangle \\ &= \langle \varphi^{(\gamma 0)}, R^{(0)}\rangle - \langle R^{(0)}, \varphi^{(\gamma 0)}\rangle = 0\end{aligned}$$

Hence, $\langle R^{(0)}, \varphi^{(\gamma 0)}\rangle \varphi^{(\gamma 0)}$ is also orthogonal to $R^{(1)}$ so that

$$\left\|R^{(0)}\right\|^2 = \left\|R^{(1)}\right\|^2 + \left\|\left\langle R^{(0)}, \varphi^{(\gamma 0)}\right\rangle \varphi^{(\gamma 0)}\right\|^2$$

At step $j + 1$, the residual $R^{(j+1)}$ is treated as $R^{(0)}$ in the first step, yielding

$$R^{(j+1)} = R^{(j)} - \langle R^{(j)}, \varphi^{(\gamma j)}\rangle \varphi^{(\gamma j)}$$

$$s^{(j+1)} = \sum_{i=1}^{j} \langle R^{(i)}, \varphi^{(\gamma i)}\rangle \varphi^{(\gamma i)}$$

After M such steps, one has a representation of the form of additive decomposition:

$$v(t) = \sum_{i=1}^{M-1} \langle R^{(i)}, \varphi^{(\gamma i)}\rangle \varphi^{(\gamma i)} + R^{(M)}$$

The adaptive algorithm will stop when the residual sum of squares is less than a small threshold at step M (i.e., $\left\|\boldsymbol{R}^{(M)}\right\| < \varepsilon$). An intrinsic feature of matching pursuit algorithm is that when the dictionary has orthogonal bases, it works perfectly after a few steps

yielding a sparse adaptive representation using only a few basis functions. An example of fitting high-dimensional nonlinear profile using the superposition of basis functions is shown in Fig. 9.5. It may be noted that the basis function model (red/solid) effectively represents the original data (blue/dashed).

It may be noted that optimal representation of 3D VCG topology in the MESH system will lead to the following benefits:

- *Feature extraction*: The model parameters such as weights, shifting, and scaling factors in the basis functions can be potentially used as features for the diagnostic application. As a result, large amount of VCG and ECG data is reduced to a limited amount of features (i.e., model parameters) while preserving the same information.
- *Data compression*: It is well known that hundreds of gigabytes of VCG and ECG data will be stored in the real-time cardiac monitoring. Since the basis function model yields a good representation (>99%) of real-world VCG signals, model parameters can be saved instead of long-term VCG signals.
- *Algorithm evaluation*: This proposed basis function model is data-driven and can be fitted to ECG signals from different kinds of cardiovascular diseases. The fitted model for different pathologies can generate large amount of VCG/ECG signals that can be

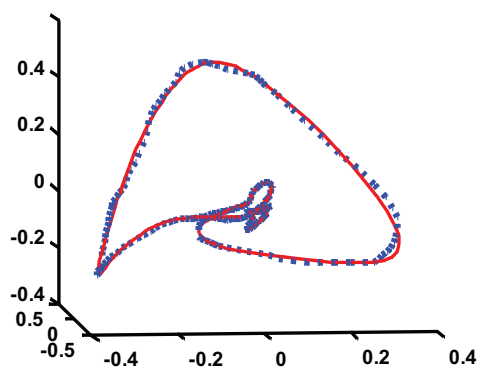


Figure 9.5. 3D trajectory of VCG signals from basis function model (red/solid) and real-world data (blue/dashed) [16].

used to test the algorithms of QRST cancellation, adaptive filtering, and classification.

- *Disease prognostics*: Because the basis function model captures all the characteristics from actual data, real-time ECG monitoring signals can be compared with the model representation trained in healthy condition. The differences of pattern similarity can be used as a performance measure for the prognostic purpose.

The model parameters and their derivatives can be used as features for the detection of process faults. However, the dimensionality of these features is high and can potentially lead to sensitive predictive models. Thus, we further utilize lasso-penalized logistic regression model [16] to investigate the "redundancy" and "relevancy" properties between these parameter-based features and fault patterns to identify a sparse set of sensitive predictors from a large number of features for fault diagnostics.

Let $p(\boldsymbol{x}, \boldsymbol{\beta})$ be the probability for y to be a success ($y = 1$) and, thus, $1 - p(\boldsymbol{x}, \boldsymbol{\beta})$ is the probability for y to be a fault ($y = 0$), where $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2,...,\beta_p)^T$ is the coefficient vector. The logistic regression model is:

$$\log\left(\frac{p(\boldsymbol{x},\boldsymbol{\beta})}{1-p(\boldsymbol{x},\boldsymbol{\beta})}\right) = \boldsymbol{\beta}^T\boldsymbol{x}$$

The likelihood function of $\boldsymbol{\beta} = (\beta_0, \beta_1,...,\beta_p)^T$, given the observation data $\boldsymbol{X} = (\boldsymbol{x}_1, \boldsymbol{x}_2,...,\boldsymbol{x}_n)^T$, $\boldsymbol{y} = (\boldsymbol{y}_1,...,\boldsymbol{y}_n)^T$ is:

$$\prod_{i=1}^{n} p(\boldsymbol{x}_i,\boldsymbol{\beta})^{y_i}\left(1-p(\boldsymbol{x}_i,\boldsymbol{\beta})\right)^{1-y_i}$$

As such, the log likelihood function becomes:

$$\begin{aligned} L(\boldsymbol{\beta}|\boldsymbol{X},\boldsymbol{y}) &= \sum_{i=1}^{n}\left[y_i\log\left(p\left(\boldsymbol{x}_i,\boldsymbol{\beta}\right)\right)+(1-y_i)\log\left(1-p\left(\boldsymbol{x}_i,\boldsymbol{\beta}\right)\right)\right] \\ &= \sum_{i=1}^{n}\left[y_i\boldsymbol{\beta}^T\boldsymbol{x}_i-\log\left(1+e^{\boldsymbol{\beta}^T\boldsymbol{x}_i}\right)\right] \end{aligned}$$

The lasso-penalized logistic regression is formulated to minimize the following objective function with the constraint that the upper limit of L_1-norm of $\boldsymbol{\beta}$ is less than C,

$$\min_{\boldsymbol{\beta}} - L(\boldsymbol{\beta} | \boldsymbol{X}, \boldsymbol{y})$$

$$\text{subject to } \|\boldsymbol{\beta}\|_1 \leq C$$

This is equivalent to solve the following unconstrained optimization problem, with λ be the regularization parameter:

$$\min_{\boldsymbol{\beta}, \lambda} - L(\boldsymbol{\beta} | \boldsymbol{X}, \boldsymbol{y}) + \lambda \|\boldsymbol{\beta}\|_1$$

The optimal solution $\boldsymbol{\beta}$ of the unconstrained optimization problem given λ also solves the constrained minimization problem with $C = \|\boldsymbol{\beta}\|_1 = \sum_{i=1}^{p} |\beta_i|$. To solve this constrained optimization problem, let us first obtain the solution to the general logistic regression model. The objective function of general logistic regression model is as follows:

$$\min_{\boldsymbol{\beta}} - L(\boldsymbol{\beta} | \boldsymbol{X}, \boldsymbol{y})$$

From the Newton-Raphson algorithm, it may be noted that the update of parameters is obtained by approximating the objective function with the second-order Taylor expansion. Let $\boldsymbol{\beta}^{(k)}$ be the current parameters, then Newton–Raphson method finds the new set of parameters $\gamma^{(k)}$ based on the quadratic approximation:

$$\boldsymbol{\gamma}^{(k)} = \left(\boldsymbol{X}^T \boldsymbol{W} \boldsymbol{X}\right)^{-1} \boldsymbol{X}^T \boldsymbol{W} \boldsymbol{z},$$

where $\boldsymbol{z} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{W}^{-1}(\boldsymbol{y} - \boldsymbol{p})$ and $\boldsymbol{W}$ is the diagonal matrix with $(\boldsymbol{W})_{ii} = p(\boldsymbol{x}_i, \boldsymbol{\beta})\left(1 - p(\boldsymbol{x}_i, \boldsymbol{\beta})\right)$. As such, solving for $\gamma^{(k)}$ is equal to finding the solution to the following weighted least squares problem:

$$\boldsymbol{\gamma}^{(k)} = \arg\min_{\boldsymbol{\gamma}} \left\| \left(\boldsymbol{W}^{\frac{1}{2}} \boldsymbol{X} \right) \boldsymbol{\gamma} - \boldsymbol{W}^{\frac{1}{2}} \boldsymbol{z} \right\|_2^2$$

For lasso-penalized logistic regression, there is a need to add the L_1 constraint to the unregularized logistic regression to ensure $\|\boldsymbol{\gamma}\|_1 \leq C$, that is,

$$\min_{\boldsymbol{\gamma}} \left\| \left(\boldsymbol{W}^{\frac{1}{2}} \boldsymbol{X} \right) \boldsymbol{\gamma} - \boldsymbol{W}^{\frac{1}{2}} \boldsymbol{z} \right\|_2^2$$
$$\text{Subject to} \left\| \boldsymbol{\gamma} \right\|_1 \leq C$$

As a result, the lasso-penalized logistic regression is transformed to an iteratively reweighted least square problem. At each iteration, we update the $\boldsymbol{W}^{\frac{1}{2}} \boldsymbol{X}$ and $\boldsymbol{W}^{\frac{1}{2}} \boldsymbol{z}$, based on the new estimate of coefficients. After $\boldsymbol{\gamma}^{(k)}$ is obtained, we update $\boldsymbol{\beta}^{(k)}$ by:

$$\boldsymbol{\beta}^{(k+1)} = (1-\theta)\boldsymbol{\beta}^{(k)} + \theta \boldsymbol{\gamma}^{(k)}$$

where $\theta \in [0,1]$ is the learning rate for the parameter update. In this study, we adopted the coordinate descent algorithm to solve the regularized problem. If we write $\boldsymbol{W}^{\frac{1}{2}} \boldsymbol{X} = \overset{V}{\boldsymbol{X}}$ and $\boldsymbol{W}^{\frac{1}{2}} \boldsymbol{z} = \overset{V}{\boldsymbol{y}}$, only one β_j is changed at each time, while the other parameters $\beta_k (k \neq j)$ stay the same.

The lasso penalized logistic regression model is implemented in MESH to investigate the "redundancy" and "relevancy" properties between features and fault patterns, thereby identifying a sparse set of sensitive predictors for fault diagnostics. This model was evaluated in our previous study, and the experimental results showed that more than 60% of features had the KS statistic greater than the critical value 0.17, indicating significant differences between control and fault conditions. Furthermore, the lasso-penalized logistic regression model yields a superior accuracy of 97.13%, with a parsimonious set of 81 features. The proposed approach facilitates the modeling and characterization of high-dimensional nonlinear profiles and provides effective predictors for real-time fault detection, thereby promoting the understanding of fault-altered spatiotemporal patterns in the complex cardiovascular systems.

9.3.3. *Stochastic network modeling and online diagnosis*

A remarkable feature of MESH is its information-processing capability to perform spatiotemporal recognition of disease patterns using 3D trajectories of cardiac electric signals. As shown in Fig. 9.6, there is spatiotemporal dissimilarity between the 3-lead VCGs of MI (red dashed loops) and HC (blue solid loops) subjects. The quantification of such dissimilarity will provide a great opportunity for the identification of cardiovascular diseases. However, it is challenging to measure the spatiotemporal dissimilarity between two functional signals in both space and time. Due to phase shift and discrete sampling, two VCG signals can be misaligned, for example, both signals show a typical pattern and yet there are variations in shape, amplitude, and phase between them. In the clinical practice, various methods are developed to measure the dissimilarities between misaligned signals. Figure 9.7 illustrates some of them using simple two-dimensional (2D) ECG signals. To compare the ECG signals (blue and red), the intuitive way is to directly take the difference between them (see Fig. 9.7a). As such, the difference may be huge even for similar signal patterns because of the misalignment. For example, the QRS wave (ventricular depolarization) of the blue ECG may be compared to the P wave (atrial depolarization) of the red ECG, which

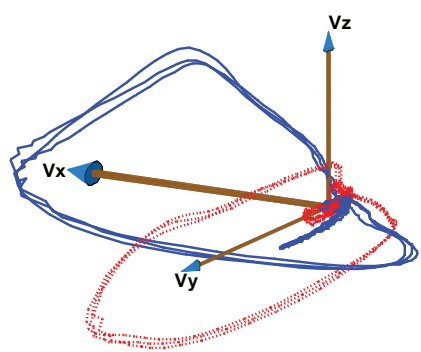


Figure 9.6. Spatiotemporal VCG signals of control (blue/solid) and diseased subjects (red/dashed).

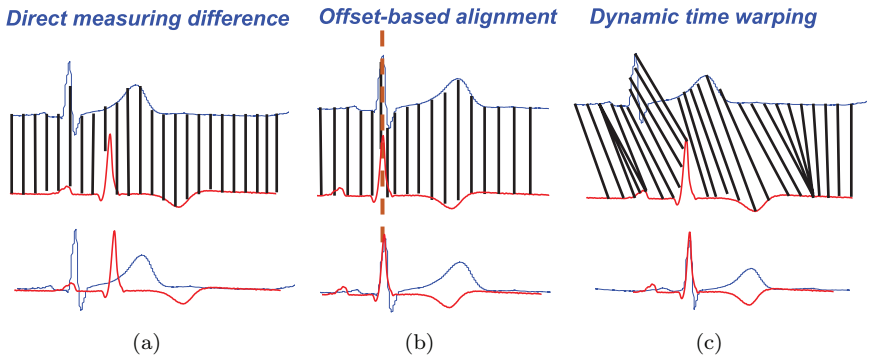


Figure 9.7. Measuring dissimilarities between misaligned ECG cycles: (a) Direct difference, (b) Offset based alignment, and (c) Dynamic time warping.

generates misleading results. For years, physicians used offset-based alignment to improve the solution. In other words, R peaks from two ECGs are first aligned together and then take the difference (see Fig. 9.7b). In this way, the ventricular depolarization of two subjects are compared together, but the atrial depolarization (P wave) and ventricular repolarization (T wave) are still misaligned. Finally, dynamic time warping [26, 27] is a viable method that may help optimally align two ECG signals (see Fig. 9.7c). Such an alignment is critical to compare the corresponding electrical activity of heart chambers. For example, we are comparing the ventricular depolarization (i.e., QRS complex) for two subjects, as opposed to the incorrect comparison between atrial depolarization (P waves) from one subject and ventricular depolarization from the other subject.

Importantly, the first step of stochastic network modeling is to implement our dynamic spatiotemporal warping approach to measure the dissimilarities between space-time functional recordings [3, 28]. As opposed to traditional time-domain warping (see Fig. 9.7c), spatiotemporal warping is innovatively created to solve the problem of misalignment in both space and time. As shown in Fig. 9.8, P, QRS, and T loops are aligned for two subjects in both space and time. Notably, little work has been done to measure the differences between VCG loops by means of dynamic time warping. However, 3-lead VCG signals are analogous to the voice from the heart.

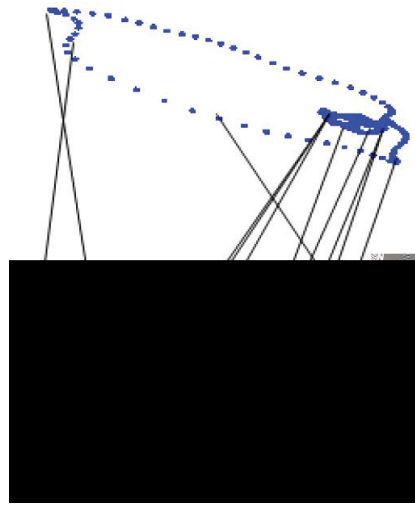


Figure 9.8. Spatiotemporal alignment of 3-lead VCG signals [3].

Our algorithm is the first of its kind to utilize space-time warping of VCG signal patterns for the identification of disease patterns and has been granted two patents [29, 30].

Given two 3D VCG signals $\overrightarrow{v_1}(t)$ and $\overrightarrow{v_2}(t)$, the time-normalized spatial distance between $\overrightarrow{v_1}(t)$ and $\overrightarrow{v_2}(t)$ is calculated as $\sum_{(t_i,t_j)\in p}\left\|\overrightarrow{v_1}(t_i)-\overrightarrow{v_2}(t_j)\right\|$ by alignment p. The warping path $p(i, j)$ connects (1, 1) and (N_1, N_2) in a 2D square lattice as well as satisfying constraints such as monotonicity condition and step size condition. To find the optimal path, an exhaustive search of alignment path is intractable and computationally expensive. However, this problem is solved efficiently using dynamic programming (DP) methods. The DP algorithm is started at the initial condition: $g(1,1)=d(1,1)=\left\|\overrightarrow{v_1}(t_1)-\overrightarrow{v_2}(t_1)\right\|$ and the warping window $|i-j|<r$. The algorithm is searching forward as follows:

$$g(i,j)=min\begin{pmatrix} g(i,\ j-1)+\ d(i,\ j) \\ g(i-1,\ j-1)+\ d(i,\ j) \\ g(i-1,\ j)+\ d(i,\ j) \end{pmatrix}$$

Finally, the time-normalized spatial distance is calculated as follows:

$$\Delta\left(\overrightarrow{v_1}, \overrightarrow{v_2}\right) = \frac{g\left(N_1, N_2\right)}{N_1 + N_2}$$

where N_1 and N_2 are the length of $\overrightarrow{v_1}(t)$ and $\overrightarrow{v_2}(t)$, respectively. The $\Delta\left(\overrightarrow{v_1}, \overrightarrow{v_2}\right)$ represents the spatiotemporal dissimilarity between two multidimensional functional recordings. Therefore, disease-altered characteristics of 3-lead VCG signals are obtained in the warping matrix.

However, it may be noted that the warping matrix itself cannot be used as features for the identification of disease properties in classification models. In addition, the measure of Euclidean distance is not directional and can mix the distances that are equal in magnitudes but along different spatial directions. A novel method needs to be developed to transform the warping matrix into feature vectors that preserve the warping distances among functional recordings. The spatial embedding method represents the functional recordings as the points in a high-dimensional space. These points can be used as feature vectors that recover not only the distance matrix but also directional differences between functional recordings [28].

This is similar to a network problem, that is, how to reconstruct the location of nodes in a high-dimensional space if the node-to-node distance matrix is known. As shown in Fig. 9.8, a network comprises a number of nodes that are connected by edges. Each node stands for an individual component in the system, and the edges show the relationship (e.g., distances or causal relationships) between nodes. As given in Fig. 9.9a, assume the distance matrix Δ for five nodes is known. If the network is reconstructed in the 3D space, this is analogous to optimally identify the coordinate vector $\boldsymbol{x}_i = (x_{i1}, x_{i2}, x_{i3}),\ i = 1, 2, \ldots, 5$ for five nodes that can preserve the distance matrix Δ. As shown in Fig. 9.9b, all the nodes and their connections preserve the dissimilarities matrix Δ. The matrix D is the pairwise distances between reconstructed nodes in the 3D space.

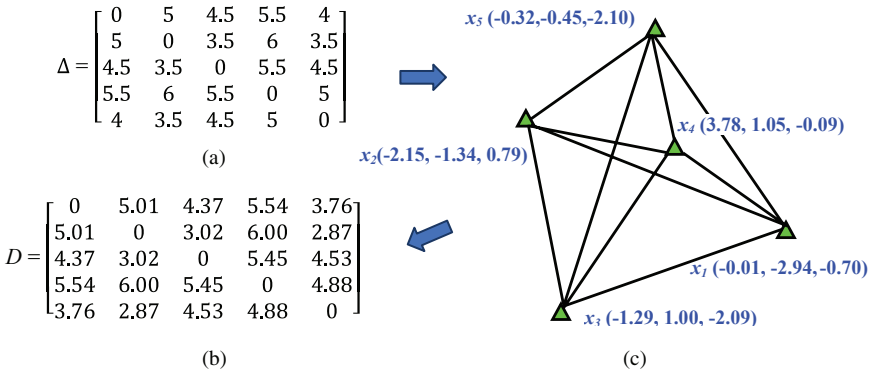


Figure 9.9. (a) Original distance matrix Δ, (b) reconstructed network and nodes in the 3D space, and (c) reconstructed distance matrix D [3].

It may be noted that we are bridging from functional signals to the distance matrix to feature vectors (nodes in the network). The feature vectors will approximately preserve the distance matrix Δ between functional signals.

Let us assume that δ_{ij} denotes the dissimilarity between i^{th} and j^{th} functional recordings in $n \times n$ warping matrix Δ, x_i, and x_j denotes the i^{th} and j^{th} feature vectors in a high-dimensional space. Then, the objective function of feature embedding algorithm can be formulated as follows:

$$\min \sum_{i<j} \left(\left\| \boldsymbol{x}_i - \boldsymbol{x}_j \right\| - \delta_{ij} \right); \; i, j \in [1, n]$$

where $\| \ \|$ is the Euclidean norm. To solve this optimization problem, the Gram Matrix B is firstly reconstructed from the $n \times n$ distance (dissimilarity) matrix Δ:

$$B = -\frac{1}{2} H \Delta^{(2)} H$$

where the centering matrix $H = I - n^{-1} 1 1^T$ and $\mathbf{1}$ is a column vector with n ones. The $\Delta^{(2)}$ is a squared matrix and each element in $\Delta^{(2)}$ is

$\delta_{ij}^{\,2}$ (i.e., the squares of δ_{ij} in the matrix Δ). The element b_{ij} in matrix B is:

$$b_{ij} = -\frac{1}{2}\left[\delta_{ij}^{\,2} - \frac{1}{n}\sum_{k=1}^{n}\delta_{ik}^{\,2} - \frac{1}{n}\sum_{k=1}^{n}\delta_{kj}^{\,2} + \frac{1}{n^2}\sum_{g=1}^{n}\sum_{h=1}^{n}\delta_{gh}^{\,2}\right]$$

It is known that the Gram Matrix B is defined as the scalar product $B = XX^T$, where the matrix X minimizes the aforementioned objective function. The Gram Matrix B can be further decomposed as $B = V\Lambda V^T = V\sqrt{\Lambda}\sqrt{\Lambda}V^T$, where $V = [\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_n]$ is a matrix of eigenvectors and $\Lambda = \text{diag}\,(\lambda_1, \lambda_2, \ldots, \lambda_n)$ is a diagonal matrix of eigenvalues. Then, the matrix of feature vectors is obtained as $X = V\sqrt{\Lambda}$. The algorithm embeds each functional recording as a feature vector in the d-dimensional space ($d = 2, 3, 4, \ldots$).

To this end, a network is optimally constructed in the high-dimensional space. Notably, such network is not static. It is a dynamic network that contains both within-a-patient and between-patient stochastic behaviors. For example, each cycle of the 10-second ECG signal from an individual patient (see Fig. 9.1a) is represented as a node in the network. It may be noted that the node location is changing over time due to the cycle-to-cycle stochastic dynamics. As shown in Fig. 9.10, network nodes are located closely when ECG cycles have similar morphology. However, when there is a significant change, for example, cycle #6, the node moves far away from the previous cycles. Such stochastic network reveals the cycle-to-cycle dynamics and provides physicians useful information pertinent to the underlying changing of cardiac conditions of an individual patient.

Figure 9.11 demonstrates the stochastic network for different patients. Like Fig. 9.10, two nodes are distributed closely when two patients share similar cardiac conditions. The positions of nodes are changing if cardiac conditions vary with respect to time. For example, when patient P1 also gets myocardial infarction symptoms as P3, the corresponding node will move toward P3. As such, physicians are quickly alerted and are able to deliver life-saving therapies in time.

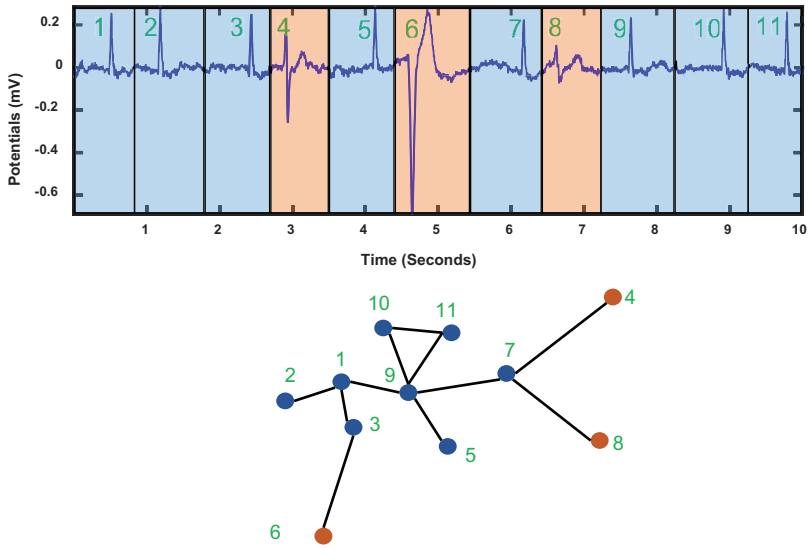


Figure 9.10. Stochastic network for monitoring cycle-to-cycle dynamics of an individual patient.

The proposed stochastic network model can be readily used for online diagnosis. As shown in Fig. 9.12, when a new VCG recording is presented, the pattern dissimilarity will be measured against the database of N patients. Then, a new row and column will be obtained in the warping matrix, and a new feature vector will be embedded in the high-dimensional space. Finally, the classification model will predict cardiac conditions with this feature vector [31].

However, the large number of patients in MESH poses great challenges for real-time analytics and management. On one hand, MESH is aimed at integrating patients all over the world to reduce the risk of cardiac diseases and improve the quality of life. More than 17.5 million people die from cardiac diseases every year, and this number is expected to grow to over 23.6 million by 2030. It is extremely expensive to process millions and billions of patients and provide feedbacks within a reasonably short time. On the other hand, MESH is aimed at long-term monitoring of patients' cardiac

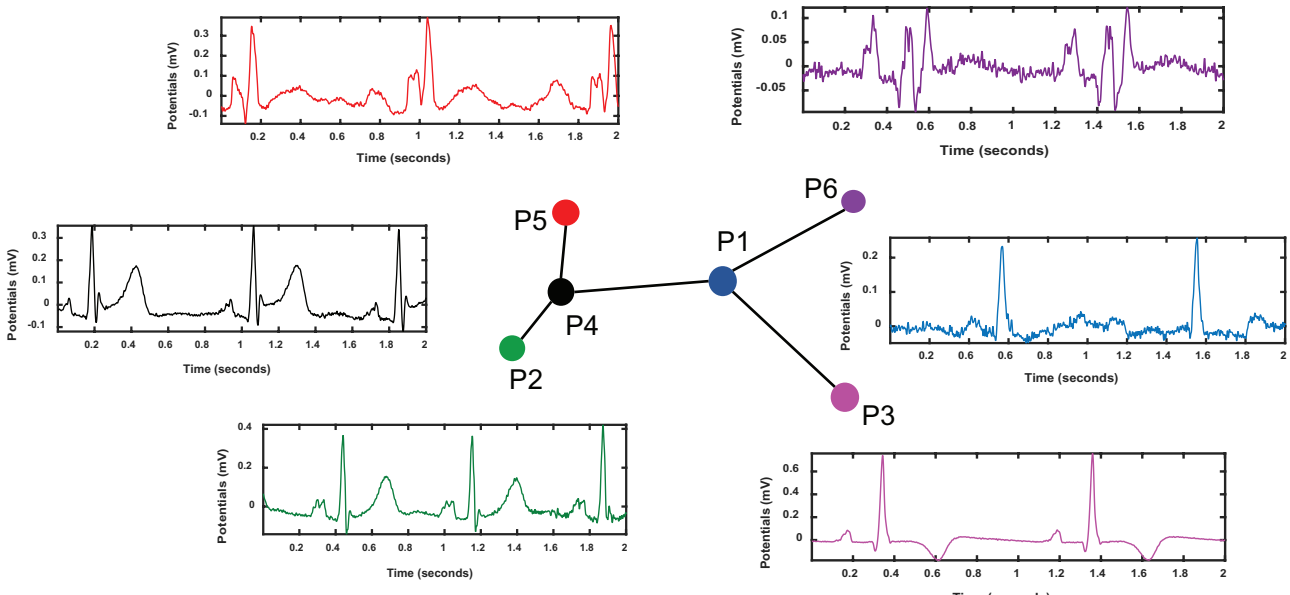


Figure 9.11. Stochastic network for monitoring patient-to-patient dynamics.

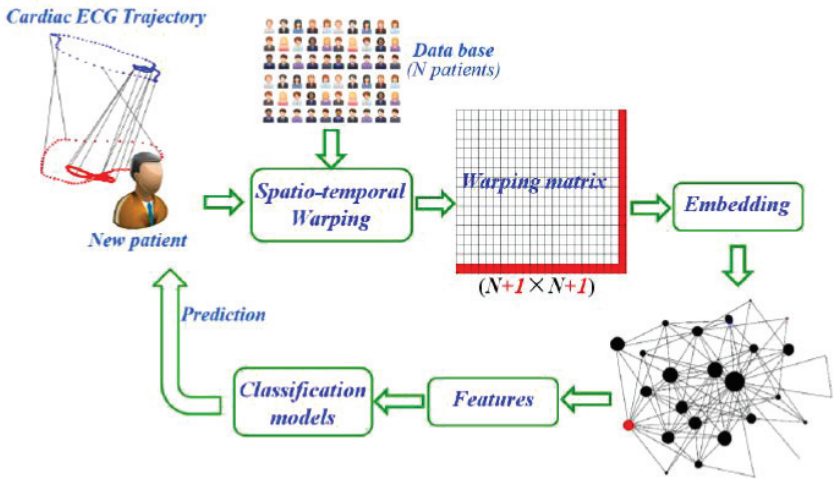


Figure 9.12. The flowchart of stochastic network modeling and online diagnosis.

conditions for personalized cardiac care. Continuous monitoring of an individual patient generates a large amount of data when performed in hours, days, months, and years. There is lack efficient tools to handle such ever-increasing volume of data.

Therefore, we further have developed a new map-reduce framework in MESH for large-scale computing. That is, we have decomposed the large-scale stochastic network optimization problem into local networks and resolved them in a parallel manner [32]. By applying stochastic gradient descent, local networks are optimally casted. Then, the global stochastic network is built by optimally piecing together the local ones. Notably, the proposed strategy facilitates the implementation of parallel computing on a multitude of processors and significantly improve the computation efficiency of the MESH system.

9.4. MESH Design

As shown in Fig. 9.13, the proposed MESH system integrates wearable ECG sensors and mobile computing with network analytics for smart heart health management. The wearable sensing device will

Figure 9.13. The overall framework of the designed MESH prototype [32].

continuously monitor cardiac conditions. Patients will be able to install the *MESH App* onto their smartphones and tablets to register and get connected to the system. After proper authorization, physicians will be able to access patients' data, review results in each analytical module, and communicate with patients and other physicians for timely cardiac care.

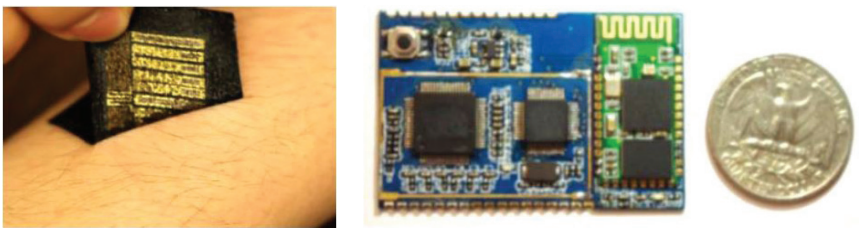
In the past decade, the Internet of Things (IoT) was hailed as a revolution in health care. The IoT system deploys a multitude of wireless sensors, mobile computing units, and physical objects in an Internet-like infrastructure. This provides an unprecedented opportunity to realize a smart automated system that consists of medical devices and analytical modules to advance connected cardiac care. Connected care has been advocated by the Office of the National Coordinator for Health Information Technology for years. As opposed to traditionally isolated care, a highly connected cardiac care system resembles a large-scale network, which seamlessly connects physicians, patients, devices, databases, and other entities. Optimizing the connectivity in cardiac care provides a data-rich environment for medical decision-making, enables smart cardiac telehealth, facilitates personalized patient-centered care, and diminishes care disparities.

However, most of the existing products focus on wearable sensing and fitness applications while being limited in the capability for cardiac sensing and clinical applications. Very little work has been done to develop advanced IoT technologies for smart monitoring and maintain heart health. Therefore, the proposed MESH system is developed to fill this gap. MESH is a new IoT technology specific to the heart, and it is aimed at realizing the next-generation of the cardiac mobile health system (namely the *Internet of Hearts*), proposed by our research group.

9.4.1. *Wearable sensing device*

The existing electrodes are foam-made, fixed-shape, and attached to the skin by electrolyte gel. They do not adhere well to the irregular body surface, thereby, resulting in artifacts during body movement. In this study, we have exploited microdevices assembled on stretchable substrates to develop a new generation of ECG sensors that can stretch, fold, twist, and wrap around the complex surface of the skin. Furthermore, we embedded wireless module (e.g., Bluetooth LE) into the ECG sensor. Thin film circuits of the wireless module were patterned on the soft material so that they can accommodate to large deformations. Moreover, the skin-like substrate architecture quantitatively reproduces mechanics of the non-linear property of the real skin. This, in turn, significantly improved the wearability and facilitate unobtrusive long-term monitoring. As shown in Fig. 9.14a, stretchable sensors have been developed to measure EMG signals in the state of the art [33, 34]. Also, we have developed an ECG sensing board with Bluetooth LE module (Fig. 9.14b) to wirelessly transmit sensing data to mobile devices [13].

Furthermore, the sensor-skin contact can be oftentimes influenced by sweating, motion, among other factors. Thus, the contact is not only static but also dynamic. Notably, the performance of ECG sensors with microelectrodes deteriorates significantly in dynamic contact. As such, the segments of ECG signals or even an entire lead can be missing. In other words, it is not uncommon to



(a) (b)

Figure 9.14. (a) Stretchable bio-sensors [33, 34], (b) Wireless ECG sensing board.

encounter sensor failures in body area sensor networks. For example, a subset of sensors often loses contact with the skin surface in ECG sensor networks because of body movements. Maintaining strict skin contacts for hundreds of sensors is not only challenging but also greatly deteriorates the wearability of ECG sensor networks. Therefore, we have proposed a novel strategy, named stochastic sensor network, which allows a subset of sensors at varying locations within the network to transmit dynamic information intermittently [35]. Notably, the new strategy of stochastic sensor networks is generally applicable in many other domains. For example, a wireless sensor network is often constrained by finite energy resources. Hence, optimal scheduling of activation and inactivation of sensors is imperative to realize long-term survivability and reliability of sensor networks. This information-theoretic approach is integrated with sparse particle filtering to impute missing ECG segments and compensate missing lead(s). In our previous study, we implemented sparse particle filtering for modeling space-time dynamics in an cardiac activity with stochastic sensor networks. The wearable sensing device of MESH will yield an efficient hardware-software solution to ensure the extraction of sufficient diagnostic information from ECG sensor networks.

9.4.2. *MESH database*

An advanced cloud database, that is, MESHDB, is developed to store user data of the proposed MESH system. The cloud platform

optimally allocates the memory among the cluster of servers, which enable nearly unlimited space for storage. At the same time, the MESH system will protect the information stored in the MESHDB. The objective of data management is to specifically focus on optimal management and handling of cyber security issues of cloud database. Notably, the MESH system will only allow the use of MESH app (please refer to Section 9.4.3 for details) and the cloud database from registered users. The users will also be allowed to add notes for each patient and send alert information to the care group. In addition, MESH is designed to connect to ECG data management systems hosted in each hospital. For example, GE MUSE system is a central database that stores all the patients' data and information in the cardiology unit at hospitals. The GE MUSE system provides rich information on cardiology assessments, making administrative workflow and sharing and securing information.

The MESH technology will realize smart and connected cardiac health, once it is available to everyone in the world. It is well-known that the large-scale database is critical to big data analytics, which has the potential to transform the next-generation health care [36]. Big data presents a "gold mine" of this era (21st century). Toward this end, cardiac health care in the future is envisioned to be equipped with the mobile technology, mobile-based data acquisition and cloud database and big data analytics. With new wearable ECG sensing devices, users can directly collect and upload cardiac signals to the MESH system. Each recording will be automatically analyzed by MESH and stored in a cloud database. The more users involved, the bigger the database is, the more powerful the MESH will be. Notably, low-dimensional embedding of a large-scale network can include millions of patients around the world.

Figure 9.15 shows the data flow in the MESH system. Note the arrows indicate the direction of data flows. Primary physicians and care providers in hospitals and home care services have access to their assigned patients in the GE MUSE database hosted by hospitals and home care facilities, as well as in the cloud database hosted by MESH. They can review real-time cardiac recordings for analysis and send back instant feedbacks and care alerts. This will greatly

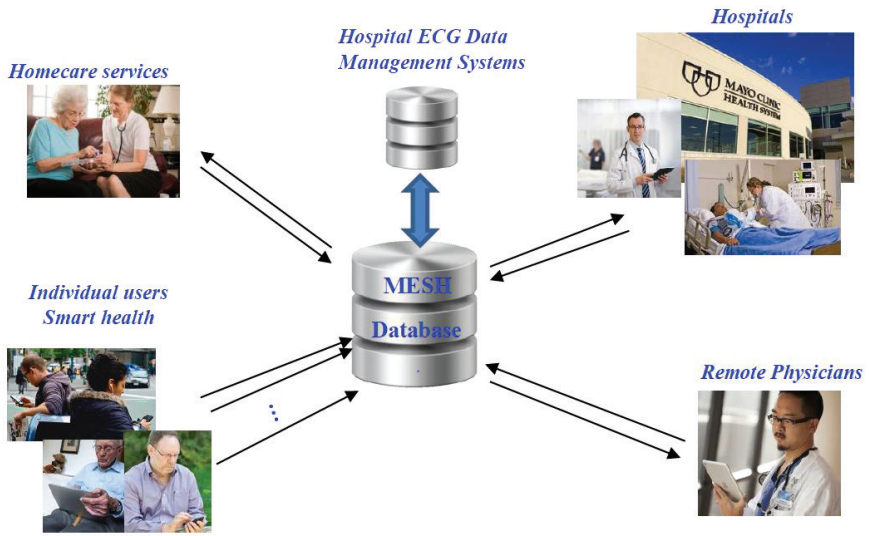


Figure 9.15. Database design of the proposed MESH system.

promote early identification and diagnosis of life-threatening cardiac events (e.g., heart attacks and cardiac arrest). Furthermore, if the patient wants to seek diagnosis results and treatment advice from cardiac experts all around the world, the MESH system can also enable remote physicians to review and analyze the patient's data. In this way, better treatments of cardiovascular diseases can be achieved by teamed efforts from physicians with different background and expertise. Individual users worldwide will be able to monitor their cardiac electric activity in real time, upload data into the cloud database, and consult the physician online. It should be noted that MESH realizes the patient-centered cardiac care anywhere and anytime with the mobile technology and the internet. It is expected that the MESH system will provide an indispensable enabling tool for realizing smart health and wellbeing for the population worldwide.

9.4.3. *MESH smartphone application*

We have developed a mobile application to implement partial functions of the proposed MESH system. This application is developed

in the world's most widely used iOS mobile operating system (which is compatible with iPhone, iPad, and iPod Touch devices). It enables physicians to access the patients' ECG signals in real time, remotely interact with patients, and rapidly respond to life-threatening cardiac disorders.

Screenshots of designed MESH application are shown in Fig. 9.16. Figure 9.16 guides the user through login and patient selection. First, the Login page allows the authorized users to enter their username and password to log into the MESH system. This guarantees the security of the data stored in MESH and protects the privacy of the users. Then, the users such as physicians will be directed to the Sites page that lists hospitals and homecare services. The patients'

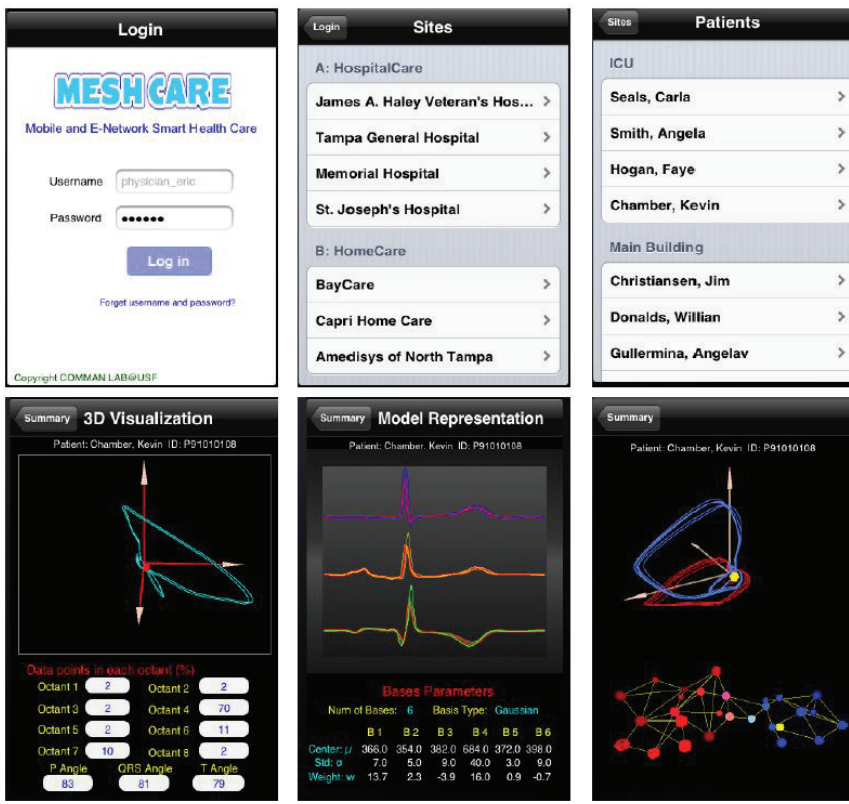


Figure 9.16. Screenshots of designed MESH APP on iPhone.

profiles and data are categorized by the hospital or homecare service. The user can select one site to list his/her assigned patients associated with that site. On the Patients page, all patients associated with the selected healthcare site are listed. Patients are organized by their categories. If a patient is not shown in the list, the doctor needs to go back and select the correct healthcare site. This can be done by clicking on the Sites button on the navigation bar.

Figure 9.16d–f demonstrate three major functions of the MESH system, that is, dynamic visualization of space-time VCG signals, optimal model-based representation, and stochastic network analytics. On "3D visualization" page, dynamic space-time VCG signals are displayed on the upper panel. The red point gives the current position of the cardiac vector. The cyan loops record the full history pertinent to complete the topological shape of the VCG state space. The plot is automatically rotating counter-clockwise on the z-axis. The rotation facilitates a 360° view of spatiotemporal signals. Spatiotemporal features are updated in real time in the lower panel, including the percentage of data points in each of the eight octants, and the angle of P, QRS, and T axis.

On "Model Representation" page, multiple cycles are collected from each of the three VCG channels and displayed on the upper panel (blue → X channel, yellow → Y channel, and green → Z channel). The red curves (with large line width) are the basis function models obtained from the summation of six adaptive Gaussian functions. It is noteworthy that the models effectively capture the morphology of signals. The parameters of basis functions, including center, standard deviation, and weight, are listed in the lower panel for basis 1 (B1) to basis 6 (B6).

On the last page, that is, "dynamic network analytics", 3D visualization of VCG loops are shown in the upper panel. The blue trajectory is from a normal subject, and the red trajectory is from myocardial infarction. The yellow indicator moving along the VCG cycles represents the current cycle we are looking at. The plot is automatically rotating counter-clockwise on the z-axis, providing a 360° view of spatiotemporal cardiac patterns. The embedded network is displayed on the lower panel. Nodes are the patients in the database: red nodes are myocardial infarction patients and blue

nodes are healthy subjects. The yellow node in the network indicates the current position of the patient (e.g., Kevin Chamber in this screenshot). When the yellow indicator in the upper panel is moving along the blue cycles, the yellow node on the lower panel is within the group of healthy subjects (i.e., blue nodes). However, when the yellow indicator in the upper panel moves into the red cycles, the yellow node in the network is switched to the cluster of myocardial infarction patients (i.e., red nodes).

9.5. Discussion

The developed MESH system is aimed at a large market for patient-centered cardiac care. In 2013, more than 83.3 million American adults (>1 in 3) had heart diseases. The increasing prevalence of cardiac disease calls for smarter cardiac care services. The growing presence of smartphones and tablets provides an unprecedented opportunity to advance cardiac telemedicine and realize the smart cardiac care anytime anywhere, which is not only responsive but also cost effective.

In the NSF I-Corps program, which aimed at developing entrepreneurial skills to translate research results from academic laboratories, we did an extensive marketing research regarding the developed MESH system. We interviewed over 100 cardiac patients, physicians, and cardiac nurses; identified unprecedented marketing opportunities; and found the following:

(1) There is a lack of wireless sensing devices for continuous monitoring of multi-channel ECG signals. Existing products focus on portable cardiac monitors, which can only monitor a single-channel ECG and are limited in their ability to facilitate the diagnosis of complex cardiac disorders in the clinical practice. Furthermore, most of the existing monitors adopt dry electrodes. It is uncomfortable to take daily activity with them, and they may result in skin irritation. The proposed MESH system is not only able to record hospital-grade multi-lead ECG, but also comfortable, flexible, and reliable to facilitate long-term continuous monitoring.

(2) Currently, there is a great shortage of physicians in the United States, and this situation will worsen in the next decade. Patients with acute cardiac disorders need 24/7 monitoring, but physicians cannot stay in hospitals or with the patients all the time. Currently, when doctors are outside hospitals, they ask nurses to take pictures of ECG signals and send them through the phone. This is apparently not an efficient approach because certain delays are unavoidable, and the resolution of pictures is limited. Equipped with advanced cloud database, the proposed MESH system can be ready to help physicians access patients' data anywhere and anytime to give a timely diagnosis and medical intervention.

(3) There is a lack of enabling tools to extract useful information from big data that is generated from continuous cardiac monitoring. Early identification of disease patterns hinges on information-processing and data mining algorithms. The existing devices are only capable of extracting simple ECG characteristics or transferring data to physicians for visual inspection. MESH innovatively adopts stochastic network analytics for disease pattern recognition. Unlike traditional warping that can only be used to align signals in time domain, the proposed method is able to quantify the space-time dissimilarities between 3D trajectories of cardiac signals. One remarkable feature of the MESH system is that it considers both within-a-patient and between-patient stochastic dynamics for network-based pattern recognition of cardiac diseases. This will assist and enable physicians in the decision-making process.

9.6. Summary

Cardiovascular diseases are the leading cause of death around the world. According to WHO, cardiac diseases contribute to more than 30% of the global deaths each year. Optimal management and treatment of cardiac diseases hinge on the development of advanced cardiac telemedicine system for the detection of fatal disease patterns in

the early stage and timely delivery of life-saving therapies. However, the cardiac electrical activity manifests significant stochastic properties in both space and time. The existing approaches are either not concerned with underlying changes of cardiac conditions for an individual patient or not capable to effectively differentiate different cardiac conditions among patients. There is an urgent need to fully address underlying stochastic properties and uncertainties in the cardiac electrical activity.

This chapter presents new visualization and data analytics tools for stochastic modeling and analysis of cardiac electrical signals, which advance cardiac telehealth-care service with exceptional features such as personalization, responsiveness, and superior quality. Specifically, we first developed a spatiotemporal approach to capture space-time heart dynamics by displaying the real-time motion of 3D VCG cardiac vectors. Then, an optimal model-based representation algorithm was developed to facilitate the compression of cardiac signals and the extraction of features pertinent to the disease-altered cardiac activity. Then, a stochastic network model was designed for real-time patient-centered monitoring, modeling, and analysis of cardiac variations. Finally, we leveraged the developed algorithms and built the next-generation cardiac mHealth system, MESH.

MESH bridges gaps in the current cardiac telemedicine systems and serves as an enabling tool to reduce the risk of life-threatening cardiac disorders and deliver personalized therapies.

We expect that this chapter will spur further investigations in stochastic modeling and analysis of spatiotemporal ECG signals to accelerate the discovery of knowledge in cardiovascular research.

Acknowledgment

This work is supported in part by the National Science Foundation (CMMI-1646660, CMMI-1617148, CMMI-1619648, IIP-1447289, and IOS-1146882). The authors also thank Harold and Inge Marcus Career Professorship (HY) for additional financial support.

References

1. Roger, VL, AS Go, DM Lloyd-Jones, EJ Benjamin, JD Berry, WB Borden, DM Bravata, S Dai, ES Ford, CS Fox, HJ Fullerton, C Gillespie, SM Hailpern, JA Heit, VJ Howard, BM Kissela, SJ Kittner, DT Lackland, JH Lichtman, LD Lisabeth, DM Makuc, GM Marcus, A Marelli, DB Matchar, CS Moy, D Mozaffarian, ME Mussolino, G Nichol, NP Paynter, EZ Soliman, PD Sorlie, N Sotoodehnia, TN Turan, SS Virani, ND Wong, D Woo and MB Turner (2012). Heart disease and stroke statistics — 2012 update. *Circulation,* 125(1), e2–e220.
2. Jaffe, A, L Babuin and F Apple (2006). Biomarkers in acute cardiac disease. *J. Am Coll Cardiol.,* 48(1), 1–11.
3. Yang, H, C Kan, G Liu and Y Chen (2013). Spatiotemporal differentiation of myocardial infarctions. *IEEE Trans Autom Sci Eng.,* 10(4), 938–947.
4. Chen, Y and H Yang (2012). Self-organized neural network for the quality control of 12-lead ECG signals. *Physiol. Meas.,* 33(9), 1399.
5. Chen, Y and H Yang (2013). A comparative analysis of alternative approaches for quantifying nonlinear dynamics in cardiovascular system. In *2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2013, 2599–2602.
6. Yang, H (2011). Multiscale recurrence quantification analysis of spatial cardiac vectorcardiogram signals. *IEEE Trans Biomed Eng.,* 58(2), 339–347.
7. Chen, Y and H Yang (2013). Wavelet packet analysis of disease-altered recurrence dynamics in the long-term spatiotemporal vectorcardiogram (VCG) signals. In *2013 35th Annual International Conference of the IEEE Eng Med Biol Soc.,* 2013, 2595–2598.
8. Chen, Y and H Yang (2012). Multiscale recurrence analysis of long-term nonlinear and nonstationary time series. *Chaos Solitons Fract.,* 45(7), 978–987, 7.
9. Qu, Z, G Hu, A Garfinkel and J Weiss (2014). Nonlinear and stochastic dynamics in the heart. *Physics Reports,* 543(2), 61–162.
10. Lerma, C, T Krogh-Madsen, M Guevara and L Glass (2007). Stochastic aspects of cardiac arhythmias. *J Stat Phys.,* 128, 347–374.
11. Du, D, H Yang, SA Norring and ES Bennett (2014). In-silico modeling of glycosylation modulation dynamics in hERG ion channels and

cardiac electrical signals. *IEEE J Biomed Health Informat.,* 18(1), 205–214.
12. Du, D, H Yang, H Yang H, AR Ednie and ES Bennett (2015). Statistical metamodeling and sequential design of computer experiments to model glyco-altered gating of sodium channels in cardiac myocytes. *IEEE J Biomed Health Informat.,* PP(99), 1–12.
13. Kan, C, Y Chen, F Leonelli and H Yang (2015). Mobile sensing and network analytics for realizing smart automated systems towards health internet of things. In *2015 IEEE International Conference on Automation Science and Engineering* (*CASE*), pp. 1072–1077. Gothenburg: IEEE Conference Publications. [24–28 August 2015].
14. Malmivuo, J and R Plonsey (1995). *Bioelectromagnetism: Principles and Applications of Bioelectric and Biomagnetic Fields*. New York: Oxford University Press.
15. Dale, D (2000). *Rapid Interpretation of EKG's: An Interactive Course*. Tampa, FL: Cover Publishing Company.
16. Liu, G, C Kan, Y Chen and H Yang (2014). Model-driven parametric monitoring of high-dimensional nonlinear functional profiles. In *2014 IEEE International Conference on Automation Science and Engineering* (*CASE*), pp. 722–727. Taipei: IEEE Conference Publications. [18–22 August 2014].
17. Yang, H, STS Bukkapatnam and R Komanduri (2007). Nonlinear adaptive wavelet analysis of electrocardiogram signals. *Physical Review E,* 76, 026214.
18. Bukkapatnam, S, R Komanduri, V Yang, P Rao, WC Lih, M Malshe, LM Raff, B Benjamin and M Rockley (2008). Classification of atrial fibrillation (AF) episodes from sparse electrocardiogram (ECG) datasets. *J Electrocardiol.,* 41(4), 292–299.
19. Yang, H, STS Bukkapatnam, T Le and R Komanduri (2011). Identification of myocardial infarction (MI) using spatio-temporal heart dynamics. *Med. Eng. Phys.,* 34(4), 485–497.
20. Dower, GE, A Yakush, SB Nazzal, RV Jutzy and CE Ruiz (1988). Deriving the 12-lead electrocardiogram from four (EASI) electrodes. *J Electrocardiol.,* 21(1), S182–S187.
21. Dower, GE and HB Machado (1979). XYZ data interpreted by a 12-lead computer program using the derived electrocardiogram. *J Electrocardiol.,* 12(3), 249–261.
22. Dawson, D, H Yang, M Malshe, STS Bukkapatnam, B Benjamin and R Komanduri (2009). Linear affine transformations between 3-lead

(Frank XYZ leads) vectorcardiogram and 12-lead electrocardiogram signals. *J Electrocardiol.,* 42(6), 622–630.
23. Yang, H, STS Bukkapatnam and R Komanduri (2012). Spatiotemporal representation of cardiac vectorcardiogram (VCG) signals. *Biomed Eng Online,* 11(12), 16.
24. Liu, G and H Yang (2013). Multiscale adaptive basis function modeling of spatiotemporal vectorcardiogram signals. *IEEE J Biomed Health Informatics,* 17(2), 484–492.
25. Mallat, SG and Z Zhang (1993). Matching pursuits with time-frequency dictionaries. *IEEE Trans Signal Process.,* 41(12), 3397–3415.
26. Jeong, Y, MK Jeong and OA Omitaomu (2011). Weighted dynamic time warping for time series classification. *Pattern Recognition,* 44, 2231–2240.
27. Myers, C, L Rabiner and A Rosenberg (1980). Performance tradeoffs in dynamic time warping algorithms for isolated word recognition. *IEEE Trans Acoustics Speech Signal Process.,* 28(6), 623–635.
28. Kan, C and H Yang (2012). Dynamic spatiotemporal warping for detection and location of myocardial infarctions. In *The Eighth Annual IEEE International Conference on Automation Science and Engineering* (*CASE 2012*), pp. 1046–1051. Seoul, Korea: IEEE Conference Publications. [20–24 August 2012].
29. Yang, H (2013). Systems and methods for determining a cardiovascular condition of a subject. US Patent No. 14036776.
30. Yang, H (2013). Systems and methods for diagnosing cardiovascular conditions. International PCT patent No. 61700575.
31. Yang, H and F Leonelli (2016). Self-organizing visualization and pattern matching of vectorcardiographic QRS waveforms. *Comput Biol Med.,* 79, 1–9.
32. Kan, C, FM Leonelli and H Yang (2016). Map reduce for optimizing a large-scale dynamic network — Internet of hearts. In *Proceedings of 2016 IEEE Engineering in Medicine and Biology Society Conference* (*EMBC*), pp. 1–4. Orlando, FL: IEEE Conference Publications.
33. Cheng, H, Y Zhang, K Hwang, J Rogers and Y Huang (2014). Buckling of a stiff thin film on a pre-strained bi-layer substrate. *Int J Solids Structures,* 51(18), 3113–3118.
34. Jang, K, H Chung, S Xu, CH Lee, H Luan, J Jeong, H Cheng, GT Kim, SY Han, JW Lee, J Kim, M Cho, F Miao, Y Yang, HN Jung, M Flavin, H Liu, GW Kong, KJ Yu, SI Rhee, J Chung, B Kim, JW Kwak,

MH Yun, JY Kim, YM Song, U Paik, Y Zhang, Y Huang and JA Rogers (2015). Soft network composite materials with deterministic and bio-inspired designs. *Nature Communications,* 6, 6566.

35. Chen, Y and H Yang (2016). Sparse modeling and recursive prediction of space–time dynamics in stochastic sensor networks. *IEEE Trans Automation Sci Eng.,* 13(1), 215–226.
36. Yang, H and E Kundakcioglu (2014). Healthcare intelligence: Turning data into knowledge. *IEEE Intelligent Syst.,* 29(3), 54–68.

10. Using Agent-Based Interpersonal Influence Simulation to Study the Formation of Public Opinion

Yu Teng*,†, Nan Kong‡ and Torsten Reimer§

*Center of Modeling, Planning and Policy Analysis, Avenir Health, Glastonbury, Connecticut, USA
†Weldon School of Biomedical Engineering, Purdue University, West Lafayette, Indiana, USA
‡Weldon School of Biomedical Engineering, Purdue University, West Lafayette, Indiana, USA
§Brian Lamb School of Communication and Department of Psychological Sciences, Purdue University, West Lafayette, Indiana, USA

Abstract

In the past decade, computational social sciences have become a vibrant research area, partially due to rapidly advanced computer simulation tools, such as agent-based social simulation. At present, we can simulate complex social systems in detail. One important field of study looks at social dynamics that incorporates differences

among individual agents and adaptive choice behavior. In this chapter, we showcase a recent study on an agent-based social influence simulation that aimed to investigate the change of individual attitudes and the formation of public opinions over time through scale-free networks. This simulation study is expected to help facilitate the ongoing integration of systems science and behavioral and social sciences, which is of tremendous value to tackling healthcare challenges.

10.1. Introduction and Background

Systems science methodologies, such as microsimulation, system dynamic modeling, agent-based modeling, social network analysis, discrete-event analysis, and Markov modeling, have been increasingly applied in the past decades to help understand complex dynamical behavioral and social science processes relevant to health problems. The application areas include sociology [1], economics [2], social psychology [3–4], and anthropology [5]. These modeling methodologies are mainly used to understand connections between a system's structure and its collective behavior over time. Together with other novel analytic tools, they can illuminate complex and interconnected pathways between the social, economic, and environmental causes of poor health. These tools can be used to inform and support policy-making and decisions on resource allocation in health care.

Among methodologies in computational social sciences, agent-based modeling (ABM) has gained increasing popularity in recent years as it can capture population-level inference from explicitly programmed, micro-level rules over time and space [6]. ABM builds social structures within a simulated population with the use of a "bottom-up" approach to investigate social and organizational phenomena [7–9]. With ABM, a complex social system is modeled as a collection of autonomous decision-making entities called agents. ABM enables the investigation of systems in which (1) individual behavior is nonlinear and can be characterized by thresholds, if-then rules, or nonlinear coupling; (2) individual behavior exhibits memory,

path-dependence, and hysteresis, non-Markovian behavior, or temporal correlations, including learning and adaptation; and (3) agent interactions are heterogeneous and can generate network effects. Typically, in the execution of an ABM, each agent assesses its own situation, makes its rule-based decisions to optimize some payoff, and behaves accordingly. With an update on each agent's situation, the system evolves according to the collective update. Repetitive (and often competitive) interactions between agents are a key feature of ABM, which relies on the power of computers to explore dynamics out of the reach of pure mathematical models. ABM has been applied to investigate complex human systems pertaining to sociology [1], stock market [10], epidemics [11], and ancient civilization [12].

In this chapter, we present a study that applies ABM to investigate (1) how individual attitudes toward some contentious issue change over time through social influence; (2) how and when public opinion may be formed as some collection of individual attitudes; and (3) how the formation of public opinion is affected by social networking characteristics and the decision rules taken by the individuals in the network. We have thereby modeled individual agents within a social network and simulate social influence processes at the population level with the bottom-up approach. We have integrated two dimensions into an ABM: the structure of social networks and psychologically plausible models of individual decision-making. Our study rests on two main categories of assumptions: assumptions about individual decision-making and public opinion formation that are motivated through psychological models and assumptions about opinion diffusion and network structure. Based on these assumptions, we design the dynamic model for public opinion and apply agent-based simulation to investigate the dynamics numerically and its relationship with the model parameters.

When the public faces contentious issues (e.g., the measles outbreak and vaccine controversy; see DeStefano and Chen [13]), individuals hold diverse opinions at the beginning. However, the public opinion, that is, the opinions of the majority of members and significant minorities, is shaped over time through dynamic influence

processes in which the opinions of individual members are formed. These dynamic processes may have different outcomes: opinions may polarize, thus, yielding two large factions or, in other situations, a majority may form fast without strong fractionalization. To be able to predict better and explain the outcomes of the formation of public opinions, it is important to characterize and model public opinion dynamics. For general introductions to research on social dynamics, see Friedkin [14], Hegselmann and Krause [15], and Janssen and Jager [16]. For a review on computational models of public opinions and collective behaviors, we refer to Goldstone and Jassen [17].

Earlier studies in this area found that public opinion on contentious issues depends on the way autonomous but interdependent individuals process information and make decisions on the issues [18, 19]. Further studies identified several important individual factors that impact public opinion and its formation. First, public opinions depend on the distribution of initial opinions held by individuals in the system [20]. Second, public opinion is influenced by the formation of simple heuristics used by individuals [21–23] and the effectiveness of communication approaches [18, 24, 25]. In addition, public opinion is affected by the structure of the social network and associated environmental factors, including the availability of authoritative figures on the issues [26, 27]. Our study builds on this research by proposing an ABM-based study platform. Our study confirms that public opinion is sensitive to initial composition of individual opinions and collective opinions are affected by how individuals seek and process information in the social network.

In the remainder of the chapter, we will first review the literature on applying computational modeling to study social networks, with emphasis on the modeling of the dynamics of social contagion and interpersonal influence processes. Next, we will present the conceptual design of our ABM and discuss its implementation. We will report several simulation experiments and provide policy insights based on the simulation results. At the end, we will draw conclusions and outline our future research.

10.2. Computational Models of Social Contagion and Influence

There have been many studies on computational modeling of social contagion and influence processes. For literature reviews, we refer to Smith and Christakis [28] and Christakis and Fowler [29, 30]. In general, modeling and characterization of social networks have been a vibrant research area for more than two decades. For example, systems scientists, back in the 1990s, appreciated the limit on describing systems composed of nonidentical elements, which had diverse and nonlocal interactions [31, 32]. Such limit hindered the advances in many disciplines, ranging from molecular biology [33, 34] to computer science [35–37]. The limit laid partly in the topology of the systems, as many of them form rather complex networks whose vertices are the elements of the system and whose edges represent the interactions between them. Complex networks also occurred in social science. Social networks typically have vertices representing individuals and organizations and edges representing the social interactions or connections between them [38]. Social networks can vary from small interacting groups (e.g., in a workgroup or students in a classroom) to large-scale communities or societies, for which the network topology is largely unknown. What is even less known, but perhaps more important, is the dynamical and topological stability of the networks.

Traditionally, networks of complex topology have been described with the so-called Erdös–Rényi (ER) theory [39] in the literature of random graphs [40]. However, in the absence of data on large complex networks, the predictions of the ER theory were rarely tested in the real world. Driven by the computerization of data acquisition, such topological information is increasingly available, raising the possibility of understanding these networks. One noteworthy piece of the work in this area is Barbasi and Albert [41]. The authors proposed a network generation algorithm that exploited a common feature inherent in many large networks, namely that the vertex connectivity followed a scale-free power-law distribution. A network model based on this feature reproduced the observed stationary scale-free distributions,

which allowed the social science research community to conduct model-based large complex network topology study more efficiently.

This line of research became systematic around 2002 when Drs. Nicholas A. Christakis, James H. Fowler, and their colleagues explored previously unused paper records held by the Framingham Heart Study (FHS) [42, 43], a longitudinal epidemiological cohort study, to reconstruct social network ties among 12,067 individuals over 32 years. An uncommon feature found in the data was that numerous attributes of each individual in the network were longitudinally observed. In 2007, Christakis, Fowler, and their colleagues began to model social networks computationally (i.e., network topology identification and characterization), using the FHS dataset [44–48], the National Longitudinal Study of Adolescent Health (AddHealth, a public-use dataset with social network information on 90,000 children in 114 schools) [49], online social network data [50, 51], and *de novo* data they extracted [52, 53]. The researchers also examined various network phenomena with these experimental data [54, 55].

Undoubtedly, the FHS data and others offered important insights and opportunities for the study of social networks. As the research in this area deepened, the relevant investigation was divided into two categories: studies of network topology (and its determinants) and studies of the spread of phenomena across network ties. The former categories include Fowler *et al.* [56], O'Malley and Christakis [57], Christakis *et al.* [58], Onnela and Christakis [59], Onnela *et al.* [60], and Fowler *et al.* [61]. In this social-network category, the line of work identified the influence of several health determinants on the system outcomes, including genetic makeup and other health traits. This line of work also investigated the human connectivity in real-world social networks, based on empirical data.

In this chapter, we are focusing on the latter category, which includes analyses of the flows of behaviors and affective states. The work on social influence covers several application domains and relies on diverse data and modeling approaches. Data sources explored in this area include influenza [52], obesity [62, 63], smoking [44], alcohol consumption [45], health screening [64],

happiness [46], loneliness [47], depression [48], drug use [65], and food consumption [66]. Experiments that have been conducted in this area include the analysis of the development and dynamics of public opinions. In general, this line of work builds on prior research on "peer effects" and interpersonal influence by examining data from networks containing large cohorts. In this area, we have witnessed the fast growth of network statistics, which provides viable methodological options. For useful reviews, we refer to Wasserman and Faust [38], Jackson [67], Goyal [68], O'Malley and Marsden [69], Newman [70, 71], Easley and Kleinberg [72], and Kolaczyk [73]. Each of these methodologies focuses on specific aspects of networks and is, thus, suited for specific situations. There is no generic methodology that will best answer every question one may want to ask with observational or experimental data. Furthermore, the research community needs to address challenging issues including the treatment of missing data (such as missing nodes, ties, covariates, and waves), sampling issues (design effects and incomplete network ascertainment), the computation of standard errors, and the interpretation of model parameters. A strength of agent-based simulations consists in the computational flexibility of modeling social network dynamics. However, the computationally efficient procedures and guidelines by which these methods are applied for network dynamics modeling and characterization remain hot research topics.

In the following section, we showcase an agent-based social influence simulation study on public opinion dynamics in scale-free networks.

10.3. An ABM Approach

The agent-based social influence simulation model aims to analyze how each individual's opinion regarding certain contentious matter changes over time in a social network and whether these changes lead to the formation of public opinion on the matter. Figure 10.1 presents a schematic overview of the developed agent-based simulation model.

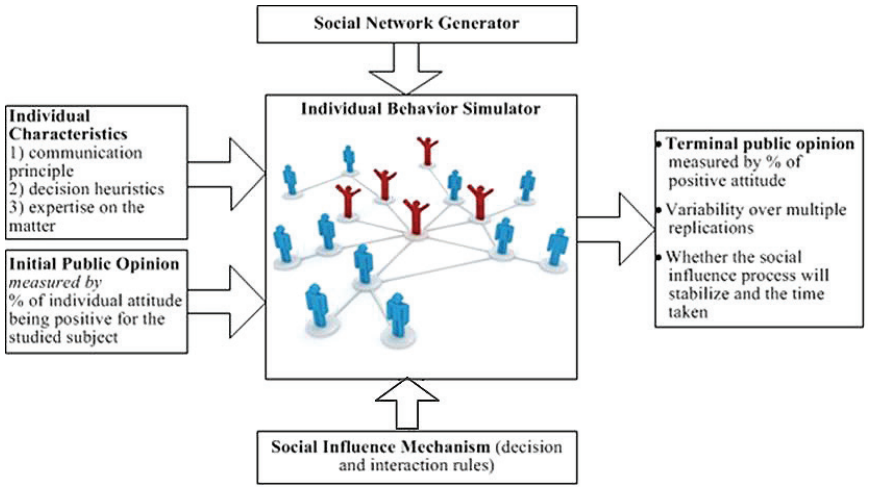


Figure 10.1. Schematic overview of the agent-based simulation model.

In the simulation, we considered a scale-free network, with nodes representing individuals and edges indicating connections between individuals in the network. Individuals that are directly connected are *neighbors*. Each individual in the network is assumed to be in contact with its neighbors in the network according to some criterion, termed *contact criterion*. This criterion defines with whom the individual communicates about the risk of the contentious matter. Influenced by the communication, each agent updates his/her opinion according to some *decision rule*, which reflects how the agent synthesizes the opinions of their neighbors.

Moreover, we assigned an expertise index to each agent in the social network to indicate its expertise on the subject. Individuals often adapt opinions of others who are assigned a high expertise [74, 75]. We used AnyLogic simulation software (*www.anylogic.com*) for the implementation. AnyLogic supports a decentralized, individual-centric approach to modeling. It also provides a visual language to simplify the development of ABMs. The network used to simulate social influence processes was a randomly generated scale-free network of 200 agents, with each one of them connected to seven neighbors on an average.

We used the Barabási–Albert (BA) algorithm [41] to generate the network. Scale-free networks are networks that have a power-law distribution of number of links connecting to an agent, that is, a majority of agents have less-than-average links, while a small fraction of agents are connected to many other agents. Many networks in the real world are conjectured to be scale-free, including social networks and many kinds of computer networks.

We analyzed the impact of individual decision rules and contact criteria on the social influence dynamics. Table 10.1 summarizes the decision rules and contact criteria used in this study. These rules and criteria describe psychologically viable information processing and social influence mechanisms that have been observed across a variety of different groups and social settings (see e.g., Schwenk and Reimer [23]).

Table 10.1. A summary of the social influence parameters.

Decision Rule	*Simple Majority* (SM)	An agent takes the opinion of the majority of its neighbors.
	Weighted Majority (WM)	An agent weights and integrates the opinions of its neighbors based on their expertise index.
	Follow the Leader (FL)	An agent takes the opinion of its neighbor with the highest expertise index.
Contact Criterion	*All Neighbors* (AN)	An agent seeks opinion from all its neighbors.
	More Expertise (ME)	An agent seeks opinion only from those who have higher expertise indices.
Initial Opinion Distribution	*Uncorrelated* (UC)	The initial opinion of each agent is uncorrelated with its expertise.
	Positively Correlated (PC)	The initial opinion of each agent is positively correlated with its expertise index.

We assigned identical decision rules and contact criteria to all agents. As an additional dimension that has been shown to influence the effectiveness of social influence mechanisms, we systematically varied the assumed expertise of agents (e.g., see the follow-the-expert rule, Reimer and Hoffrage [25]). We randomly selected a subset of agents to be "experts" and assigned their expertise indices, which were higher than the rest of the population. Further, we considered two scenarios that specified whether each agent's initial opinion was correlated to its assigned expertise. The first scenario, termed uncorrelated (UC), assumed that each agent's initial opinion is uncorrelated to its expertise index. The second scenario, termed positively correlated (PC), assumed that each agent's initial opinion is positively correlated with its expertise index. The second scenario was motivated by the assumption that expert opinions are often not independent from each. Often, the opinions of experts are to some extent correlated when one course of action is more appropriate or effective than another course of action [22].

To initialize the simulation, the initial opinion for each agent was generated. At each tick (or generation) in the simulation, all agents were activated and updated their state based on a randomly generated sequence. When an agent was activated, it could take one of the two actions: maintaining its opinion or changing it to the opposite. The action taken was based on the interaction with other agents (as described in the contact criterion) and how the agent formed its opinion (as described in the decision rule). We terminated the simulation when the percentage of the population holding positive opinions oscillated within a threshold or the maximum number of the generation was reached. After each simulation replication, we recorded the terminal percentages of the population that held positive opinion and the number of generations. See Fig. 10.2 for a snapshot illustration of the simulation execution. For detailed description of the system dynamics induced by the contact criteria and decision rules as well as the initial individual opinion distribution, we refer to the Appendix.

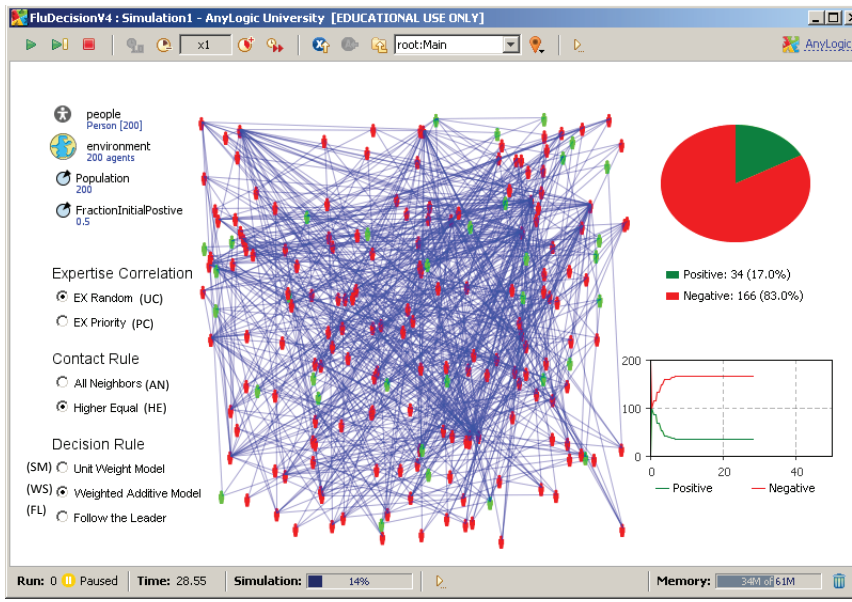


Figure 10.2. A snapshot of the simulation execution.

10.4. Simulation Results and Discussions

With the agent-based simulation developed, we investigated the association between the initial individual opinions and the terminal individual opinions, from which we concluded whether stable public opinions could be formed in a reasonable amount of time. Furthermore, we analyzed the impact of alternative decision rules and contact criteria on the social influence dynamics, as well as effects of the correlation between agent's initial opinions and their expertise on the subject.

In our numerical studies, there were nine sets of simulation experiments (see Table 10.2). Note that once the agents took the decision rule "Follow the Leader," there was no difference between the two contact criteria they might apply since each agent identified the same leader from its two different social networks. Hence, we combined UC_FL_AN and UC_FL_ME, and PC_FL_AN and PC_FL_ME.

Table 10.2. A summary of the test cases.

Case Label	Initial Opinion Distribution	Decision Rule	Contact Criteria	Index in Figures
SM_AN	Uncorrelated with the expertise index (UC)	Simple Majority (SM)	All Neighbors (AN)	(a)
UC_SM_ME			More Expertise (ME)	(d)
UC_WM_AN		Weighted Majority (WM)	All Neighbors (AN)	(b)
UC_WM_ME			More Expertise (ME)	(f)
UC_FL		Follow the Leader (FL)	—	(h)
SM_AN	Positively Correlated with the expertise index (PC)	Simple Majority (SM)	All Neighbors (AN)	—
PC_SM_ME			More Expertise (ME)	(e)
PC_WM_AN		Weighted Majority (WM)	All Neighbors (AN)	(c)
PC_WM_ME			More Expertise (ME)	(g)
PC_FL		Follow the Leader (FL)	—	(i)

In addition, regardless of how initial opinion distribution was correlated with the expertise index, each agent's social network remained roughly the same for the homogeneity among the agents. Meanwhile, when "Simple Majority" decision rule and "All Neighbors" contact criterion were applied, the expertise index was essentially not used in the social simulation dynamics. As a result,

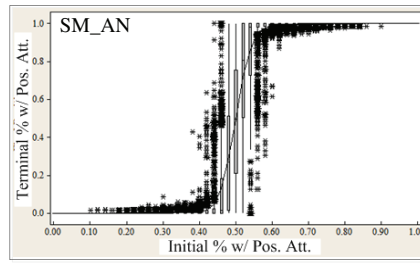
the two cases UC_SM_AN and PC_SM_AN achieved the same results, with sufficiently many simulation runs. We, thus, combined the two cases and simply called them both SM_AN.

In each of the 10 experiment sets, we varied the percentage of agents in the network that held a positive initial opinion in favor of the subject or contentious action, from 0% to 100% (with 2% increment). For example, if this percentage was 60%, 60% of agents in the network had a positive opinion regarding the contentious matter at the outset and 40% held a negative opinion. To terminate a simulation run, we either stopped the simulation when the system reached relative stable state (i.e., percentage oscillation was within 2% of agents, that is, the number of members with positive opinions stayed within a range of +4 and −4 for 10 consecutive time units) or after sufficiently long simulation duration (i.e., 1,000 time units) even if the system state continued to oscillate noticeably. For each experimental specification, we ran the simulation for 1,000 replications.

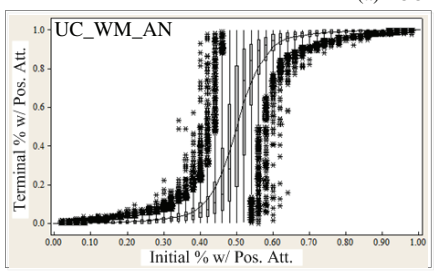
10.4.1. *Simulation results*

Figures 10.3 and 10.4 present simulation results. Each figure contains nine subfigures. In each subfigure, the x-axis represents the percentage of agents holding positive opinion on the studied subject at the beginning of the simulation, and the y-axis represents the percentage of agents holding positive opinion at the termination. Each subfigure of Fig. 10.3 presents a box plot for each given initial condition on the x-axis (i.e., percentage of positive opinion). In addition, Fig. 10.3 shows the Pearson product-moment correlation coefficient to measure the linear correlation between the initial percentage of agents holding positive opinion and average terminal percentage over the 1,000 replications.

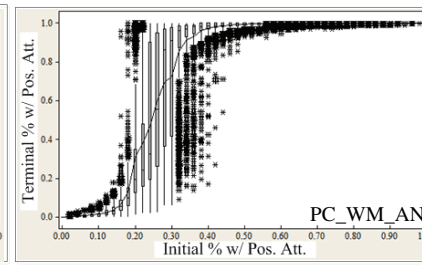
Each subfigure of Fig. 10.4 either reports the time (number of time units or ticks in AnyLogic) that it took for the public opinion in the social network to stabilize or indicates that the public opinion still oscillated even after a sufficiently long time (i.e., 200-unit time intervals).



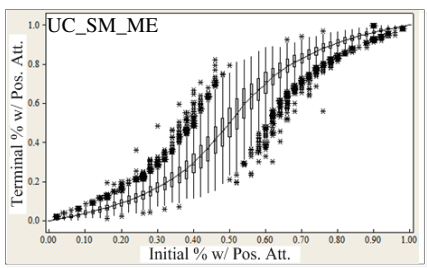
(a) PCC = 0.898



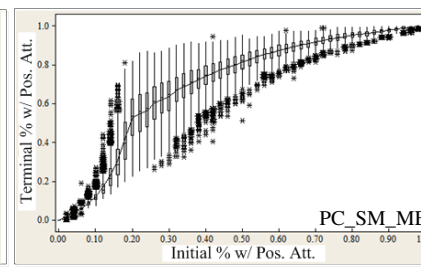
(b) PCC = 0.928



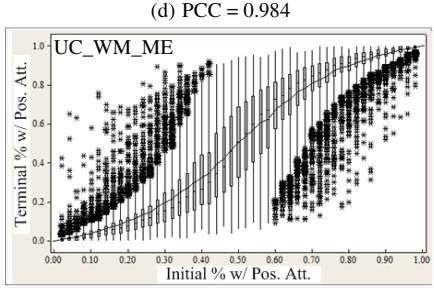
(c) PCC = 0.825



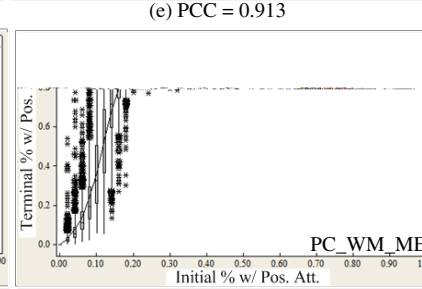
(d) PCC = 0.984



(e) PCC = 0.913

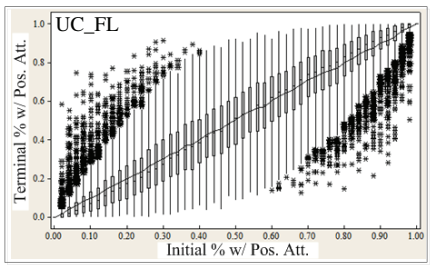


(f) PCC = 0.989

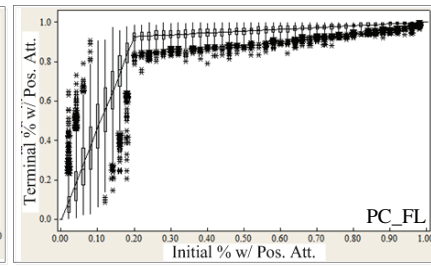


(g) PCC = 0.688

Figure 10.3. (*Continued*)

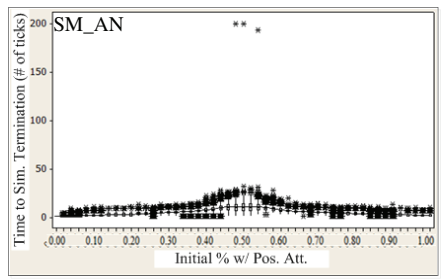


(h) PCC = 0.999

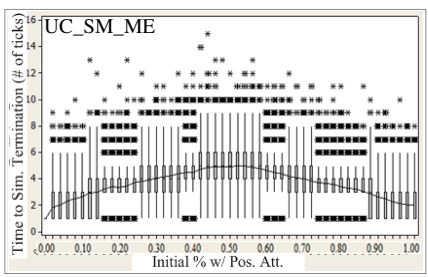


(i) PCC = 0.701

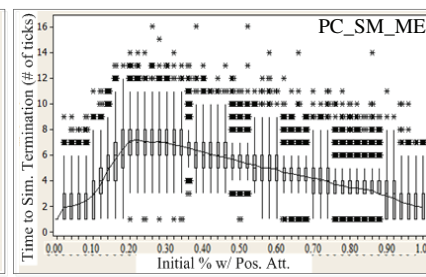
Figure 10.3. The correlation between initial and terminal public opinions in the nine cases (PCC = Pearson correlation coefficient).



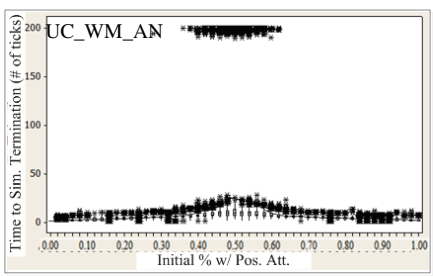
(a)



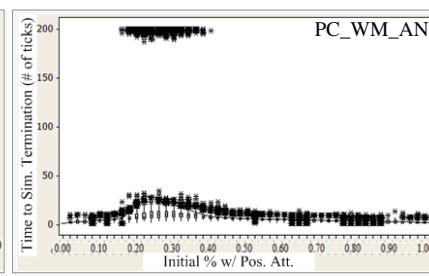
(b)



(c)



(d)



(e)

Figure 10.4. (*Continued*)

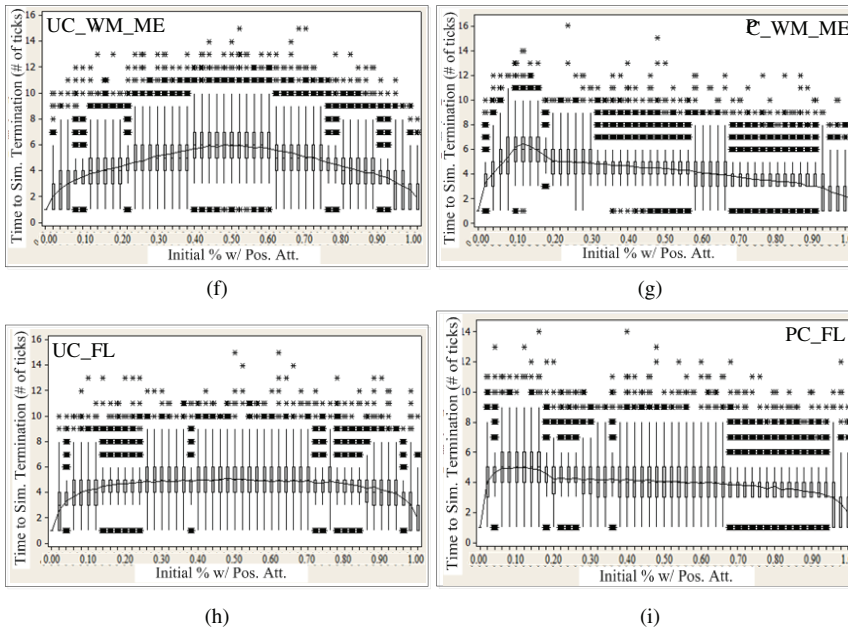


Figure 10.4. The public opinion stabilization results in the nine cases.

Correlation between Initial and Terminal Public Opinions: We analyzed the correlation between initial public opinion (x-axis) and terminal public opinion (y-axis) over multiple simulation runs. We conducted three investigations on how different social influence parameters impacted the influence dynamics.

The first investigation was intended to study the impact of the correlation between expertise index and initial individual opinions (UC vs. PC), by comparing Figs. 10.3b and c, 10.3d–f and 10.3g, 10.3h and i. The comparison suggests that the terminal public opinion was more likely to be in favor of the subject when more experts on the subject held positive opinion initially. This suggestion can be explained by the difference between the Pearson correlation coefficients (e.g., PCC = 0.98 in case UC_SM_ME vs. PCC = 0.91 in case PC_SM_ME). Moreover, in UC cases, the average percentage of agents holding positive opinion at the termination exceeds the initial percentage, until the initial percentage is above 50% (i.e., a population

is already in favor of the subject at the beginning). In contrary, in PC cases, this happens even when there are less than 20%–25% of agents holding positive opinions at the beginning. Finally, we compared the variances on the simulated terminal percentage over the multiple simulation replications. The comparison suggested that noticeable differences only appeared while comparing UC_FL and PC_FL.

The second investigation was intended to study the impact of the decision rule (SM vs. WM vs. FL) by comparing Figs. 10.3a, b, and h; Figs. 10.3d, e, and h; Figs. 10.3a, c, and i; Figs. 10.3a, e, and i. The comparison suggests that the relationship between initial and terminal public opinion will look like a stepwise function with the decision rule of SM, follow a S-shaped curve with WM, and will follow a straight diagonal line in the UC cases. SM ignored expertise completely but aggregated the opinions of all agents contacted in the social network, whereas FL took expertise into account and ignored the opinions of most agents connected in the social network. This implies that in the former cases, the dominant individual opinion (i.e., opinion held by more than 50% of the agents) can propagate among the agents effectively through the social influence mechanism; whereas in the latter cases, the dominant individual opinion cannot be magnified for its homogenous distribution among experts and nonexperts. In terms of linear correlation between initial and terminal public opinions, there are differences between the UC cases and PC cases (e.g., in UC cases UC_SM_AN, UC_WM_AN, and UC_FL_AN, the PCC value increases from 0.898 to 0.928 to 0.999; in PC cases PC_SM_AN, PC_WM_AN, and PC_FL_AN, the PCC value decreases from 0.898 to 0.825, 0.701). These differences can be explained by the nonlinear social influence from the experts in the social network. WM is a hybrid decision rule between SM and FL. The results associated with WM were usually between the corresponding SM and FL cases. Finally, we compared the variances in the simulated terminal percentage. The comparison suggests that the variance is generally small in SM cases but may be larger in WM and FL cases when the initial individual opinion is uncorrelated with the expertise.

The third investigation was intended to study the impact of the contact criterion (AN vs. ME), by comparing Figs. 10.3a–c with Figs. 10.3d–g. The comparison suggests that the terminal public opinion is more likely to be linearly related to the initial public opinion when each individual in the social network only seeks opinion from people having higher expertise indices. This is particularly true when the initial individual opinion is uncorrelated to the expertise. In other words, when the social network is close to even split between the two sides on the subject, the contact criterion of AN tends to help form more unanimous public opinion than the contact criterion ME (e.g., see the comparison between Figs. 10.3b and f). In addition, we compared the variances in the simulated terminal percentage. The comparison suggests that noticeable difference only appears while comparing the two cases UC_WM_AN and UC_WM_ME. Moreover, when the initial public opinion is evenly split, the variance in the terminal percentage of holding positive opinions in the population is smallest when the contact criterion ME is used; moreover, when the initial public opinion is noticeably strong on one side, the variance in the terminal percentage is rather high.

Stabilization of Social Influencing Process and Formation of Public Opinion: To a large extent, the stabilization process coincides with the variation on the terminal percentage of individuals holding the positive opinion on the subject. In the majority of the cases and for most of the initial conditions, the public opinion can stabilize quickly (i.e., within 20–30 iterations). In few cases and initial conditions, the system continues to oscillate up to the pre-specified limit on simulation duration (i.e., 200 iterations). More specifically, significant public opinion oscillation only appears in cases where the contact criterion is AN (e.g., see Figs. 10.4a, d, and e).

10.4.2. *Discussion on policy implications*

Simulation experiments offer valuable insights into facilitating social influences for altering individuals' opinions on contentious topics related to health risks. In the following, we use the example of vaccination intake risk perception to discuss the relevant policy implications.

First, it is important to assess individuals' initial opinions when some outbreak has been identified. The current simulation suggests that if most of the population already has positive opinions on vaccine intake, spending much effort on public health campaigns persuading them to undergo vaccination is unnecessary. Within the constraints and limitations of the current framework, majority opinions typically prevail, unless in cases in which agents use an expert-based decision rule. If most of the population holds negative opinions at the onset, one promising strategy may be to foster the use of expertise-based decision strategies among consumers and to target expert multipliers and key personnel in social networks and persuade them to change their opinions. The simulation study suggests that it is important to understand how individuals seek opinions from their neighbors in social network and how these opinions are processed.

The results of the presented simulation demonstrate that even a small percentage of initial positive opinion of experts may be sufficient to trigger the formation of stable positive opinions among the large majority of the population. The percentage of the population holding positive opinion is sensitive to the decision rule. If the decision rule is SM, the relationship between the terminal public opinion and the initial public opinion is close to a step function, which implies that (1) most populations, after some time, will hold either positive opinion or negative opinion and (2) there is a relatively sharp threshold, around 50%, for the percentage of initial positive opinion. If the decision rule is FL, this relationship is close to a straight, linear function, which implies that the percentage of terminal public opinion is similar to that of initial public opinion. If the decision rule is WM, the relationship can be best described by a sigmoid function, which implies that the percentage of terminal public opinion is somewhat between the other two decision rules.

The public opinion in the cohort typically reached the steady state in 30–40 decision cycles. The configurations with which the public opinion in the population could not quickly reach the steady state were those where initially about half of the population held positive opinions while the other half held negative opinions.

10.5. Conclusions and Future Work

We applied agent-based simulations to capture the dynamics of social influence processes after some potentially disastrous events, such as a disease outbreak. We simulated and assessed how individual opinions were altered and to what extent the public opinion was a function of given initial individual opinions. We also measured how quickly the public opinion was formed. In the model, we assumed the topology of a scale-free social network and tested the impact of several representative decision rules and contact criteria. We also assessed the effect of having initial individual opinions positively correlated with the expertise in the social network. The main take-home messages are as follows: (1) the distribution of the final public opinions are positively correlated with the initial distribution; (2) the relationship between the distribution of final public opinions and initial distributions is systematic but typically not linear, reflecting social influence processes; (3) under certain circumstances (namely, when the majority holds positive opinion), the social influence process can quickly alter the opinions for others holding negative opinions and, thus, it is not necessary to make huge effort in public campaigns for adopting the health intervention (e.g., vaccine intake); and (4) when the initial assessment indicates significant obstacle in adopting the intervention, it is critical to influence and alter the opinions of certain experts at the beginning of the campaigns.

One main benefit of our work lies in the suggestion to consider infectious disease control immediately following a disease outbreak when the disease is relatively unknown to the general public. The work may also be translated to the prevention of infectious diseases at the endemic stage. For example, research has shown that adherence to measles-mumps-rubella vaccination has been suboptimal due to low perceived risk infection and lack of immediate benefit. In summary, this work could potentially benefit policy makers in making informed resource-allocation decisions on public health campaigns for infectious disease prevention and control.

There are several limitations in this work. First, it ignores the impact of several social elements, such as social events. Second, it

ignores the impact of information filtering techniques in recommendation systems such as mass media. Third, it does not differentiate individuals in the social network with heterogeneous decision rules and contact criteria, as well as with different risk perceptions and efficacy beliefs on the matter of interest. Finally, in the presented simulation, individuals in the network are assumed to be static and not adaptive in choosing their contact and decision rules during the social influence process. Future studies may model individual deciders that adapt their decision and contact rules and heterogeneous populations that involve agents using different strategies.

Appendix

Detailed Description of the Social Influence Dynamics Modeling

We consider a social network that contains a set of agents, denoted by N. For each agent $i \in N$, we denote $N(i)$ to be the set of agents immediately connected to agent i (or say agent i's neighbors). In addition, we associate each agent i with an expertise index, denoted by p_i. Without loss of generality, we assume this index ranges between 0 and 1, inclusively, that is, $p_i \in [0, 1]$ for all $i \in N$, with 1 implying the absolute authority on the subject, and 0 implying the complete novice on the subject. We let $N'(i) \subseteq N(i)$ be the set of agent i's neighbors whose expertise indices are greater than p_i, that is, $N'(i) = \{j \in_N (i) \mid p_j > p_i\}$.

At any time $t >= 0$, we denote each agent i's attitude to be $xi(t)$. Let $x(t) = (x_1(t), \ldots xN(t))$. We assume that each agent holds dichotomous attitude toward the subject, that is, each agent can only be in favor of the subject or against it. We label $x_i(t) = 1$ if agent i's attitude is for the subject; and $x_i(t) = 0$ otherwise. To specify the initial attitude of each agent, that is, $x_k(0)$ for each $k \in N$, we further consider two scenarios, depending on whether the initial attitude is formed based on the individual's expertise. If the scenario is IN, we generate $x_i(0)$ randomly and independently of p_i. If the initial attitude is assumed to be positively correlated to the expertise, termed

scenario PC, we generate $x_i(0)$ with the following procedure. We randomly select 20% of the agents to be "experts" and assign their expertise indices according to a uniform distribution between 0.8 and 1. We assign each of the remaining 80% (i.e., nonexperts) an expertise index according to a uniform distribution between 0.1 and 0.5. Note that experts constitute 20% of the entire population. So, if the percentage of agents holding positive attitude initially is given to be less than 20%, those agents that hold initial positive attitude are randomly generated among the experts only. If such percentage is given to be greater than 20%, all the agents representing experts are assigned with initial positive attitude. Then, some agents from the remaining 80% nonexpert population are randomly selected and assigned with initial positive attitude. Once we complete the assignment of initial positive attitude, we assign the remainder of the population with initial negative attitude.

At time $t > 0$, we assume that individual attitudes are updated following some pre-specified order, denoted by list L. That is, we update $x_i(t)$ until we have updated the attributes of all entities in $S(t)$ preceding i. Even though an agent's attitude has been updated at time t, we assume that it does not influence the agents that are updated later (i.e., later in the list L) at time t. The updates are simulated in a discrete fashion. Thus, we use integers to index the time points at which the system is updated. We present the system update algorithm as follows.

Environment Parameters: $N(i)$, p_i for all $i \in N$; L; individual decision rule and contact criterion that determine the updating scheme.

Input: $x(t-1)$

Output: $x(t)$

Step 0: Let $x(t) \leftarrow x(t-1)$ and $k = 1$.

Step 1: Select the kth agent in L to update and use i to represent the label of the agent. Update $x_i(t)$ based on one of the following updating schemes.

Updating Scheme I: Decision rule is "SM"; contact criterion is "AN":

If $\sum_{j \in N(i)} x_j(t-1) \leq \frac{|N(i)|}{2}$ (i.e., among agent i's neighbors, if more

neighbors are against the subject), then $x_i(t) = 0$; otherwise, $x_i(t) = 1$.

Scheme II: Decision rule is "SM"; contact criterion is "ME":

If $\sum_{j \in N'(i)} x_j(t-1) \le \frac{|N'(i)|}{2}$ (i.e., among agent I's neighbors whose expertise indices are higher than *i*'s, if more neighbors are against the subject), then $x_i(t) = 0$; otherwise, $x_i(t) = 1$.

Scheme III: Decision rule is "WM", and contact criterion is "AN":

If $\sum_{j \in N(i)} x_j(t-1)p_j \le \sum_{j \in N(i)} (1 - x_j(t-1))p_j$ (i.e., among agent *i*'s neighbors, if the expertise index weighted sum of individual attitudes for those against the subject is greater than the weighted sum for those for the subject), then $x_i(t) = 0$; otherwise, $x_i(t) = 1$.

Scheme IV: Decision rule is "WM", and contact criterion is "ME":

If $\sum_{j \in N'(i)} x_j(t-1)p_j \le \sum_{j \in N'(i)} \left(1 - x_j(t-1)\right)p_j$ (i.e., among agent *i*'s neighbors whose expertise indices are higher than *i*'s, if the expertise index weighted sum of individual attitudes for those against the subject, is greater than the weighted sum for those for the subject), then $x_i(t) = 0$; otherwise, $x_i(t) = 1$.

Scheme V: Decision rule is "FL", and contact criterion is "AN":

Let $j^* = \arg\max_{j \in N(i)} p_j$ (i.e., j^* has the largest expertise index among agent *i*'s neighbors), then $x_i(t) = x_{j^*}(t-1)$.

Scheme VI: Decision rule is "FL," and contact criterion is "ME":

Let $j^{**} = \arg\max_{j \in N'(i)} p_j$ (i.e., j^{**} has the largest expertise index, among agent *i*'s neighbors whose expertise indices are higher than *i*'s), then $x_i(t) = x_{j^{**}}(t-1)$ (Note that this scheme yields the same updates as Scheme II. So, we ignore it.)

Step 2: $k = k + 1$. If $k = |N|$, STOP, otherwise go to Step 1.

References

1. Macy, MW and R Willer (2002). From factors to actors: Computational sociology and agent-based modeling. *Ann Rev Sociol.*, 28, 143–166.
2. Kirman, A and J Zimmermann (2001). *Economics with Heterogeneous Interacting Agents*. Berlin: Springer-Verlag.

3. Latané, B and MJ Bourgeois (2000). Dynamic social impact and the clustering, correlation, and continuing diversity of culture. In *Handbook of social psychology: group processes*, RS Tindale and M Hogg (eds.), Vol. 4, pp. 235–258. Hoboken: Wiley-Blackwell.
4. Kenrick, DT, NP Li and J Butner (2003). Dynamic evolutionary psychology: Individual decision rules and emergent social norms. *Psychol Rev.*, 110(1), 3–28.
5. Kohler, T and G Gumerman (2002). *Dynamics in Human and Primate Societies*. Oxford: Oxford University Press.
6. Bonabeau, E (2002). Agent-based modeling: Methods and techniques for simulating human systems. *PNAS*, 99(suppl 3), 7280–7287.
7. Axelrod, R (1997). *The Complexity of Cooperation: Agent-Based Models of Competition and Collaboration*. Princeton, NJ: Princeton University Press.
8. Epstein, JM and R Axtell (1996). Growing artificial societies: Social science from the bottom up. Brookings Institution. Boston: MIT Press.
9. Goldspink, C (2002). Methodological implications of complex systems approaches to sociality: Simulation as a foundation for knowledge. *JASSS*, 5(1), 3.
10. Arthur, WB, SN Durlauf and DA Lane (eds.) (1997). *The Economy as an Evolving Complex System II, SFI Studies in the Sciences of Complexity*. Reading, MA: Addison-Wesley.
11. Bagni, R, R Berchi and P Cariello (2002). A comparison of simulation models applied to epidemics. *JASSS*, 5(3), 5.
12. Kohler, TA, GJ Gumerman, and RG Reynolds (2005). Simulating ancient societies. *Scientific American*, 293(1), 77–84.
13. DeStefano, F and RT Chen (2001). Autism and measles-mumps-rubella vaccination: Controversy laid to rest? *CNS Drugs*, 15(11), 831–837.
14. Friedkin, NE (1999). Choice shift and group polarization. *Am Sociol Rev.*, 64, 856–875.
15. Hegselmann, R and U Krause (2002). Opinion dynamics and bounded confidence: Models, analysis and simulation. *JASSS*, 5(3), 2.
16. Janssen, MA and W Jager (2001). Fashions, habits and changing preferences: simulation of psychological factors affecting market dynamics. *J Economic Psychol.*, 22, 745–772.
17. Goldstone, RL and MA Janssen (2005). Computational models of collective behavior. *TRENDS in Cognitive Sciences*, 9(9), 424–430.

18. Cho, H, T Reimer and KA McComas (eds) (2015). *The SAGE Handbook of Risk Communication*. Thousand Oaks, CA: SAGE Publications.
19. Novak, A, J Szamrej and B Latané (1990). From private attitude to public opinion: A dynamic theory of social impact. *Psychol Rev.*, 97(3), 362–376.
20. Kerr, NL and RS Tindale (2004). Group performance and decision making. *Ann Rev Psychol.*, 55, 623–655.
21. Gigerenzer, G, Todd PM and the ABC Research Group (1999). *Simple Heuristics that Make Us Smart*. New York: Oxford University Press.
22. Luan, S, K Katsikopoulos and T Reimer (2012). When does diversity trump ability (and vice versa) in group decision making? A simulation study. *PLoS One*, 7(2), 1–9, e31043.
23. Schwenk, G and T Reimer (2008). Simple heuristic in complex networks: Models of social influence. *JASSS*, 11(3), 4.
24. Fischhoff, B, NT Brewer and JS Downs (eds.) (2011). Communicating risks and benefits: An evidence-based user's guide. Washington, DC: Food and Drug Administration, U.S. Department of Health and Human Services.
25. Reimer, T and U Hoffrage (2012). Ecological rationality for teams and committees: Heuristics in group decision making. In *Ecological Rationality: Intelligence in the World*, PM Todd, G Gigerenzer and the ABC Research Group (eds.), pp. 266–286. New York: Oxford University Press.
26. Carley, KM, MJ Prietula and Z Lin (1998). Design versus cognition: The interaction of agent cognition and organizational design on organizational performance. *JASSS*, 1(3), 4.
27. Sun, R and I Naveh (2004). Simulating organizational decision-making using a cognitively realistic agent model. *JASSS*, 7(3), 5.
28. Smith, KP and NA Christakis (2008). Social network and health. *Ann Rev Sociol.*, 34, 405–429.
29. Christakis, NA and JH Fowler (2013). Social contagion theory: Examining dynamic social networks and human behavior. *Stat Med.*, 32, 556–577.
30. Christakis, NA and JH Fowler (2009). *Connected: The Surprising Power of Our Social Networks and How They Shape Our Lives*. New York: Little Brown and Co.
31. Gallagher, R and T Appenzeller (1999). Beyond reductionism. *Science*, 284(5411), 79.

32. Service, RF (1999). Exploring the systems of life. *Science*, 284(5411), 80–83.
33. Weng, G, US Bhalla and R Iyengar (1999). Complexity in biological signaling systems. *Science* 284(5411), 92–96.
34. Koch, C and G Laurent (1999). Complexity and the nervous system. *Science*, 284(5411), 96–98.
35. Chakrabarti, S, B Dom and Members of the Clever Project (1999). Hypersearching the web. *Sci Am.*, 280(54), 54–60.
36. Albert, R, H Jeong and AL Barabási (1999). Diameter of the worldwide web. *Nature*, 401, 130–131.
37. Barabási, A-L, R Albert and H Jeong (1999). Mean-field theory for scale-free random networks. *Physica A*, 272, 173–187.
38. Wasserman, S and K Faust (1994). *Social Network Analysis*. Cambridge, UK: Cambridge University Press.
39. Erdös, P and A Renyi (1960). On the evolution of random graphs. *Publ Math Inst Hung Acad Sci A.*, 5, 17–61.
40. Bollobás, B (1985). *Random Graphs*. London: Academic Press.
41. Barabási, A-L and R Albert (1999). Emergence of scaling in random networks. *Science*, 286(5439), 509–512.
42. Dawber, TR (1980). *The Framingham Study: The Epidemiology of Atherosclerotic Disease*. Cambridge, MA: Harvard University Press.
43. Feinleib M, WB Kannel, RJ Garrison, PM McNamara and WP Castelli (1975). The Framingham offspring study. Design and preliminary data. *Prev Med.*, 4, 518–525.
44. Christakis, NA and JH Fowler (2008). The collective dynamics of smoking in a large social network. *New Eng J Med.*, 358, 2249–2258.
45. Rosenquist, JN, J Murabito, JH Fowler and NA Christakis (2010). The spread of alcohol consumption behavior in a large social network. *Ann Intern Med.*, 152, 426–433.
46. Fowler, JH and NA Christakis (2008a). Dynamic spread of happiness in a large social network: Longitudinal analysis over 20 years in the Framingham heart study. *Brit Med J.*, 337, a2338.
47. Cacioppo, JT, JH Fowler and NA Christakis (2009). Alone in the crowd: The structure and spread of loneliness in a large social network. *J Pers Soc Psychol.*, 97, 977–991.
48. Rosenquist, JN, JH Fowler and NA Christakis (2011). Social network determinants of depression. *Mol Psychiatry*, 16, 273–281.
49. Harris, KM, PS Bearman and JR Udry (2010). The national longitudinal study of adolescent health: Research design, 2010. http://www.cpc.unc.edu/projects/addhealth/design.

50. Lewis, K, J Kaufman, M Gonzalez, A Wimmer and N Christakis (2008). Tastes, ties, and time: A new social network dataset using Facebook.com. *Soc Netw.*, 30, 330–342.
51. Bond, RM, CJ Fariss, JJ Jones, AKE Settle, C Marlow and JH Fowler (2012). A massive scale experiment in social influence and political mobilization. *Nature*, 489, 295–298.
52. Christakis, NA and JH Fowler (2010). Social network sensors for early detection of contagious outbreaks. *PloS One 5*, e12948.
53. Apicella, CL, FW Marlowe, JH Fowler and NA Christakis (2012). Social networks and cooperation in hunter-gatherers. *Nature*, 481, 497–501.
54. Fowler, JH and NA Christakis (2010). Cooperative behavior cascades in human social networks. *Proc Nat Acad Sci.*, 107, 5334–5338.
55. Rand, DG, S Arbesman and NA Christakis (2011). Dynamic social networks promote cooperation in experiments with humans. *Proc Nat Acad Sci.*, 108, 19193–19198.
56. Fowler, JH, CT Dawes and NA Christakis (2009). Model of genetic variation in human social networks. *Proc Nat Acad Sci.*, 106, 1720–1724.
57. O'Malley, AJ and NA Christakis (2011). Longitudinal analysis of large social networks: Estimating the effect of health traits on changes in friendship ties. *Stat Med.*, 30, 950–964.
58. Christakis, NA, JH Fowler, GW Imbens and K Kalyanaraman (2010). An empirical model for strategic network formation. National Bureau of Economic Research Working Paper Series. 16039.
59. Onnela, J-P and NA Christakis (2011). Spreading paths in partially observed social networks. *Phys Rev E*, 85, 036106.
60. Onnela, J-P, S Arbesman, MC González, A-L Barabási and NA Christakis (2011). Geographic constraints on social network groups. *PloS One*, 6, e16939.
61. Fowler, JH, JE Settle and NA Christakis (2011). Correlated genotypes in friendship networks. *Proc Nat Acad Sci.*, 108, 1993–1997.
62. Christakis, NA and JH Fowler (2007). The spread of obesity in a large social network over 32 years. *New Eng J Med.*, 357, 370–379.
63. Fowler, JH and NA Christakis (2008b). Estimating peer effects on health in social networks: A response to Cohen-Cole and Fletcher and Trogdon, Nonnemaker, and Pais. *J Health Econ.,* 27, 1400–1405.
64. Keating, NL, AJ O'Malley, JM Murabito, KP Smith and NA Christakis (2011). Minimal social network effects evident in cancer screening behavior. *Cancer*, 117, 3045–3052.

65. Mednick, SC, NA Christakis and JH Fowler (2010). The spread of sleep loss influences drug use in adolescent social networks. *PloS One*, *5*, e9775.
66. Pachucki, MA, PF Jacques and NA Christakis (2011). Social network concordance in food choice among spouses, friends, and siblings. *Am J Public Health*, 101, 2170–2177.
67. Jackson, MO (2008). *Social and Economic Networks*. Princeton, NJ: Princeton University Press.
68. Goyal, S (2007). *Connections: An Introduction to the Economics of Networks*. Princeton, NJ: Princeton University Press.
69. O'Malley, A and P Marsden (2008). The analysis of social networks. *Health Ser Outcomes Res Methodol.*, 8, 222–269.
70. Newman, MEJ (2003). The structure and function of complex networks. *SIAM Rev.*, 45, 167–256.
71. Newman, MEJ (2010). *Networks: An Introduction*. Oxford, UK: Oxford University Press.
72. Easley, D and J Kleinberg (2010). *Networks, Crowds, and Markets: Reasoning about a Highly Connected World*. Cambridge, UK: Cambridge University Press.
73. Kolaczyk, ED (2009). *Statistical Analysis of Network Data: Methods and Models*. New York, NY: Springer.
74. Cialdini, RB (2001). The science of persuasion. *Sci Am.*, 284(2), 76–81.
75. Reimer, T, R Hertwig and S Sipek (2012). Probabilistic persuasion: A Brunswikian theory of argumentation. In *Simple Heuristics in a Social World*, R Hertwig, U Hoffrage and the ABC Research Group (eds.), pp. 33–55. New York, NY: Oxford University Press.

11. Growth Curves of American Children Differ Significantly from CDC Reference Standards

Xiang Zhong*, Jingshan Li†, Goutham Rao‡ and K. P. Unnikrishnan§

*Department of Industrial and Systems Engineering, University of Florida, Gainesville, FL 32611, USA
†Department of Industrial and Systems Engineering, University of Wisconsin, Madison, WI 53706, USA
‡Center for Biomedical Research Informatics, NorthShore University HealthSystem, Evanston, IL 60201, USA
§Global Data Insight and Analytics (GDIA), Ford Motor Company, Dearborn, MI, 48126, USA

Abstract

Background: Anthropometric measurements such as weight, height (stature), and body mass index (BMI) provide reliable indicators of children's growth. The national standards in USA for these measurements are the 2000 Centers for Disease Control and Prevention (CDC) growth charts, which were generated using data from 1963 to 1994. In this paper, methodologies identical to that of the CDC were used to generate growth charts from more recent datasets.

These charts, derived from both publically available and hospital system datasets, provide a glimpse of the current growth of American children.

Methods: The datasets were from the National Health and Nutrition Examination Survey (NHANES) for years 1999 to 2010 and from NorthShore University HealthSystem's Enterprise Data Warehouse (NS-EDW) for years 2006 to 2012. The weight-for-age, stature-for-age, and BMI-for-age percentile curves and the associated L, M, and S parameters for both boys and girls aged 2–20 years were generated. A free-standing software program to achieve this was also created.

Results: The weight and BMI percentile curves generated from NS-EDW and NHANES data differ substantially from the CDC percentile curves, while those for stature do not. The weight and BMI of children at each of the percentiles are significantly higher at all ages compared with the CDC reference standard curves.

Conclusion: A software program that generates growth curves of any population of children using the CDC method was developed and successfully applied to two recent datasets. Meaningful comparisons between the growth curves generated from these datasets, and the CDC's reference curves were provided. These charts provided a visual representation of how dramatically the weight and BMI growth curves of today's children differ from the CDC's reference standard curves.

11.1. Introduction

Children's growth curves are very useful and a variety of reference curves have recently been developed and published [1–8]. The growth curves developed by the CDC are the most widely used ones in USA for clinicians to track the growth of children, serving as a reference standard, upon which a diagnosis of overweight and obesity can be made [9]. The curves, which were last updated in 2000, are based on a composite of data collected from five national health and examination surveys between 1963 and 1994) and five supplemental data sources [9]. Thus, the included data are relatively old and may not reflect the rapidly changing demographic trends in the US population. Furthermore,

given the increasing prevalence of obesity, it is not clear whether the shape or trajectory described by the CDC curves can be extrapolated to describe the growth patterns of children today. It is well known that today's children are heavier than those of prior generations [10, 11]. However, it is not clear to what extent the shape or growth trajectory of today's children differs from the reference standard curves [12]. The better representation of the current patterns of weight gain among children will require more up-to-date growth curves.

The overall purpose of this study is to replicate the methodology used by the CDC and apply it to more recently collected sources of pediatric growth data using an innovative software program. This paper applied the methodology to data collected from a large repository of electronic health records (EHR) information — NorthShore University HealthSystem Enterprise Data Warehouse (NS-EDW) [13,14]. NorthShore University HealthSystem is an integrated healthcare delivery system that serves patients throughout the Chicago metropolitan area [15]. As data from this source may not represent American children as a whole, the CDC's methods were also applied to data from the National Health and Nutrition Examination Survey (NHANES), carried out between 1999 and 2010 [18,19]. The prevalence of obesity in USA has been evaluated based on the NHANES data [19, 20]. Since CDC curves used data collected prior to 1994, the curves from the more recent NHANES data represent the current condition of child population in a better way. The specific objective here was to identify significant differences between the growth curves constructed from the two recent sources and the CDC curves. This is the first time that a practical, software-based strategy applying the CDC's methodology to data collected from clinical sources has been carried out.

11.2. Methods

As stated above, the aim of the study was to reproduce the CDC's methodology to create growth charts and apply it to two current sources of data: NHANES (1999–2010) and NS-EDW (2006–2012).

The methods used to generate the 2000 CDC growth charts were described by Kuczmarski *et al.* [9]. Statistical procedures were applied to the observed data in two stages: first, to generate initial smoothed curves for selected major percentiles, and second, to generate the parameters that were used to construct the final smoothed curves and additional percentiles. In the smoothing stage, selected empirical percentiles were smoothed with a variety of parametric and non-parametric regression procedures. In the transformation stage, the smoothed curves were approximated using a modified LMS estimation procedure to provide the transformation parameters, λ, μ, and σ (LMS), and compute additional percentiles and z-scores. The details of the statistical procedures, including smoothing (Procedure 1) and transformation (Procedure 2), are shown in Appendix A. Discussions regarding the performance of the CDC's LMS transformation [9] and the LMS methods by Cole [3–6] are presented [21].

11.2.1. *Data sources*

The general information of each data source and the corresponding charts they generated are summarized in Table 11.1. The timeline of the three data sources is illustrated in Fig. 11.1. The information, stratified by sex and race/ethnicity for the NS-EDW and NHANES datasets, is summarized in the Data Statistics section. A detailed description of the CDC dataset is given by Kuczmarski *et al.* [9].

11.2.1.1. *NorthShore University HealthSystem Enterprise Data Warehouse*

To enhance the capability of carrying out research using information from EHRs, NorthShore University HealthSystem (NS) developed a state-of-the-art clinical informatics system, an Enterprise Data Warehouse (EDW) [13, 14]. The EDW captures clinical and administrative data for quality improvement and research. Data from more than 400,000 encounters from years 2006 through 2012 are collected [15]. All data available for research are fully de-identified.

Table 11.1. Data characteristics.

Dataset	Year	Subject	Sex	Chart
NS-EDW Data				
NS-EDW	2006–2012	Age:2–20	M,F	W,S,BMI
NHANES Data				
NHANES	1999–2010	Age:0–26	M,F	W,S,BMI
NHANES	1999–2002	Age:0–26	M	W
NHANES	2003–2006	Age:0–26	M	W
NHANES	2007–2010	Age:0–26	M	W
CDC Data				
NHES2	1963–1965	Age:6–12	M,F	W,S,BMI
NHES3	1966–1970	Age:12–18	M,F	W,S,BMI
NHANES1	1971–1974	Age:1–20	M,F	W
NHANES1	1971–1974	Age:2–25	M,F	S,BMI
NHANES2	1976–1980	Age:1–20	M,F	W
NHANES2	1976–1980	Age:2–25	M,F	S,BMI
NHANES3	1988–1994	Age:1–6	M,F	W
NHANES3	1988–1994	Age:2–25	M,F	S
NHANES3	1988–1994	Age:2–6	M,F	BMI

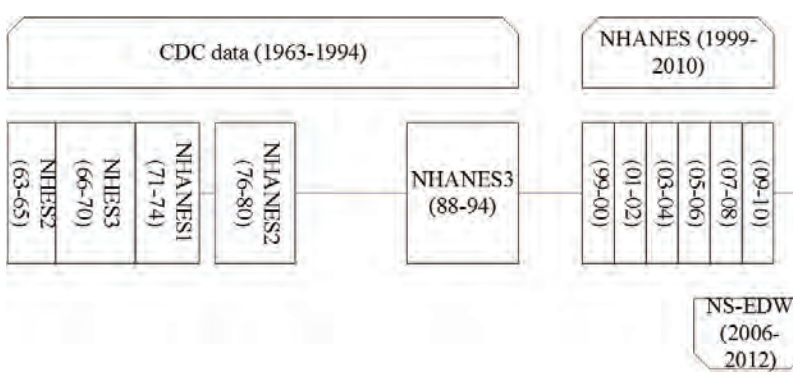


Figure 11.1. Timeline for data sources.

11.2.1.2. *National Health and Nutrition Examination Survey*

The National Health and Nutrition Examination Survey (NHANES) is a program of studies designed to assess the health and nutritional status of adults and children in USA [16]. The survey combines interviews, including demographic, socioeconomic, dietary, and health-related questions; physical examinations, consisting of medical, dental, and physiological measurements; and laboratory tests administered by highly trained medical personnel [16, 17].

11.2.2. *Inclusion criteria*

Data from six NHANES national surveys (99–00, 01–02, 03–04, 05–06, 07–08, 09–10) were pooled to construct growth charts. To achieve better precision of empirical percentiles, pooling was introduced to enhance the number of subjects for each age group, thereby increasing the stability of the outlying percentile estimates. Only data from children and adolescents aged 2 to 20 years were included for all data sources.

11.2.3. *Exclusion criteria*

Several exclusions were made prior to data processing. NS-EDW data without weight, stature, or BMI information were excluded. NS children with less than five weight and stature measurements were also excluded. Weights greater than 200 kg and statures greater than 242 cm were assumed to be outliers due to inaccurate measurements or recordings and, hence, were excluded. BMI values less than 6 kg/m^2 or greater than 100 kg/m^2 were also excluded.

Similarly, the NHANES data were first filtered by encounters whose ages were within the scope of the study (2–20 years). Then, data with missing weight, stature, or BMI information were excluded. Weights greater than 200 kg and BMIs greater than 100 kg/m^2 were also excluded.

11.2.4. *Data statistics*

NS-EDW data were extracted from the EHR information database. The query results after data exclusion were 7,592 boys and 6,878 girls between the ages of 2 and 20 years, each with a minimum of five BMI measurements separated in time. The BMI values ranged from 6.7 to 63.1 kg/m^2.

The NHANES data had 62,160 encounters in total, with approximately 10,000 encounters for each bi-yearly dataset. After the exclusion of missing values, there were 11,820 encounters from boys and 11,538 encounters from girls with complete information regarding age at sampling, weight, stature, and BMI. The BMI values ranged from 7.99 to 66.32 kg/m^2.

To generate the CDC-like growth charts with NS-EDW and NHANES data for boys and girls, the statistical procedures using custom written computer programs were replicated in R [21]. The *quantile()* function [22] was applied to generate empirically selected percentiles, and the *lowess()* function [23] was applied for locally weighted regression. After smoothing, the *lm()* function [24] was used to generate generalized linear regression models for weight-for-age and BMI-for-age percentile curves, and *nls2()* function [25] was used to construct non-linear regression models for stature-for-age percentile curves. To compare to the CDC growth charts, we referred to Martino [26], which implemented a function in R to re-create the CDC growth charts according to the data provided by the CDC [27]. Additional details about the R software programs are available from the authors upon request.

11.3. Results

The detailed demographic information for the filtered NS-EDW and NHANES data are summarized in Tables 11.2 and 11.3, respectively. The age distributions of the two datasets are shown in Fig. 11.2. Two-sample KS tests were conducted to compare the age distribution of the two datasets for boys and girls. Both boys' and

Table 11.2. Demographic information for NS-EDW data.

	Boys	**Girls**
Encounters	50,775	45,390
Patients	7,592	6,878
Race/Ethnicity		
African American	5.9%	5.5%
American Indian	0.6%	0.7%
Asian	3.6%	4.5%
Caucasian	57.8%	57.9%
Hispanic/Latino	5.0%	4.8%
Other	27.0%	26.6%
Measurement		
Wt (kg)	7.2–180.1	4.2–163.2
St (cm)	60.9–241.3	53.0–221.0
BMI (kg/m^2)	6.73–63.11	6.76–56.69

Table 11.3. Demographic information for NHANES data.

	Boys	**Girls**
Encounters	11,820	11,538
Ethnicity		
Mexican American	31.2%	31.8%
Other Hispanic	6.3%	6.4%
Non-hispanic White	28.5%	27.9%
Non-hispanic Black	29.2%	28.5%
Other multi-racial	4.8%	5.3%
Measurement		
Wt (kg)	9.7–239.4	8.9–174.8
St (cm)	79.0–204.4	78.0–187.2
BMI (kg/m^2)	11.98–66.32	7.99–62.08

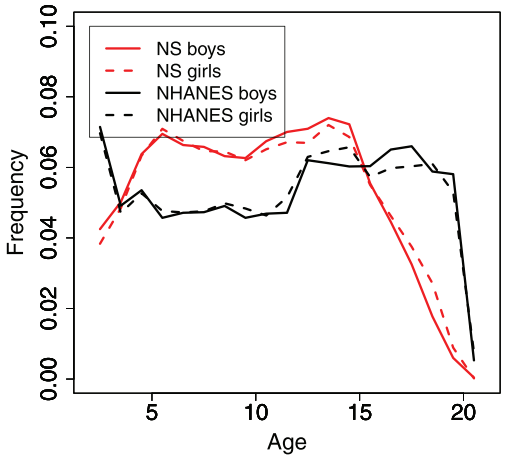


Figure 11.2. Age distributions of NHANES and NS-EDW datasets.

girls' age distributions of the two populations were significantly different ($p < 0.001$). Chi-square tests were conducted to compare the sex distribution of the two datasets. The difference between the two datasets with respect to sex distribution was significant ($p < 0.001$).

For each dataset, weight-for-age, stature-for-age, and BMI-for-age percentiles were generated separately for boys and girls and compared with the 2000 CDC growth curves (see Figs. 11.3–11.8). Furthermore, to illustrate the variation trend over years of sampling, the NHANES data were separated into three non-overlapping datasets of four years (NHANES 99–02, NHANES 03–06, and NHANES 07–10). Five weight-for-age growth charts for boys were generated from the 2000 CDC data, Using the three non-overlapping four-yearly NHANES datasets, and 2000 CDC data and NS-EDW 06–12 data, five weight-for-age growth charts for boys were generated and compared in Fig. 11. 9. The curves show an ascending trend from datasets one to five. To better illustrate the difference, three selected representative percentiles (3rd, 50th, and 97th) were compared, as shown in Fig. 11.10.

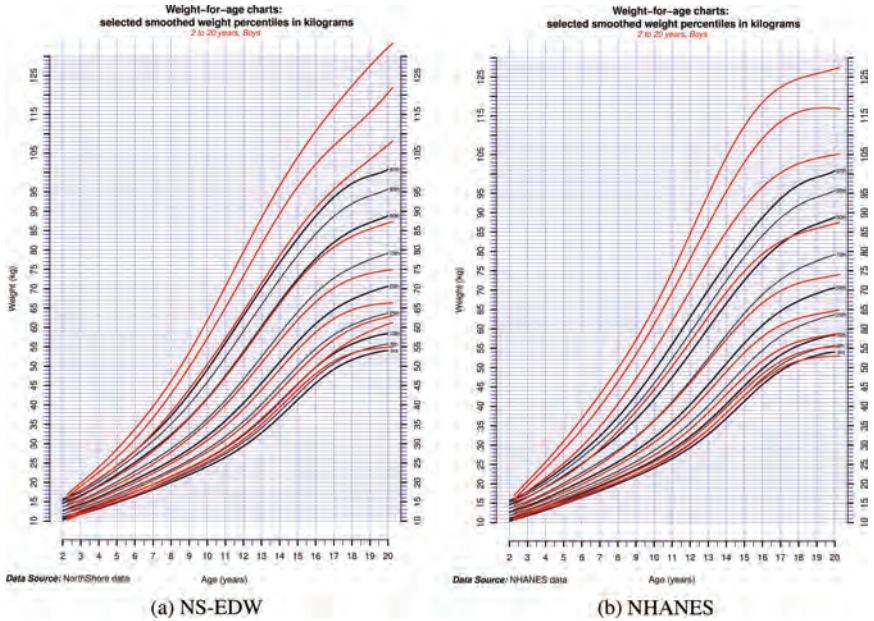


Figure 11.3. Boys' weight-for-age (2–20 years).

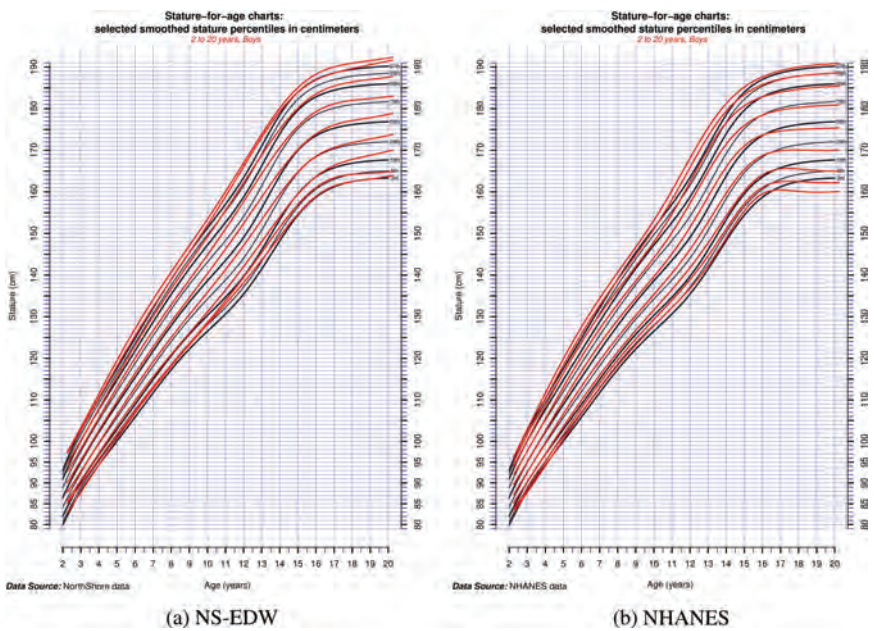


Figure 11.4. Boys' stature-for-age (2–20 years).

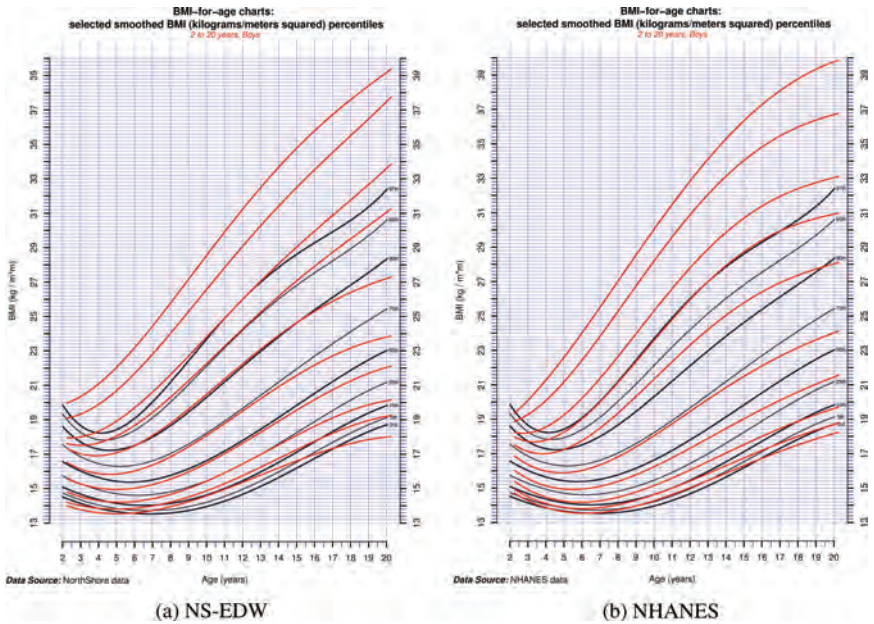


Figure 11.5. Boys' BMI-for-age (2–20 years).

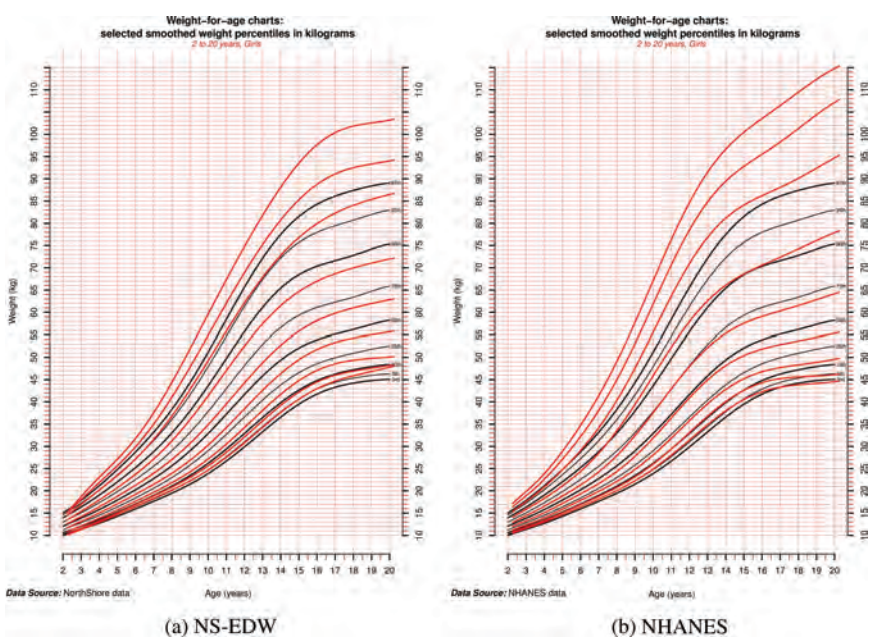
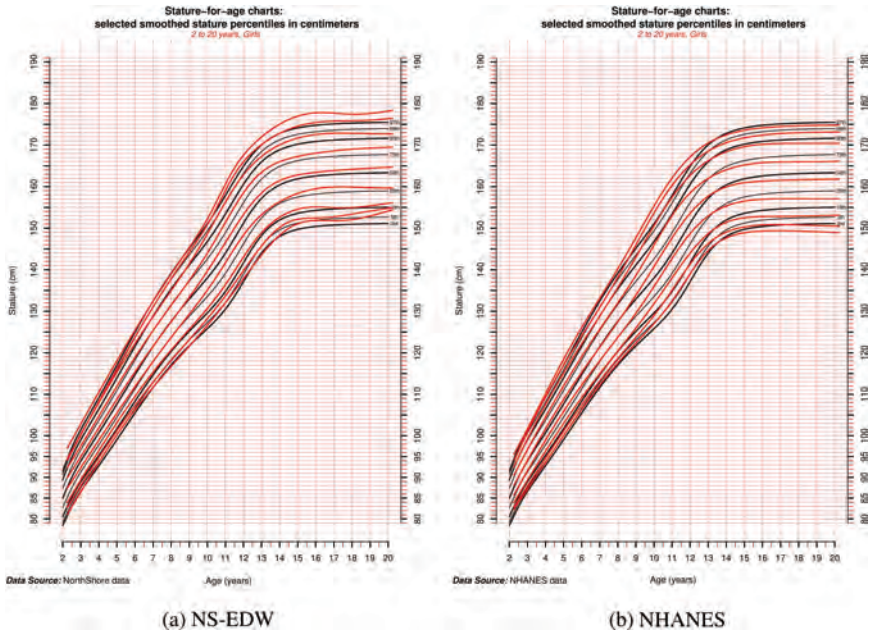
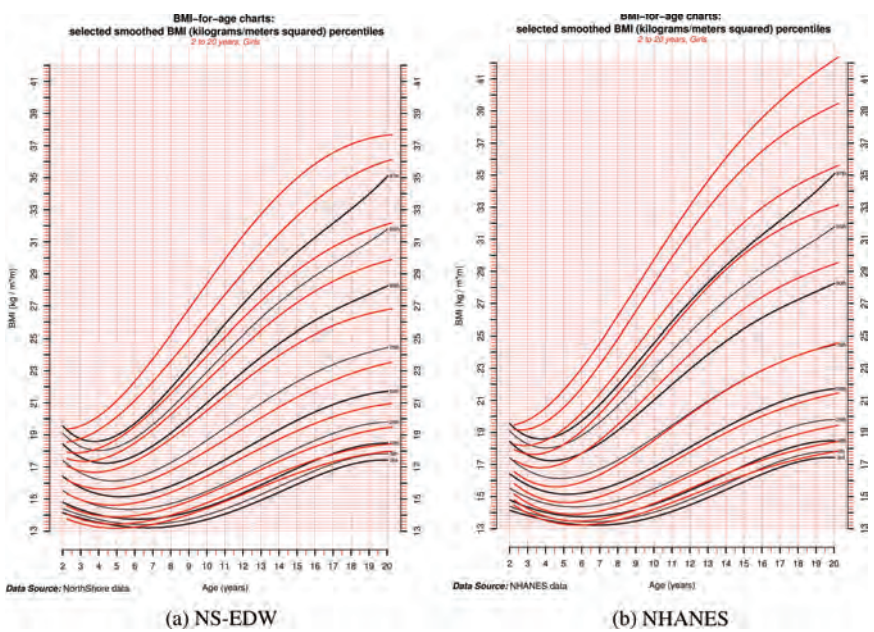


Figure 11.6. Girls' weight-for-age (2–20 years).



(a) NS-EDW (b) NHANES

Figure 11.7. Girls' stature-for-age (2–20 years).



(a) NS-EDW (b) NHANES

Figure 11.8. Girls' BMI-for-age (2–20 years).

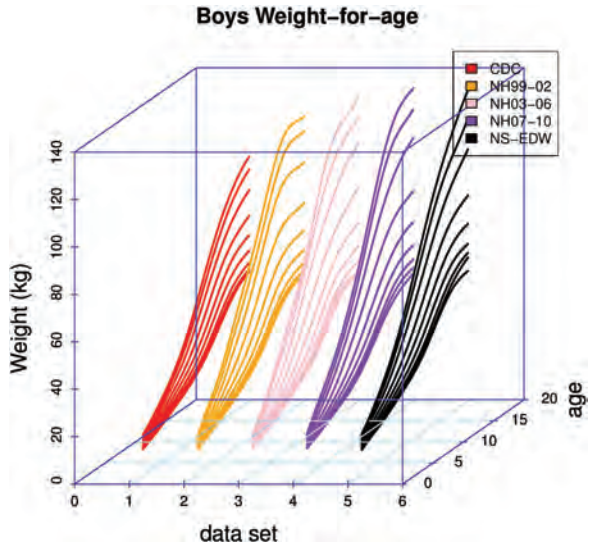


Figure 11.9. Boys' weight-for-age comparison.

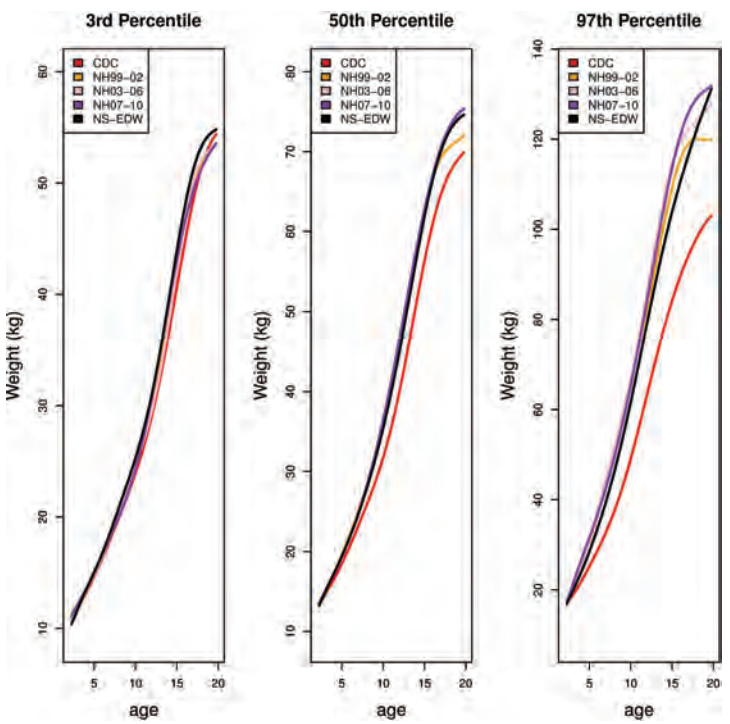


Figure 11.10. Selected weight-for-age percentiles.

11.3.1. *Stature-for-age*

The comparison of stature-for-age charts of the CDC, NS-EDW, and NHANES data for boys and girls show that these curves do not differ much, i.e., the stature growth pattern of American children has been fairly consistent over the past 50 years (see Figs. 11.4 and 11.7, where the black curves were generated by the published LMS value from the CDC website [27], and the red curves were generated from NS-EDW and NHANES datasets).

11.3.2. *Weight-for-age*

The weight-for-age charts from both NS-EDW and NHANES datasets (red curves) differ substantially from the CDC data (black curves). The curves generated from NS-EDW and NHANES are shifted upward, indicating a progressive increase in American children's weight through these years for both girls and boys. The upward shifts also became more and more significant with age. The accelerating trends for boys' weight in the NS-EDW and NHANES data were similar, while the girls' weight in the NHANES data increased even more significantly than in the NS-EDW data (see Figs. 11.3 and 11.6).

11.3.3. *BMI-for-age*

The BMI-for-age charts from both NS-EDW and NHANES datasets (red curves) also differed from the CDC data (black curves). As previously mentioned, the children's stature growth pattern did not change much, while the weight growth pattern accelerated substantially. A similar acceleration of BMI is, thus, expected since BMI is directly proportional to weight. The shape of the BMI-for-age curves of both NS-EDW and NHANES datasets also changed. In childhood, BMI typically increased through the first year of life, dropped off and declined to a minimum later in childhood, and then increased again, as shown in the CDC BMI-for-age charts. This phenomenon is known as adiposity rebound (AR) [28, 29].

However, as shown in the NHANES and NS-EDW curves, there was no obvious AR phenomenon for the 95th and 97th percentiles of children aged 2 to 20 years. Unlike the CDC data, the nadirs of those percentiles were not very obvious. This might partly be due to the insufficiency of accurate sampling for children aged 1 to 2 years. Furthermore, AR points for other percentiles also shifted to the left, which was consistent with the observations for the 95th and 97th percentiles (whose AR could be regarded as occurring even before the age of 2).

11.3.4. *LMS statistics*

Comparisons with respect to the L, M, and S parameters of weight, stature, and BMI data for boys and girls were conducted among the three data sources. Figures comparing the L, M, and S parameters of the CDC, NS-EDW, and NHANES weight-for-age for boys are shown as illustrative examples (see Figs. 11.11–11.13). The L parameter is the power in the Box-Cox transformation. Curves of L parameters are significantly different, reflecting different degrees

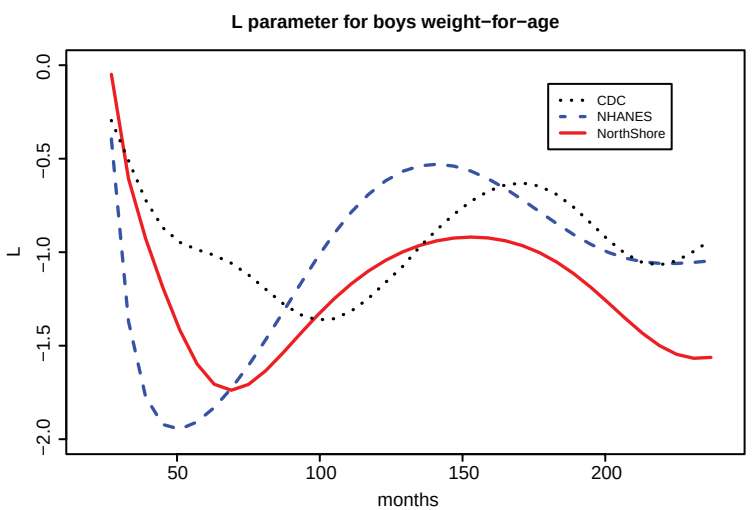


Figure 11.11. L parameters of boys weight-for-age.

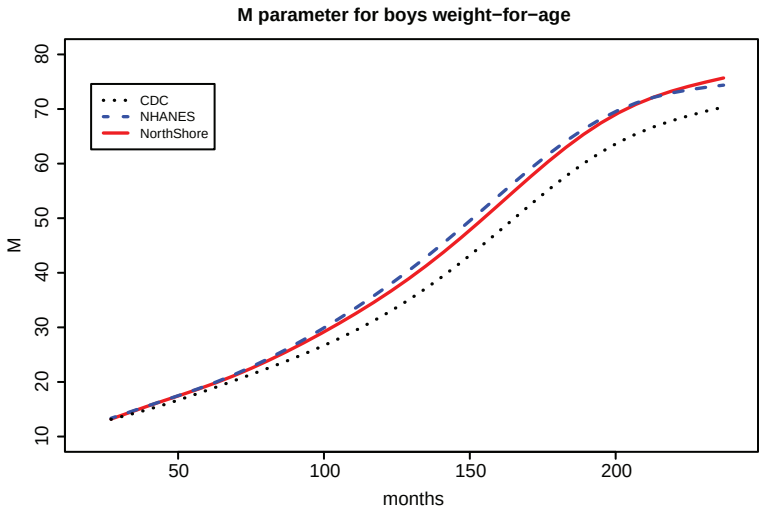


Figure 11.12. M parameters of boys weight-for-age.

Figure 11.13. S parameters of boys weight-for-age.

of skewness along ages among the three datasets. The M parameter stands for the median along ages. The median weight for boys is found to be considerably higher in the NHANES and NS-EDW datasets

than the CDC. The S parameter stands for the generalized coefficient of variation. The shapes of the S curves of the three data sources are similar, and the NHANES data show the highest coefficient of variation, while the CDC data have the lowest. The values of the L, M, and S parameters that were generated after transformation stages are not shown in detail.

11.3.5. *Curve analysis*

The smoothed percentile curves for boys and girls from the NHANES, NS-EDW, and CDC datasets were compared graphically as shown in Figs. 11.3 through 11.8. Moreover, attention is drawn to the extent of the differences between the shape or growth trajectory of children growing up in recent years and the CDC reference standard curves. Tables 11.4 through 11.7 summarize the BMI percentile shifts for boys in NS-EDW and NHANES compared to the CDC at ages 4, 9, and 15 years. These findings indicate an upward shifting trend of percentiles in the two more recent datasets. For example, for boys aged 9 and 15 years, the 50th BMI percentile in

Table 11.4. BMI percentile shift (Boys): Comparing NS and CDC.

Boys	Age 4	Age 9	Age 15
NS Percentiles	CDC Percentile	CDC Percentile	CDC Percentile
3	below 3	10	5
5	3	10	10
10	10	25	25
25	25	50	50
50	**50**	**75**	**75**
75	90	90	90
85	95	95	95
90	97	97	97
95	above 97	above 97	above 97
97	above 97	above 97	above 97

Table 11.5. BMI percentile shift (Girls): Comparing NS and CDC.

Girls	Age 4	Age 9	Age 15
NS Percentiles	CDC Percentile	CDC Percentile	CDC Percentile
3	below 3	10	5
5	3	10	10
10	10	25	25
25	25	50	50
50	**50**	**75**	**75**
75	85	90	85
85	90	95	95
90	95	97	97
95	97	above 97	above 97
97	above 97	above 97	above 97

Table 11.6. BMI percentile shift (Boys): Comparing NHANES and CDC.

Boys	Age 4	Age 9	Age 15
NHANES Percentiles	CDC Percentile	CDC Percentile	CDC Percentile
3	3	10	5
5	5	10	10
10	10	25	25
25	25	50	50
50	**50**	**75**	**75**
75	85	90	85
85	95	95	95
90	97	97	97
95	above 97	above 97	above 97
97	above 97	above 97	above 97

Table 11.7. BMI percentile shift (Girls): Comparing NHANES and CDC.

Girls	Age 4	Age 9	Age 15
NHANES Percentiles	CDC Percentile	CDC Percentile	CDC Percentile
3	3	5	5
5	5	10	10
10	10	25	25
25	25	50	50
50	**50**	**75**	**75**
75	85	90	90
85	95	97	97
90	97	above 97	above 97
95	above 97	above 97	above 97
97	above 97	above 97	above 97

the NS dataset is located around the 75th BMI percentile in the CDC dataset. Additionally, Table 11.8 shows the age in months where the minimum BMI for boys is reached in the CDC dataset, and shifts from this standard can be seen in the NHANES and NS datasets. In the two more recent datasets, the mimimum BMI for each percentile is reached at earlier ages. For instance, at the 90th percentile for boys, the minimum BMI occurs at month 57 in the CDC dataset, at month 33 in the NHANES dataset, and at month 27 in the NS dataset. Moreover, Table 11.9 shows the ascending trends in the minimum BMI values, comparing NS and NHANES data with the CDC data.

11.4. Conclusions

The CDC growth curves, which were last updated in 2000, are used as a reference standard and are not intended to reflect children's current growth trends. In this study, a software program that generates

Table 11.8. Age (months) shift of minimum BMI.

Percentiles	3	5	10	25	50	75	85	90	95	97
Boys										
CDC min (months)	75	75	75	75	69	65	57	57	51	45
NH min shift (months)	0	0	6	12	12	20	18	24	24	18
NS min shift (months)	12	12	12	18	12	20	18	30	14	18
Girls										
CDC min (months)	69	69	69	63	63	57	51	51	39	39
NH min shift (months)	0	0	0	0	6	12	12	18	6	12
NS min shift (months)	12	12	12	6	12	12	12	18	12	12

growth curves using the CDC's method was developed and successfully applied to two other datasets (NS-EDW and NHANES). By comparing the CDC's reference curves to the growth curves from more recent datasets, it is confirmed that the CDC curves do not accurately reflect the weight or BMI of today's American children. It has been shown that children in the NS-EDW and NHANES 1999–2010 datasets are heavier at any given age compared with children in the CDC dataset. In addition, adiposity rebound [30] occurs at an earlier age, or may not even exist, in these two groups of children. These findings suggest a progressive fattening of American children, and the growth charts generated in the past as standards for measuring growth might no longer be applicable to today's population.

11.4.1. *Significance*

The significance is two-fold. First, this work successfully reproduced the CDC's methodology for the creation of growth curves and developed a straightforward algorithm and software program

Table 11.9. Minimum BMI value for each percentile.

Percentiles	3	5	10	25	50	75	85	90	95	97
Boys										
CDC BMI (kg/m^2)	13.35	13.63	13.994	14.67	15.40	16.29	16.88	17.38	18.18	19.01
NHANES BMI (kg/m^2)	13.53	13.80	14.17	14.92	15.88	16.99	17.67	18.17	18.86	19.34
NS BMI (kg/m^2)	13.53	13.81	14.21	14.95	15.83	16.90	17.51	17.94	19.09	19.97
Girls										
CDC BMI (kg/m^2)	13.04	13.31	13.66	14.33	15.14	16.15	16.85	17.46	18.36	18.93
NHANES BMI (kg/m^2)	13.22	13.46	13.90	14.63	15.61	16.79	17.62	18.17	19.14	19.46
NS BMI (kg/m^2)	13.19	13.44	13.89	14.62	15.61	16.67	17.34	17.87	18.55	19.35

that can be applied to any population of children whose growth data (weight, height, etc.) had been recorded during ages 2–20. The CDC curves do not necessarily reflect normal growth in sub-populations of children, and this is important as USA is a racially and ethnically diverse country. Secondly, meaningful comparisons between today's children and the children whose cross-sectional data make up the CDC curves have been provided. For example, based on NS-EDW data, the 50th percentile of BMI-for-age for girls is nearly equivalent to the 80th percentile on the corresponding CDC curve. These observations provide a unique way to quantify the scope of the obesity epidemic among children.

11.4.2. *Long-term goal*

The eventual goal of this study is to incorporate this program into EHR software. This tool is aimed to provide clinicians access to the overall profile of a certain population using the data collected from EHRs, allow clinicians to identify growth patterns in the population and compare with the national standards, and help clinicians determine the growth status of individual children in their practices. It is anticipated that such a tool will be useful for clinical practice, beneficial to both clinicians and parents, and helpful in conducting the study of children's growth patterns.

11.4.3. *Limitations*

NHANES data were based on accurate measurements recorded by trained researchers. NS-EDW data were collected in clinical settings and may be less accurate. Given that, the CDC methodology uses cross-sectional data, the creation of growth curves based on longitudinal data collected from NS-EDW is under investigation. NS-EDW data were largely collected from children in an economically advantaged region and may not be representative of children in USA. However, the data were found to be similar to that of more recent NHANES surveys.

Acknowledgment

This work is supported in part by NSF Grants No. CMMI-1233807 and CMMI-1536987. The authors would like to thank Cynthia Ogden for helping access the CDC data, John Komlos for helping access the NHANES data, and Chad Konchack and Justin Lakeman for extracting NS-EDW data. The authors have benefited from conversations with and comments from Katherine Flegal, Tim Sanborn, Kibaek Kim, Joyce Ho, Sanjay Mehrotra, Tony Solomonides, Yuan Ji, Jessica Chan, Nigel Parsad, and Jonathan Silverstein.

Author Disclosure Statement

No competing financial interests exist.

References

1. Owen, GM (1978). The new National Center for Health Statistics growth charts. *South Med J*, 71, 296–297.
2. WHO Multicentre Growth Reference Study Group (2006). WHO child growth standards based on length/height, weight and age. *Acta paediatr (Oslo, Norway: 1992), Supplement*, 450, 76–85.
3. Cole, TJ, JV Freeman and MA Preece (1998). British 1990 growth reference centiles for weight, height, body mass index and head circumference fitted by maximum penalized likelihood. *Stat Med*, 17, 407–429.
4. Cole, TJ and MJ Roede (1999). Centiles of body mass index for Dutch children aged 0–20 years in 1980 — a baseline to assess recent trends in obesity. *Ann Hum Biol*, 26, 303–308.
5. Cole, TJ (1988). Fitting smoothed centile curves to reference data. *J R Stat Soc*, 151, 385–418.
6. Cole, TJ and PJ Green (1992). Smoothing reference centile curves: The LMS method and penalized likelihood. *Stat Med*, 11, 1305–1319.
7. Flegal, KM and Cole TJ (2013). Construction of LMS parameters for the centers for disease control and prevention 2000 growth charts. *Natl Health Stat Report*, 63, 1–3.
8. Chen, C. *Growth Charts of Body Mass Index (BMI) with Quantile Regression*. Cary: SAS Institure Inc.

9. Kuczmarski, RJ, CL Ogden, SS Guo, LM Grummer-Strawn, KM Flegal Z Mei, R Wei, LR Curtin, AF Roche and CL Johnson (2002). 2000 CDC growth charts for the United States: Methods and development. National Center for Health Statistics. *Vital Health Stat*, 246, 1–190.
10. Flegal, KM, R Wei, CL Ogden, DS Freedman, CL Johnson and LR Curtin (2009). Characterizing extreme values of body mass index-for-age by using the 2000 Centers for Disease Control and Prevention growth charts. *Am J Clin Nutr*, 90, 1314–1320.
11. Cole, TJ, KM Flegal, D Nicholls and AA Jackson (2007). Body mass index cut offs to define thinness in children and adolescents: International survey. *BMJ*, 335, 194–202.
12. Rao, G (2013). A new tool needed for identifying and characterizing obesity. *Curr Cardiovasc Risk Rep*, 7, 85–87.
13. http://www.cio.com/cio100/detail/2263
14. http://www.healthcareitconnect.com/fall-2012-accountable-care-health-it-strategies-summit-preview-northshore-university-healthsystem/
15. http://www.healthcareitconnect.com/interview-steve-smith-cio-of-northshore-university-healthsystem/
16. http://www.cdc.gov/nchs/nhanes.htm
17. http://www.cdc.gov/nchs/nhanes/nhanesproducts.htm
18. Kuczmarski, RJ, MD Carroll, KM Flegal and RP Troiano (1997). Varying body mass index cutoff points to describe over-weight prevalence among US adults: NHANES III (1988 to 1994). *Obes Res*, 5(6), 542–548.
19. Ogden CL, Carroll MD, Kit BK, Flegal KM. "Prevalence of obesity in the United States, 2009–2010" NCHS Data Brief. 2012 Jan;(82):1–8.
20. Zhong, X, J Li and KP Unnikrishnan (2013). Analysis of children's growth patterns using the LMS method and CDC method. University of Wisconsin-Madion: Technique report, Dept. of Industrial and Systems Engineering.
21. http://www.r – project.org/
22. http://stat.ethz.ch/R-manual/R-patched/library/stats/html/quantile.html
23. http://stat.ethz.ch/R-manual/R-patched/library/stats/html/lowess.html
24. http://stat.ethz.ch/R-manual/R-patched/library/stats/html/lm.html
25. http://stat.ethz.ch/R-manual/R-patched/library/stats/html/nls.html
26. http://statistic-on-air.blogspot.com/2011/09/implementation-of-cdc-growth-charts-in.html
27. http://www.cdc.gov/growthcharts/percentile-data-files.htm

28. Rolland-Cachera, MF, M Deheeger, F Belisle, M Sempé, M Guilloud-Bataille and E Patois (1984). Adiposity rebound in children: A simple indicator for predicting obesity. *Am J Clin Nutr*, 39, 129–135.
29. Rolland-Cachera, MF, M Deheeger, P Avons, M Guilloud-Bataille, E Patois and M Sempé. Tracking adiposity patterns from 1 month to adulthood. *Ann Hum Biol*, 14, 219–222.
30. Taylor, RW, AM Grant, A Goulding and SM Williams (2005). Early adiposity rebound: Review of papers linking this to subsequent obesity in children and adults. *Curr Opin Clin Nutr Metab Care*, 8, 607–612.
31. Williams, SM (2005). Weight and height growth rate and the timing of adiposity rebound. *Obes Res*, 13, 1123–1130.
32. Janicke, DM, BJ Sallinen, MG Perri, LD Lutes, JH Silverstein and B Brumback (2009). Comparison of program costs for parent-only and family-based interventions for pediatric obesity in medically underserved settings. *J Rural Health*, 79, 319–325.
33. Waters, E, A de Silva-Sanigorski, BJ Hall, T Brown, KJ Campbell, Y Gao, R Armstrong, L Prosser and CD Summerbell (2011). Interventions for preventing obesity in children. *Cochrane Database Syst Rev*, 12, CD001871.

Index