ABSTRACT

Title of dissertation: APPLICATION OF MATHEMATICAL AND COMPUTATIONAL MODELS TO MITIGATE THE OVERUTILIZATION OF HEALTHCARE SYSTEMS

Xia Hu, Doctor of Philosophy, 2017

Dissertation directed by: Professor Bruce Golden
Professor Sean Barnes
Robert H. Smith School of Business

The overutilization of the healthcare system has been a significant issue financially and politically, placing burdens on the government, patients, providers and individual payers. In this dissertation, we study how mathematical models and computational models can be utilized to support healthcare decision-making and generate effective interventions for healthcare overcrowding. We focus on applying operations research and data mining methods to mitigate the overutilization of emergency department and inpatient services in four scenarios. Firstly, we systematically review research articles that apply analytical queueing models to the study of the emergency department, with an additional focus on comparing simulation models with queueing models when applied to similar research questions. Secondly, we present an agent-based simulation model of epidemic and bioterrorism transmission, and develop a prediction scheme to differentiate the simulated transmission patterns during the initial stage of the event. Thirdly, we develop a machine

learning framework for effectively selecting enrollees for case management based on Medicaid claims data, and demonstrate the importance of enrolling current infrequent users whose utilization of emergency visits might increase significantly in the future. Lastly, we study the role of temporal features in predicting future health outcomes for diabetes patients, and identify the levels to which the aggregation can be most informative.

APPLICATION OF MATHEMATICAL AND COMPUTATIONAL MODELS TO MITIGATE THE OVERUTILIZATION OF HEALTHCARE SYSTEMS

by

Xia Hu

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2017

Advisory Committee:
Professor Bruce Golden, Chair/Advisor
Professor Sean Barnes, Co-chair/Co-Advisor
Professor Margrét Bjarnadóttir
Professor Hector Corrada Bravo
Professor Paul Smith

Dedication

To my family.

Acknowledgments

First and foremost, I would like to thank my advisors Dr. Bruce Golden and Dr. Sean Barnes for their enduring support, indispensable guidance, and great vision. They introduced me to various challenging and interesting topics in healthcare, and gave me the freedom to explore my own research interests while providing valuable insight and wise counsel. I am deeply indebted to both of them for their careful edits that have made my research more readable and presentable. Besides advising me on my work, they remain exemplary on how to be dedicated and efficient researchers, and more importantly, upright and self-motivated human beings. My research experience with them helps to set me on a lifelong journey of exploration and learning as a scientist.

This dissertation would not have been possible without the help of Dr. Margrét Bjarnadóttir, who opened up opportunities for my two data mining projects. She provided inspiring insight during our collaboration and set an encouraging example as a female researcher with a great mind and work ethic. I am also very grateful to other members in my thesis committee—Dr. Paul Smith for his linear regression classes that constituted the fundamentals for my machine learning research, and Dr. Hector Corrada Bravo for serving as the deans representative on my committee. I thank all of them for sparing their valuable time reviewing this dissertation.

I wish to express my gratitude to the Director of the AMSC program, Dr. Konstantina Trivisa, who supported my study with several fellowships and encouraged me with heartfelt enthusiasm to pursue my goals during different stages of my

Ph.D. study. And I want to thank Dr. Lawrence Washington for always keeping his door open. I am also grateful to the following staff members who have offered me generous administrative help: Alverda McCoy, Jessica Sadler, Janet Cavanagh, and Celeste Regalado. The undergraduate students I have taught in the past six years as a math TA also deserve special thanks for acknowledging me with good teaching reviews and making my teaching experience rewarding.

I appreciate everyone I have met in UMD that makes me feel at home at College Park. I would like to thank my former housemates Wen Chen, Yue Dong, Junyi Shen, and He He, for their enjoyable company. I thank my officemates, especially Sean Ballentine, Ryan Hunter, and Rebecca Black, with whom I shared many good games and weekend outings. I appreciate all the career advice from Chen Dong and Lucas Tcheuko, and the assistance and friendship from many others at UMD.

I owe my deepest gratitude to my family for their constant support. I thank my parents, who both got their own Ph.D. while I grew up, for never pushing me to study but influencing me unnoticeably with their own diligence. Thank you for filling me with love and leading me to be the happy person that I want to be. I am also grateful to my parents-in-law, who contributed to my warm memories of many home-coming weekends and holiday celebrations in the past few years.

Finally, thanks to my husband and best friend Robert Maschal, for all the laughter, good meals, and motivations that we shared. This dissertation is not possible without you.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

A&E	Accident and emergency
ABMS	Agent-based modeling and simulation
AD	Ambulance diversion
AUC	Area under ROC curve
ARIMA	Autoregressive integrated moving average
ARMA	Autoregressive moving average
CCS	Clinical classifications software
CM	Case management
DCSI	Diabetes Complications Severity Index
DES	Discrete event simulation
DFT	Discrete Fourier Transformation
ED	Emergency department
EHR	Electronic health records
EMS	Emergency medical services
ER	Emergency room
FIFO	First-in, first-out
HIPAA	Health insurance portability and accountability act of 1996
ICD-9-CM	International Classification of Diseases, Ninth Revision, Clinical Modification system
IU	Inpatient unit
KS	Kolmogorov-Smirnov
LASSO	Least absolute shrinkage and selection operator
LOS	Length of stay
LWBS	Leave without being seen
MOL	Modified offered load
NHPP	Nonhomogeneous Poisson process
OL	Offered load
PSA	Piecewise stationary approximation
QT	Queueing theory
QED	Quality and efficiency driven
RCCP	Rough cut capacity planning
SIPP	Stationary independent period by period

Chapter 1: Introduction

The dramatic growth in healthcare system utilization has been a critical issue in the United States. Over the last half century, healthcare spending in the U.S. has steadily increased from 5% of GDP in 1960 to 17.4% in 2013 [33], and the percentage is projected to rise to 26% by 2035 [115]. Driven by the Patient Protection and Affordable Care Act—which aims to expand care access and coverage [182]—total healthcare spending in the U.S. has increased to $3.2 trillion, with $9,990 being spent per person on average [36]. As 64% of healthcare spending was paid for by the government (funded via programs such as Medicare, Medicaid, the Children's Health Insurance Program, and the Veterans Health Administration) [95, 136], this places a significant burden on the government's budget.

The overutilization of healthcare systems not only influences the government financially, but also impacts the whole society. Many consequences have arisen as a result, imposing risks on patients, providers, and individual payers (such as the insurance companies) [137]. From the perspective of patients, congestion in healthcare facilities results in prolonged lengths of stay, compromised patient safety, increased costs, and dissatisfaction [218, 53, 54, 222]; from the perspective of providers, healthcare overutilization puts a burden on the system and leads to staff stress, errors, and

morale issues [228, 77, 218]; and from the perspective of individual payers, extensive utilization of the healthcare systems increases expenditures significantly.

Motivated by political influences such as calls to action in the U.S. by the Institute of Medicine and the President's Council of Advisors on Science and Technology [193], researchers have proposed numerous methods to improve the system efficiency, promote patient health outcomes, and decrease healthcare costs. Classical operations research methods such as queueing theory (QT) [79] and simulation [185] have been extensively utilized to model patient flow and optimize resource allocation and forecasting in a healthcare setting [15]. Optimization methods such as linear programming and metaheuristics have been applied to treatment planning, resource scheduling, facility location, and organ donation and transplantation [194]. Statistical methods have been effectively used to analyze healthcare costs and utilization, for their ability to address positive skewness and heavy tails in non-negative data (such as costs or number of service claims) [162]. More recently, many efforts have been devoted toward taking full advantage of extensive electronic health records (EHR) data-so-called big data-to provide decision support for effective interventions. Facilitated by advancements in health information systems, analytical software, and computational power [126], data mining and machine learning techniques have been applied to identify high-risk patients [11], understand disease progression patterns [138], and predict diagnosis such as heart disease [178].

This dissertation is focused on the study of computational and mathematical models to support healthcare decision-making and generate effective interventions for healthcare overcrowding. The methods included in this dissertation are queueing

theory (QT), simulation, and various machine learning algorithms. QT is a classical operations research methodology that uses relevant mathematical models to obtain closed-form or recursive formulae to calculate performance metrics such as average queue length, average wait time, and the proportion of customers turned away [87]. Discrete event simulation (DES) imitates system behavior using the sequential execution of events, and offers great flexibility in testing various interventions in a virtual environment [185]. In the context of the emergency department (ED), a patient's stay includes events such as arrival, triage, diagnosis, treatment, and departure, with waiting occurring at any point in the process where all resources are currently being utilized. Patients are usually modeled as passive entities who consume resources such as physicians, nurses, and beds at different times during their stay. As a subset of simulation methods, agent-based modeling and simulation (ABMS) is a rapidly emerging methodology that determines system behavior through the aggregation of interactions among individuals or between individuals and the environment [146].

We are particularly interested in the mitigation of overutilization of ED and inpatient services, as both are generally very costly. For example, in the U.S., the average cost per inpatient stay was $10,000 in 2011 [188] and per ED visit was $2,000 in 2013 [2]. To decrease the frequency of inpatient and ED visits, it is essential to identify potential frequent utilizers (i.e., individuals who will use the healthcare system extensively in the future). Such frequent utilizers consume a disproportionate amount of healthcare resources; therefore, identifying them is critical for interventions such as case management in order to target them to potentially reduce their

usage of the healthcare system.

This dissertation is organized as follows. In Chapter 2, we systematically review research articles that apply analytical queueing models to the study of patient flow in the ED. QT has been extensively applied to general service settings (e.g., call centers). However, as the healthcare system differs from these other settings in terms of overall mission and complexity, the direct application of existing queueing models to healthcare is inappropriate. In this study, we not only examine ED queueing models from problem- and modeling-oriented perspectives, but also compare QT with simulation, focusing on the advantages and limitations of each approach when applied to similar research questions. We identify situations for which queueing models are more likely to obtain less realistic results than comparable simulation models, and highlight the latest methodological advancements in queueing theory, simulation, hybrid modeling, and big data. To our knowledge, this is the first systematic review of queueing theory focusing exclusively on ED operations, and also the first article comparing queueing models with simulation models in the application to similar ED problems. This work is originally published in *International Transactions in Operational Research* [103].

In Chapter 3, we apply ABMS to detect the outbreak of bioterrorism during its initial stages. Bioterrorism, namely the intentional release of viruses, bacteria, or other toxic biological agents, is a significant threat to the U.S. Early detection of a potential bioterrorism incident is vital for controlling diseases and limiting the damage. However, as some candidate bioterrorism diseases (e.g., anthrax and viral hemorrhagic fever) present symptoms in humans similar to those of common illnesses

(e.g., the flu), it is difficult to quickly distinguish between a bioterrorism outbreak and a natural disease. In this chapter, we propose an agent-based model to simulate the transmission patterns of diseases caused by bioterrorism attacks or epidemic outbreaks and to differentiate between these two scenarios. Focusing on a region of three cities, our goal is to detect a bioterrorism attack before a sizeable proportion of the population is infected. Further, we validate our epidemic simulation results using a two-phase equation-based model. Our results indicate that the aggregated infection and death curves in the region can serve as indicators in distinguishing between the two disease scenarios. We also conclude that for a bioterrorism outbreak, the bioterrorism source city becomes more dominant as the local working probability (p_L, defined as the probability of people working inside their home-cities) increases. By contrast, the behavior of individual cities for the epidemic model presents a "time-lag" pattern, especially when p_L is large. As p_L decreases, the individual city's dynamic curves converge as time progresses. This work is originally published in *Proceedings of the 2014 Winter Simulation Conferences* [102].

In Chapter 4, we present a novel machine learning framework—using Medicaid claims data—for effectively selecting enrollees for case management (CM). CM is expensive and operates under limited resources; therefore, it is essential for these programs to select individuals who will achieve improved health outcomes from their enrollment and, if possible, generate cost savings for the organization. Unlike traditional methods that only target frequent users who may or may not repeat their ED usage behavior, our approach selects members for enrollment based on their likelihood of frequent use and their potential benefit from the program. We de-

velop predictive models for two types of frequent users—"jumpers" whose current ED usage is low, but will increase significantly in the future, and "repeaters" whose ED usage remains consistently high. We propose a strategy to select optimal combinations of these two types of frequent users, and compare the cost effectiveness to a baseline approach that classifies frequent users only according to their aggregated usage of the healthcare system in the previous year. We demonstrate that the baseline approach works well for targeting potential repeaters, but it will not result in positive savings unless the CM program is very effective in reducing ED usage. Including jumpers helps to improve cost effectiveness because machine learning models predict jumpers better than the baseline approach, and jumpers are more likely to achieve successful outcomes from participation in a CM program. This work is accepted to *IIE Transactions on Healthcare Systems Engineering* [101].

In Chapter 5, we discuss the role of temporal features in predicting future health outcomes for diabetes patients. Traditionally, temporal information in insurance claims data is aggregated to certain periodic levels (e.g., monthly, yearly) as input features for prediction models. However, the detailed, dynamic information of patients' disease progression or healthcare consumption patterns is lost. The main objective in this chapter is to examine the role of temporal information in improving diagnosis prediction. We examine various ways of incorporating more detailed temporal information, and analyze their effectiveness in predicting diabetic-related inpatient visits. Our results indicate that granular temporal features can better predict the risk of patient hospitalization than yearly aggregated features. In addition, compared to daily/monthly/quarterly aggregation, weekly aggregation is more

suitable for claims data analysis, as it preserves a substantial amount of information while reducing dimensionality.

Finally, we conclude with insights, contributions, and future directions. Readers should refer to the list of abbreviations for frequently used acronyms in this dissertation.

Chapter 2: Applying Queueing Theory to Emergency Department Operations

2.1 Introduction

Emergency Department (ED) overcrowding is an ongoing, critical challenge to operational efficiency in the United States [249]. Between 1996 and 2006, the number of ED visits per year in the U.S. increased by 32% to 119.2 million, while the number of EDs has decreased by 4.63% to 3,833 [191]. ED overcrowding has been associated with negative effects for both patients and providers. Patients seeking care in crowded EDs are subject to higher risks of morbidity and mortality [54], prolonged wait times [53], a higher likelihood of leaving without being seen by a care provider, and higher rates of dissatisfaction [53, 54, 222]. From the provider's perspective, overcrowding can lead to higher rates of medical errors [77], miscommunication, and stress, as well as lower productivity and morale [228]. In addition, overcrowding can have negative effects on the teaching mission in academic EDs and reduce the ability of EDs to respond to mass casualty incidents [44].

Queueing theory (QT) is a classical operations research methodology that uses relevant mathematical models to "obtain closed-form or recursive formulae that al-

low system designers to calculate performance metrics such as average queue length, average wait time, and the proportion of customers turned away" [87]. First studied by Erlang in 1913 in the context of telephone facilities, QT has been extensively utilized in industrial settings to analyze how resource-constrained systems respond to various demand levels, and thus is a natural fit for modeling patient flow in a healthcare setting [15, 179, 126]. Many researchers have used QT because the resultant closed-form solutions minimize data requirements and facilitate implementation in practice via spreadsheet models [43]. Such simplicity and speed enable QT to quickly evaluate system performance and compare alternatives for process improvement.

Queues are ubiquitous in general service settings (e.g., call centers), and a considerable body of research exists for these applications. However, healthcare settings differ from other service settings both in terms of their overall mission and complexity, thus making the direct application of existing queueing models inappropriate. Unlike call centers, whose priority is to attract and retain customers (measured by abandonment rate or mean number of revisits), the main purpose of the ED is to provide timely access to healthcare services by prioritizing the health of patients (measured by short wait time or small wait probability), given the criticality of the service offered [122]. In addition, the healthcare system is often more complex, with large variability in patient care pathways and processing times. Below we list a few characteristics of EDs that are more difficult to model than general healthcare settings (e.g., hospital inpatient units, primary care providers) or general service systems:

1. The rate of patient arrivals to the ED varies as a function of time [83, 157]. In a general healthcare setting, variability in demand can be mitigated through appropriate capacity planning [199], wait lists (e.g., for organ transplant or surgery), or suitable appointment systems [88]. The last two methods are not feasible for EDs; therefore, other control options must be leveraged to ensure timely service to the patients.

2. Patient flow throughout the ED and ancillary services, such as radiology or phlebotomy, can vary significantly from one patient to another [121]. Figure 2.1 provides an example of multiple care pathways for ED patients. Note that even for a fixed path, the service time, service protocols, and number of resources also vary along the way [87]. Such diversity in patient routes also makes the estimation of physician service times difficult, as these times are usually discontinuous due to physicians repeatedly ordering test results and waiting before deciding on the next course of action [83].

3. ED patients are prioritized and treated according to their assessed triage level (e.g., based on urgency or complexity), not according to their arrival time or a pre-determined schedule [76, 164]. Typically, patients with more severe conditions are treated sooner. In addition, the patient triage level estimated by nurses at presentation may not be accurate (with misclassification rates as high as 25%), which complicates the analysis even further [205].

4. EDs operate on a different time scale than other service systems, making the direct application of long-run, steady-state analysis inappropriate. As physi-

cians' service times are usually long in the ED (up to hours, versus minutes in call centers and teller systems in banks), time variability in arrival has a more compounding effect and stationary approximations cannot always be applied directly. In addition, the ED system is slow to converge to steady state in practice; therefore, average performance may not be realized and direct applications of steady-state assumptions may result in problematic solutions [83].

5. The ED often interacts with other units and departments within or around the hospital, such as inpatient units (e.g., general wards, cardiac wards, intensive care units, and post-anesthesia care units), service departments (e.g., catheterization lab, surgery, interventional radiology, and internal medicine), and other EDs through ambulance diversion [27, 87, 6]. Such interactions usually have compounding effects on ED wait times and performance.

6. The access blocking issue can be more complex in the ED system. Insufficient bed capacity in the ED, inefficient outpatient planning, and inadequate admission intensity from the ED to the wards can all cause blockage for patients into or out of the ED [143]. Luo et al. [143] reported that more than 40% of the admitted patients experienced ED blocking in a metropolitan hospital in Australia, and Schneider et al. [207] reported that such patients account for 22% of the total ED patient census in the U.S. For instance, blocking of beds (e.g., when beds are fully utilized) in the wards can cause boarding in the ED, whereby patients who are ready to transfer do not have access to a bed in the

destination unit due to the lack of bed availability. The boarded patients are at risk because the specialized services they require are not usually available in the ED. For example, Liu et al. [140] reported that 28% of patients boarded in the ED experienced some type of adverse event. In addition, the boarded patients are consuming ED resources (such as beds and medical staff) that can be utilized for other patients who are still waiting [137, 30, 44].

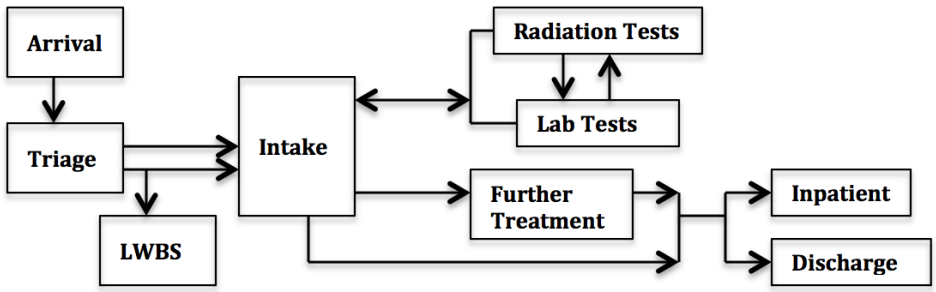

Figure 2.1: General patient flow diagram for the emergency department (Note that LWBS refers to patients who leave the ED without being seen by a physician or other care provider).

These characteristics complicate the direct application of general queueing models to the ED system, as it is theoretically and computationally challenging to develop an accurate queueing model for this system. Therefore, detailed simulation models are often used to generate results that closely agree with the observed performance [15].

There are several relevant survey articles examining the application of QT to healthcare. However, most of them either focus on applications of QT in a broad

range of operational or healthcare settings (e.g., emergency cardiac ward, intensive care unit, inpatient unit (IU), entire hospital, or general operational research settings) [192, 172, 67, 177, 79, 179, 15, 126], or examine operations in the ED using a variety of methods (e.g., regression models, time-series analysis, QT, and discrete-event simulation) [185, 249, 204] . Defraeye and Van Nieuwenhuyse [51] reviewed the approximation of time-varying systems by stationary queueing models in the ED, but they only focused on the staffing level problem. In addition, most methods discussed in this article were not supported by ED applications, as some of their selected literature was simply focused on general healthcare, industry, or theoretical settings. Saghafian et al. [204] comprehensively reviewed the contribution of operations research and management science to ED problems. However, they summarized QT applications more generally (and along with several other common operations research methodologies), and did not compare QT with simulation when applied to similar problems. Furthermore, many articles claim to use QT without distinguishing simulation-based queueing models from analytical models requiring traditional QT [18, 147]. To our knowledge, there is no detailed review of analytical QT models focusing exclusively on ED operations; hence, the motivation for this study. In addition to surveying QT applied to ED settings, we also examine how QT compares with other methods used in this context, specifically simulation approaches. The aim of this review is to highlight the contributions of QT and its uniqueness compared to simulation, as well as describe key trends of its application in ED settings. We find that queueing models provide important insights into ED operations, but they also have significant limitations when compared with other

modeling techniques.

The remainder of this chapter is organized in the following manner. Section 2.2 describes our search strategy for the survey and provides descriptive summaries of the selected articles, including the most commonly used performance measures. Section 2.3 examines ED queueing models from a problem-oriented perspective, whereas Section 2.4 characterizes them from a modeling-oriented perspective. Section 2.5 compares QT and simulation—with a focus on the advantages and limitations of each approach when applied to similar research questions—and includes a comparison of data acquisition and challenges for each method. We recommend those who are already familiar with the ED QT research to skip to Section 2.5. In Section 2.6, we summarize insights gained from these studies, highlight any limitations, and provide some directions for future research.

2.2 Survey Methodology and Literature Summary

2.2.1 Article Selection

This review examines 48 articles published since 1970 that apply analytical queueing models to the study of the ED. We use a three-stage approach to identify these relevant studies. In the first stage, we search the ACM Digital Library, Proquest, INFORMS, IEEE, PubMed, Science Direct, and Medline databases from 1970 to 2015, as well as Winter Simulation Conference Proceedings since 2000. We also include relevant Master's and Doctoral theses and working papers. These sources represent a comprehensive body of literature within the computer science, math-

ematics, operations management, operations research, engineering, and healthcare fields. In this stage, we identify the papers with "queueing", "queuing", or "queue" in the title, keywords, or abstract and one of the phrases "emergency department (ED)", "emergency room (ER)", or "accident and emergency (A&E)" in the abstract. In the second stage, we include papers that meet the following criteria: (1) the paper describes a queueing model based on a mathematical formulation and analysis, and (2) the paper calibrates a QT model to a specific ED environment (and possibly surrounding departments) in order to inform decision-making or improve operational efficiency. In other words, we focus on analytical applications of QT—not simply queueing analysis based upon simulation experiments—on operations conducted within the ED or a directly connected department. Applications of QT to general hospital departments and other clinical units are not included in our survey. In the third stage, we examine the references of the articles retained from the second stage and include any additional relevant articles.

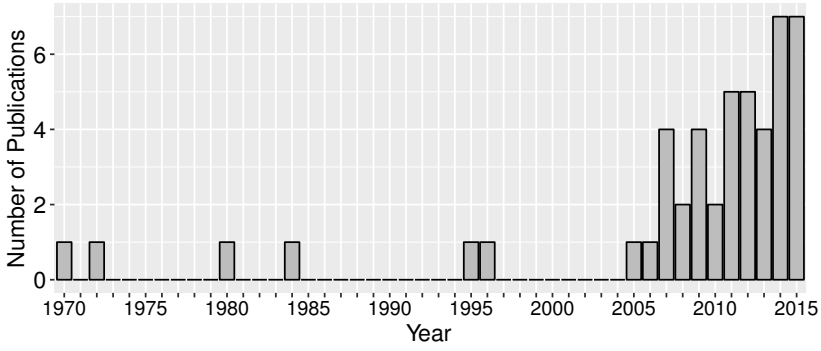

Figure 2.2: Number of articles published per year that apply queueing theory to emergency department operations.

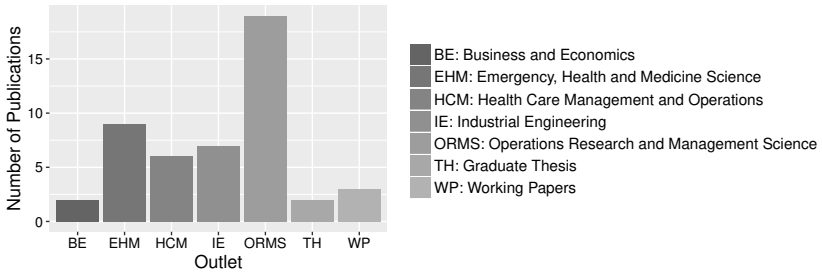

Figure 2.3: Number of publications by outlets.

2.2.2 Descriptive Analysis

In Figure 2.2, we summarize the number of selected publications in our survey as a function of the year of publication. We observe that the number of publications is limited before 2007, whereas more than half of the total contributions appeared in 2011 or later. This trend illustrates the increasing interest of researchers in this domain, which has been motivated by political influences such as calls to action in the U.S. by the Institute of Medicine and the President's Council of Advisors on Science and Technology [193] and facilitated by advancements in health information systems, analytical software, and computational power [126].

In Figure 2.3 , we summarize by research discipline the number of QT articles published in journals and other publication outlets from 1970 to 2015. OR/MS journals have published the most QT-related ED articles, followed by the EHM and IE journals. To shed some light on the development and evolution of QT in ED, we also list the year of the first appearance of an ED QT article in each above field: BE (2012), IE (1972), EHM (1970), HCM (2007), OR/MS (2007). We observe that ED QT methods are attracting increased attention from operations research, traditional

healthcare areas, engineering, and healthcare management (in that order).

2.2.3 Summary of ED Performance Measures

Defining a set of performance measures that can best capture the primary outcome is important for evaluating any operational interventions and decisions. There are numerous ways of choosing appropriate ED performance measures [244]. We list in Table 2.1 the most commonly used performance measures within the surveyed QT articles. For a definition and discussion of specific measures, please refer to Appendix 2.A.

From Table 2.1, we observe that expected wait time has been the most widely used measure for ED performance, followed by length of stay and wait probability. It is worth noting that the same paper may use several measures to independently or jointly evaluate service performance. For instance, Saghafian et al. [205] uses the weighted average of LOS (for discharged patients) and expected wait time (for admitted patients) to measure the effectiveness of various patient streaming models. It is also worth noting that in an ED queueing system, optimizing the timeliness of service—best reflected by patient wait times or rates of abandonment—and the utilization of resources (e.g., doctors, nurses, beds) are conflicting goals. Providers and administrators in the ED are constantly attempting to balance the tradeoff between these two objectives.

Research focusing on studying performance measures can even be used to evaluate governmental policies. For instance, Mayhew and Smith [156] used a queueing

model to evaluate the length of stay in A&E departments in the U.K. in light of the government-mandated target of serving and discharging 98% of patients within 4 hours. They demonstrated how the model could be used to evaluate the practicality of A&E targets. They found that without some form of patient flow re-designation, the current target would be unachievable. Furthermore, the authors found that the target was so ambitious that the integrity of reported performance was questionable.

2.3 Problem-oriented Perspective

In this section, we review the analysis of queueing models from the perspective of ED-specific management problems. In the U.S., EDs must be able to provide timely and efficient care in order to continue attracting patients to their services, as well as guarantee the well-being of patients. Efficient patient flow is characterized by high patient throughput and short lengths of stay, while simultaneously maintaining sufficient resource utilization rates and minimizing staff idle time [111]. Two primary factors that impact patients in the ED are patient arrival rates and resource capacity. Therefore, we first classify ED QT research into two problem-specific subgroups, namely from the demand (i.e., patient) and supply (i.e., resource) perspectives. Figure 2.4 describes all problems that we will discuss in this section. For the rest of the chapter, we employ Kendall's notation [116] to describe queueing models. Please refer to Appendix 2.B for details on Kendall's notation and common definitions in QT.

Table 2.1: ED performance measures used in the surveyed articles.

ED Performance Measures		Papers	Count
Time	Expected wait time	Almehdawe et al. [7], Broyles and Cochran [30], Cochran and Roche [43], Haussmann [92], Komashie et al. [120], Lin et al. [137], Madsen and Kofoed-Enevoldsen [147], Saghafian et al. [205, 206], Sharif et al. [212], Silberholz et al. [216], Taylor and Templeton [226], Vass and Szabo [235], Yom-Tov and Mandelbaum [255].	14
	Length of stay (LOS)	de Bruin et al. [48, 49], Mandelbaum et al. [150], Mayhew and Smith [156], Saghafian et al. [205, 206], Siddharthan and Jones [215], Zeltyn et al. [258].	8
	Expected boarding time	Broyles and Cochran [30], Lin et al. [137].	2
	Fraction of time on diversion	Allon et al. [6]	1
Queue	Average queue length	Almehdawe et al. [7], Madsen and Kofoed-Enevoldsen [147], Silberholz et al. [216], Vass and Szabo [235], Yankovic and Green [254], Zonderland et al. [259].	6
	Leave without being seen (LWBS) rate	Cochran and Broyles [42], Green et al. [84], Wiler et al. [248]	3

continued on next page

continued from previous page

ED Performance Measures		Papers	Count
Probability	Wait probability	Allon et al.[6], de Vericourt and Jennings [50], Izady and Worthington [108], Maman [149], Yom-Tov and Mandelbaum [255], Zeltyn et al.[258].	6
	Area overflow probability	Au et al. [17], Cochran and Roche [43], Taylor and Templeton [226].	3
	Blocking probability to inpatient unit	Lin et al. [137]	1
	Probability of adverse events	Saghafian et al. [206]	1
Resource	Resource uilization	de Bruin et al. [49], Mandelbaum et al. [150], Yom-Tov and Mandelbaum [255], Zeltyn et al. [258].	4
	Additional resource requirements	Palvannan and Teow [179]	1

2.3.1 Demand-oriented Problems

From the demand perspective, we categorize the ED QT literature according to three dimensions based on the patient's position in the system: (1) patient arrival, (2) patient flow through the ED, and (3) discharge and departure. In Table 2.2, we summarize the literature that applied QT to problems in emergency departments related to demand management.

Table 2.2: ED QT literature on demand management.

Arrival Management	**Arrival Pattern**	Green et al. [84], Zeltyn et al. [258], Yom-Tov and Mandelbaum [255]
	Ambulance Diversion	Taylor and Templeton [226], Au et al. [17], Hagtvedt et al. [89], Enders [62], Deo and Gurvich [52], Gupta [87], Allon et al. [6], Almehdawe et al. [7], Xu and Chan [252]
Patient Flow Management Priority Queue		Cochran and Roche [43], Fiems et al. [65], Haussmann [92], Huang et al. [105], Lin et al. [137], Panayiotopoulos and Vassilacopoulos [180], Roche and Cochran [201], Saghafian et al. [205, 206], Sharif et al. [212], Siddharthan and Jones [215], Stanford et al. [223], Taylor and Templeton [226], Zayas-Caban et al. [257].
Departure Management	**ED to IU**	Au et al. [17], Broyles and Cochran [30], Mandelbaum et al. [150], Allon et al. [6], Lin et al. [137], Yom-Tov and Mandelbaum [255](2014), Armony et al. [15], Zonderland et al. [259]
	LWBS	Roche and Cochran [201], Cochran and Roche [43], Cochran and Broyles [42], Wiler et al. [248], Zayas-Caban et al. [257], Xu and Chan [252]

2.3.1.1 Management of Patient Arrival

Arrival pattern

Motivated by the success of Erlang models applied to call centers [122], researchers in healthcare have sought out simple queueing models that best approximate the complexities of the ED. However, the simple queueing models often assume that patient arrivals stay constant over time, whereas, in reality, time-varying arrivals are observed in many ED systems. Figure 2.5 shows the hourly arrival rates recorded in an ED in New York City [84], where there is a peak at noon and a low arrival rate during the night. In reality, however, the situation can be more complicated, as patient arrivals may fluctuate around those expected arrival rates [83]. The uncertainty in demand may overburden the resources, leading to an overcrowding of patients waiting in the ED. In addition, arrival rates to the ED have an enduring effect over a patient's LOS (several hours forward), and occupancy levels can vary significantly during this time [15]. Figure 2.6 illustrates the time-lag between arrivals and occupancy levels in the ED [15]. Such a phenomenon can be explained by the time-varying version of Little's Law and renewal theory [25, 83].

Several approaches have been proposed to model the time-varying arrivals; among them are the modeling of a Nonhomogeneous Poisson Process (NHPP), and approximation methods such as Piecewise Stationary Approximation (PSA) and Stationary Independent Period by Period (SIPP) approaches [117, 109, 82, 83]. As a generalization of the ordinary Poisson process for which events occur randomly over time at a constant rate, NHPP allows for this rate to vary over time. Multiple

estimates have been developed to approximate the time-varying arrival rate for a NHPP, with piecewise-constant estimation being the most commonly used [133, 154]. Provided that the arrival rate of the NHPP is approximately piecewise-constant over a set of pre-determined intervals [117], the Kolmogorov-Smirnov (KS) test can be used to identify a NHPP by analyzing data from separate subintervals. Kim and Whitt [117] discuss scenarios for which the KS test fails, and offer strategies for coping with these failures. The PSA and SIPP methods divide the time-horizon into small intervals and estimate the staffing level for each interval by a time-invariant queueing process, assuming each interval is independent and the system is operating at steady-state conditions. However, PSA first estimates the staffing level required for each time point, then sets the overall staffing level to be the maximum of these staffing requirements over the time interval of interest. By contrast, SIPP first computes the mean arrival rate over the entire time interval, and then determines the averaged staffing level needed to serve this demand [82, 83]. The limitation of PSA is that it is most suitable if the staffing intervals are short, yet this is not always the case for the ED. Also, SIPP (and its variations) can result in overstaffing, particularly for the high-volume weekdays [83]. In order to model the time-lag between arrivals and occupancy, there are lagged variants of the aforementioned models (i.e., Lag-PSA, Lag-SIPP) that align these components in order to satisfy steady-state conditions [61, 81]. We will discuss these models in detail in Section 2.3.2.1.

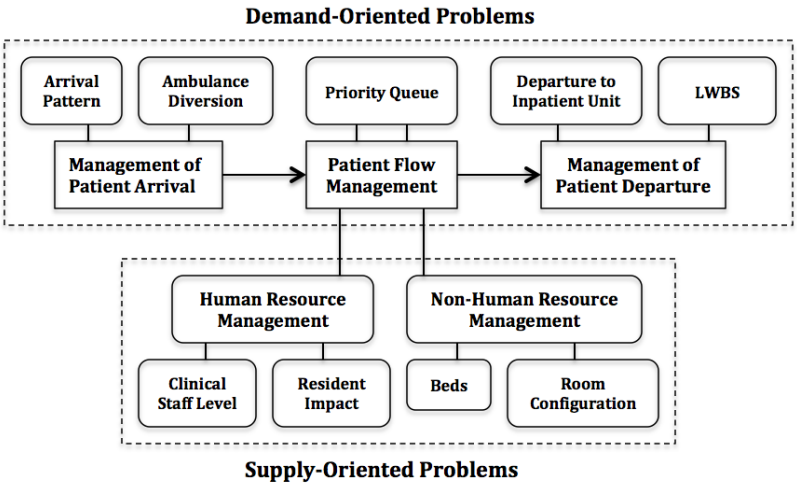

Figure 2.4: ED QT problem overview.

Ambulance diversion

Motivated by the variability in patient arrivals and the time-lag between arrival rates and occupancy, researchers have proposed several remedies to manage fluctuations in demand; among them are adaptive staffing [83, 64, 258, 255] and admission control policies such as ambulance diversion [52]. Adaptive staffing refers to matching staffing levels to accommodate variations in arrival patterns, and we will discuss this approach in detail in Section 2.3.2.1. Ambulance diversion (AD, or ambulance bypass) is a practice commonly adopted to alleviate ED congestion [174, 31], under which EDs request Emergency Medical Services (EMS) to divert incoming ambulances to hospitals nearby during periods of overcrowding [189]. The EMS agency will accept this request if not all neighboring EDs are diverting ambulances at the same time [52]. While AD can decrease the load on an ED, it potentially puts patients at risk of worse outcomes [208] and leads to lost revenue for the hospitals [158]. Almehdawe et al. [7] investigated a regional Emergency Medical Services (EMS) provider interacting with multiple EDs. Using a queueing

network, they studied the offload delays (i.e., delays in care caused by an ED's lack of available beds for incoming ambulance patients) and wait times for ambulance patients when the walk-in patients are also present.

To reduce the ED service load while maximizing the revenue, Hagtvedt et al. [89] modeled a baseline ED without AD as an $M/M/\infty$ queueing system, and then compared this system with one with a dynamic AD policy based on three thresholds ($M < K < N$), where M is the number of patients when diversion is unnecessary and K represents the number of patients in the system where a partial diversion strategy is activated, for which the hospital could selectively receive patients. Hence N is the total number of beds in the ED, and if the number of patients surpasses N, then the ED is enforced to divert all patients (e.g., full diversion) until the number of patients in the system falls to M. The authors used an ergodic continuous-time Markov chain, for which the states of each hospital were represented by the number of occupied beds, and whether or not the hospital was on full diversion. The authors suggested the potential for cooperative strategies among hospitals and the need for a centralized form of ambulance routing. Xu and Chan [252] investigated a proactive diversion strategy based on QT to utilize the predictions and proactively divert patients before congestion forms. They demonstrated that for all traffic intensities, the proposed strategy quantifies the "noise tolerance" and shortens wait times, while ensuring that the total rate of diversion and LWBS does not exceed those in the standard policies used in practice.

Studies have shown that failing to move patients from the ED to an IU (i.e., bed blocking) is the major cause for AD [17, 6]. Au et al. [17] studied this linkage

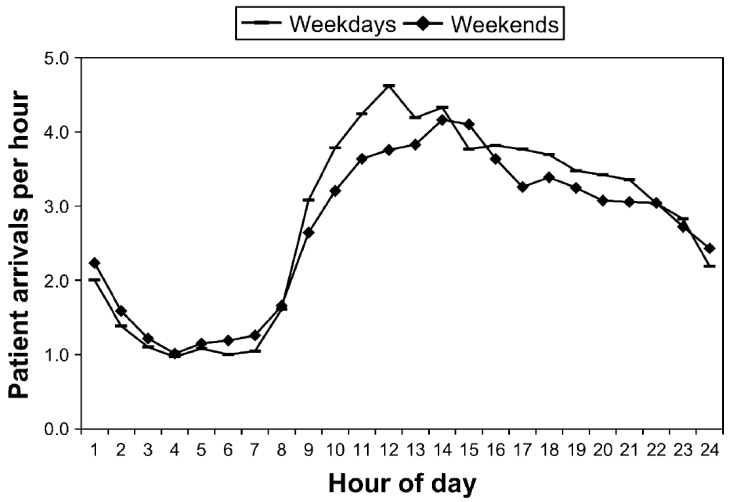

Figure 2.5: Arrival rates at NY Emergency Department (reproduced from [84] with permission).

by modeling the ED as a queue for treatment. Arrival rates to the treatment queue were assumed to be non-stationary. Given the current time and number of patients in the queue, the authors computed the conditional probability of reaching some predetermined maximum capacity level by time t, and then compared the observed and expected AD frequencies under various capacity constraints. Allon et al. [6] also studied the impact of hospital size and occupancy on the use of AD. In contrast to Au et al. [17], they used two sequential queueing models, and found that the capacity of the inpatient unit was negatively correlated with the fraction of time when the ED diverted ambulances. In other words, excess capacity in the inpatient unit leads to decreased ED diversion. In addition, the authors found that the minimum number of beds—defined as the threshold for AD—was positively correlated with the fraction of time spent on diversion.

In contrast to operations within a single hospital, AD is practiced across a

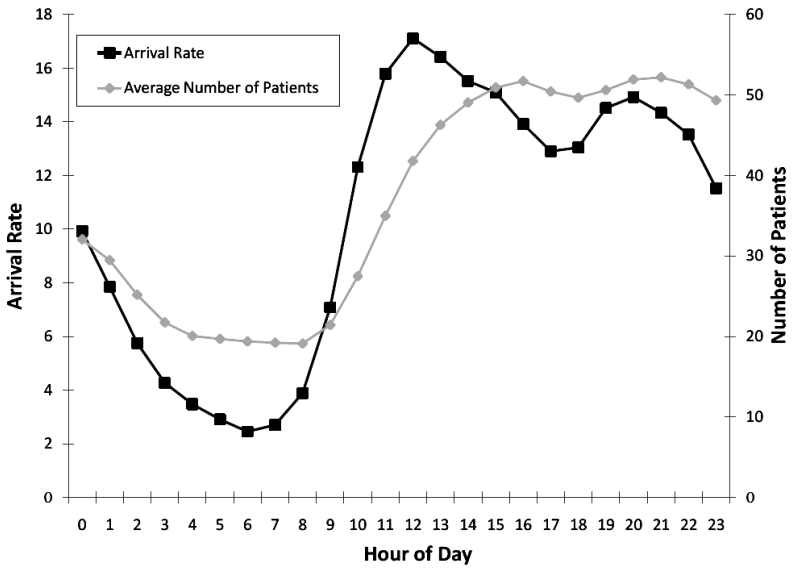

Figure 2.6: Arrival rates and average number of patients in the system by hour of the day (reproduced from [15] with permission).

network of EDs. Deo and Gurvich [52] studied a coordinated diversion mechanism between two EDs, and modeled these EDs as independent $M/M/c$ queues. By identifying the existence of a defensive equilibrium, wherein each ED stops accepting diverted ambulances from the other, the authors found that individual diversion decisions lead to poor resource pooling. This defensiveness results in the isolation of resources in the network and increased delays in comparison with those observed under coordinated diversion.

Most of these applications employ simple queueing models to describe the healthcare system (e.g., $M/M/c$), assuming that the arrival and service rates are independent of the system state (e.g., occupancy); yet they do not reflect reality very well [15]. For example, Figure 2.7 illustrates how arrival and service rates vary as a function of occupancy in an $M/M/1$ system [15]. As we can see, arrival and service

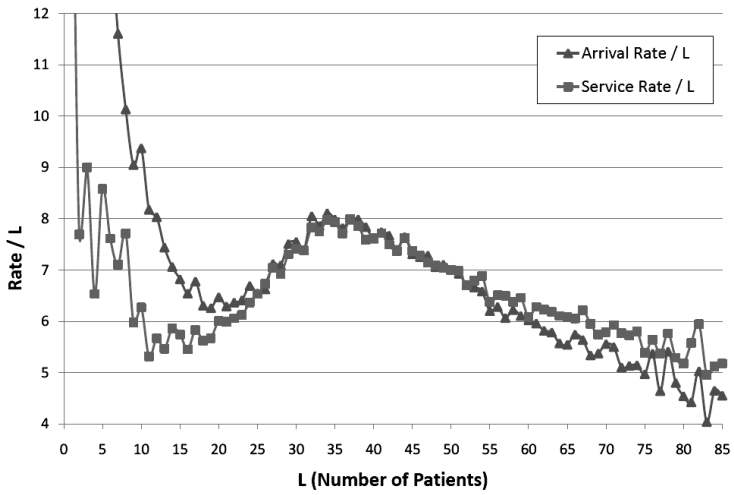

Figure 2.7: Arrival rate and service rate as a function of number of patients (reproduced from [15] with permission).

rates fluctuate with the number of patients in the system. Such a phenomenon is due to the fact that a high occupancy level can lead to increased rates of LWBS and AD, and varied rates of service [15]. In an interview-based study from two EDs in Sweden, registered nurses reported higher perceived efficiency and higher job satisfaction when the patient load was high and multitasking was needed [69]. Kc and Terwiesch [114] demonstrated that in a more general hospital environment (e.g., patient transport service and cardiothoracic surgeries), the servers might accelerate their service rate as the workload increased at first, but eventually present lower efficiency and reduced quality of care.

2.3.1.2 Patient Flow Management

In healthcare settings without appointment systems, the queueing discipline is either first-in, first-out (FIFO) or prioritized according to assessed patient classes

[67]. The queueing discipline is an important intervention that may significantly affect a patient's wait time and ultimate health outcome. In the baseline case, patients arriving at an ED are first assigned a triage number, color, or letter to reflect their severity or priority. Patients with a higher priority will usually be treated sooner. For example, triage numbers for the Emergency Severity Index system used in most U.S. EDs range from 1 to 5, with 1 being the most urgent [76]. However, such an arrangement means that patients with minor illnesses (i.e., patients with lower priority) will wait the longest. In order to balance between urgency and efficiency, several strategies have been implemented in the ED, which involve either splitting patient flow by acuity or by function [43]. For example, the fast track intervention adds a separate service stream—with dedicated beds served by a team of physicians and nurses—to serve non-urgent patients who require fewer resources and less complex treatments [46]. Studies have also shown that when utilization is high, wait times can be reduced by assigning higher priority to patients who require shorter service times [159].

QT has been used extensively in analyzing and comparing different queueing disciplines. Haussmann [92] studied the relationship between priority queues and patient wait times. Their study found that wait times for low priority patients increased when nurses were assigned more patients or a patient mix with more complex conditions. Taylor and Templeton [226] investigated a threshold service strategy for which beds are reserved for high priority patients. When the number of occupied servers exceeds a predetermined threshold, patients with low priority are rejected so as to keep the rest of the servers available for incoming patients with

high priority. They assumed Markovian arrivals and service rates, and considered two models when all servers were busy: one in which high priority patients queued for service and one in which these patients were diverted. They used the model to estimate the required number of ambulances to transfer both patient types based on the probability of all servers being busy and the wait times for the low priority patients. Fiems et al. [65] explored how emergency requests affect the wait times of scheduled patients with fixed service times. They modeled the system as a pre-emptive priority queue in which the emergency patients interrupt ongoing service of the scheduled patients. The primary effect was measured by the prolonged wait time in radiology of scheduled patients in an ED. Huang et al. [105] studied the prioritization by physicians of patients in triage and patients in process (i.e., who periodically demand the physician). They developed a multiclass queueing system with deadlines and feedback to model the flow of these respective patient classes in the ED, and proposed scheduling policies that attempted to balance the needs of these two groups. They established the asymptotic optimality of their policy under heavy ED traffic, and, additionally, developed some congestion principles that support forecasting of wait times and LOS. Zayas-Caban et al. [257] investigated the benefits of optimal control during an ED triage and treatment process. They studied how to prioritize the work of the providers to balance initial delays using a two-stage tandem queueing model with multiple servers for the triage and treatment processes with abandonments. Based on the optimal solution, they proposed new threshold-based policies as alternatives to priority rules.

There are some side effects of priority queueing. Siddharthan and Jones [215]

studied the increased wait times caused by non-emergency patients inappropriately seeking ED care. They proposed a first-in, first-out queueing model that reduced the average wait time; however, the wait time for higher priority patients was reduced at the cost of prolonged wait times from the lower-priority patients. A similar finding was presented in [137] for fast tracking. In this work, the authors explored the influence of the fast track on patient wait times and requirements for ED and IU resources. They found that although fast track shortened the overall wait time for patients from all priority classes, such a reduction was accomplished at the expense of increased wait times for patients from level three who were not eligible for the fast track. Therefore, a fast track could, in reality, decrease an ED's capacity to offer timely treatment for patients whose clinical conditions could potentially progress to a more serious level.

Split flow is a more recent approach that attempts to mitigate the aforementioned side effect of fast tracking. Unlike fast track, split flow reserves traditional beds only for high priority patients. Instead of having resources delivered to all patients, split flow requires the low priority patients to move to the resources (e.g., for diagnostic testing). To enlarge the ED's capacity to serve more patients, Cochran and Roche [43] investigated this novel ED design via a queueing network by incorporating hospital-specific characteristics in patient acuity mix, arrival patterns and volumes, and operational performance measures. They determined the required capacity of each area in the new split flow model and successfully decreased the LWBS rate. Using wait time and area overflow probability (i.e., the steady-state probability that the queue size exceeds a certain threshold) as performance targets,

they derived queueing equations that provided ED managers with real-time estimations of ED utilization. There are several other patient flow rules that are not merely based on patient acuity, but also based on a patient's appraised disposition, complexity of condition, or estimated wait time. Saghafian et al. [205] proposed a "virtual streaming" patient flow design in which patients are assigned to separate tracks based on predictions (by a triage nurse) of patients' final dispositions (admit or discharge). They provided a detailed queueing-based analysis on this design and investigated situations in which rooms and physicians could be shared across different tracks. They demonstrated that this design could achieve the benefits of both streaming and resource pooling when implemented properly. Further, Saghafian et al. [206] proposed a "complexity-augmented" triage rule, for which ED patients are classified on the basis of complexity (i.e., based on required resources) as well as urgency. Their results suggested that estimating the complexity of a patient prior to classifying his urgency leads to lower risk of adverse events and decreased LOS, even when the classification is subject to error. They also observed that it is more effective to stream patients first according to their complexity and then by urgency. Stanford et al. [223] studied a time-dependent priority queue, where a patient's priority is modeled as a linear combination of his time in the queue and triage class. They theoretically derived the wait time distribution for each class, under the constraints that performance targets specified for each class must be met. Sharif et al. [212] investigated the same problem, and numerically investigated how to choose feasible accumulation rates to satisfy specified performance objectives for multi-server, multi-class queues.

It is worth noting that the previous discussions on priority queues are based on the assumption that the triage scores of patients are accurate. In reality, patient triage estimated by the nurse is usually imperfect, and the true level of priority is usually not revealed until a physician sees the patient. Saghafian et al. [205] estimated the misclassification errors in the range of 20-25%. It is important to incorporate such uncertainty into ED modeling, as some conclusions may no longer hold when there are errors in classification [13]. We discuss this issue in more detail in Section 2.5.

2.3.1.3 Patient Discharge and Departure

In this section, we focus on two ways for patients to depart the ED: 1) being transferred to an inpatient unit or 2) leaving without being seen by a physician or nurse.

Patient departure to the inpatient unit

While most research on patient flow has focused on improving efficiencies within the ED, it is important to optimize the process externally as well. One example is bed blocking in the IU, which delays patients in the ED from transferring to the IU. Bed blocking in the IU has a compounding effect throughout the ED [137, 30, 44]. From the patient's perspective, such a delay can lead to an increased likelihood for clinical deterioration and patient dissatisfaction [145]. From the hospital's perspective, bed blocking inevitably aggravates congestion in the ED. Huang et al. [104] found that a significant proportion of admitted patients experienced

delays in transfer from the ED to the IU. Transfer patients waiting in the ED not only occupy critical resources such as ED beds, but also increase the workload of staff within the ED because some of them must be examined as frequently as every 15 minutes, according to ED regulations [15]. Such additional clinical treatment, in return, results in prolonged ED LOS, increased IU cost, and extended waits for subsequent ED patients [104].

Armony et al. [15] concentrated on the care pathways of patients in the ED and their association with transfer delays and fairness, where fairness is measured from the perspective of both staff and patients. Their data indicated that staff workload changes over a patient's stay, as patients typically require more attention during the initial part of their stay. From the patient's perspective, the FIFO rule is often violated in the process, with 45% of the patients being bypassed by another patient while waiting to be transferred from the ED to the IU. Broyles and Cochran [30] quantified the relationship between inpatient LOS and ED boarding and wait times via a QT-based statistical approximation. The authors concluded that a relatively small decrease in the hospital's inpatient LOS could cause significant reductions in ED boarding and ED waiting. Mandelbaum et al. [150] also studied the fair routing of patients from EDs to various IUs. They identified heterogeneity of LOS across different IUs, and investigated a routing scheme to account for these differences. In this scheme, a patient was routed to the IU that had the most number of available beds. They showed that this scheme was as asymptotically fair as the Longest Idle Server First (LISF) policy, but unlike this approach, their proposed scheme only required information available in the system. Lin et al. [137] estimated the wait

time during transfer from the ED to the IU. They found that the required ED capacity was inversely proportional to the size of the IU, and that an increase in the arrival rate of patients to the ED led to an even larger increase in the required capacity of the IU.

Patients leaving without being seen

The rate at which patients leave without being seen by a physician (LWBS) is one of the most important measures for evaluating ED performance [219]. Affected by the current queue length and the tolerance of patients, the LWBS rate characterizes the percentage of patients who are waiting and elect to forgo service due to their unwillingness to wait any longer. Such a phenomenon in the ED is equivalent to reneging or abandonment in QT, and QT is therefore a natural tool for modeling it. When the demand to the system is greater than the number of servers and dispatching to outside systems (such as ambulance diversion) is not available, reneging is the only mechanism that prevents an ED from being overwhelmed by demand [91, 83].

Roche and Cochran [201] found that diverting non-urgent patients to a dedicated fast track reduces the LWBS rate, as waiting for tests or test results consumes most of these patients' wait time. Cochran and Broyles [42] explored the relationship between the LWBS rate and ED utilization by approximating reneging using a queueing model of the ED with balking. They suggested utilizing patient safety (instead of the traditional measures such as LWBS rate) would be a more effective approach for determining the capacity of the ED. They also derived a relationship between the LWBS rate and the balking probability in an $M/M/1/k$ queue, which

helps to generalize the model results to other EDs. Wiler et al. [248] made the first attempt to predict patients who would abandon the queue based on patients' tolerance and ED crowdedness. They examined the influence of patient crowding on LWBS rates by approximating the $M/GI/c/s + GI$ model (i.e., parallel multiple servers with finite waiting room capacity and generally distributed patient wait time tolerance) with the established $M/M/c/s + M(n)$ model (where the patient wait time tolerance follows an exponential distribution related to the queue length). They observed that ED LWBS rates increase in an exponential way as the change rate of ED patient arrivals grows, and that shortened LOS and less patient boarding reduce LWBS rates.

2.3.2 Supply-oriented Problems

Faced with rising costs, ED administration boards are practicing cost containment by restricting resources for healthcare providers while maintaining quality care for patients. A large body of research has been devoted to the study of resource allocation. We divide the allocation of resources into two general areas: human (e.g., clinical and administrative staff) and non-human (e.g., beds, medical equipment, operating rooms) resource management.

Studies focused on ED resource planning can be classified into two types: (1) steady-state resource requirements and (2) short-term resource adjustments. The first type often uses QT, whereas the second type often involves adjustments by a manager to account for the demand fluctuations [91, 84, 51, 137]. Approaches com-

monly used in the second type include simulation models [19, 118, 258], time series models [209, 1, 152] and Markov Decision Processes [183, 230]. In this literature review, we examine articles that apply steady-state analysis for resource allocation using QT. Table 2.3 summarizes the literature that has applied queueing theory to the ED for the purpose of improving resource management.

Table 2.3: ED QT literature for resource management.

Human Resources	**Clinical Staff Level**	de Vericourt and Jennings [50], Green et al. [83, 84], Izady and Worthington [108], Komashie et al. [120], Maman [149], Panayiotopoulos and Vassilacopoulos [180], Saghafian et al. [206], Yankovic and Green [254], Yom-Tov and Mandelbaum [255], Zeltyn et al. [258].
	Resident Impact	Silberholz et al. [216]
Non-Human Resources	**Beds**	de Bruin et al. [48, 49], Gupta [87], Huang [106], Lin et al. [137], Saghafian et al. [206], Yankovic and Green [254].
	Room Configuration	Cochran and Roche [43], Mandelbaum et al. [150], Palvannan and Teow [179], Zeltyn et al. [258].

2.3.2.1 Human Resource Management

An important index for measuring ED service quality is its promptness of emergency care. Unfortunately, providing adequate staffing often proves difficult, as the demand for care can vary substantially throughout the day [83]. As Green

et al. [84] suggested, matching staffing levels to accommodate these variations is difficult for two reasons. First, variability in the arrival and treatment times for patients can cause significant delays even when the overall staff capacity is sufficient (i.e., greater than the average demand). Second, the magnitude of delays is difficult to predict directly from demand and resource levels.

Due to the time-varying nature of the ED, system parameters such as arrival rates are not constant. Therefore, traditional QT analysis is not directly applicable, as the steady state of the system cannot be achieved over these short periods [108]. In order to deal with the variation in patient arrivals, researchers have implemented several techniques to transform the varying arrival rate into a stationary service rate for the system [51]. In the following subsections, we discuss the use of QT to determine appropriate clinical staffing levels (including residents) and their impact on ED performance measures.

Nursing plays a significant role in determining hospital costs, care quality, and patient satisfaction [113]. The inadequate supply of nurses is associated with medical errors and ED overcrowding [74]; however, the most common method of determining nurse staffing levels is to use minimum nurse-to-patient ratios [50, 254]. Queueing models, on the other hand, have the flexibility to capture the stochastic nature of patient demands; therefore, they are a natural tool to determine nurse staffing levels [254]. de Vericourt and Jennings [50] examined fixed nurse-to-patient ratios from a queueing perspective. Treating medical units as closed multi-server queueing systems, they demonstrated that the fixed nurse-to-patient ratio policy cannot achieve high service quality across different unit sizes. Yankovic and Green

[254] developed a bivariate Markov model with state space (X_b, X_n) to model the relationship between bed occupancy and nursing demand, where X_b represents the number of occupied beds plus the number of patients requiring a bed, and X_n represents the number of inpatients under nursing care plus the number of patients needing a nurse. By viewing each independent clinical unit as a finite-source queueing system with two types of servers (nurses and beds), they derived formulae for a series of ED performance measures during time intervals with fixed staffing levels. They demonstrated the impact of unit size, occupancy rate and LOS on nursing levels, and concluded that fixed nurse-to-patient ratios can lead to either under- or overstaffing. Their results showed that even with sufficient bed capacity, inadequate nursing levels can cause significant boarding in the ED. Komashie et al. [120] developed a variant of the $M/G/1$ queueing model of patient and staff satisfaction levels, in which patient and staff satisfaction levels were represented by wait times and service times, respectively. They derived the Effective Satisfaction Level (ESL), for which the patient and staff satisfaction levels were maximized. Their proposed method enabled ED systems to quantify service quality for better capacity planning. By examining a system's deviation from its ESL, the authors provided guidance for clinical staffing for a desired level of patient satisfaction. Maman [149] developed a Poisson mixture model with the $M/M/c+G$ queue to study optimal staffing levels while meeting a pre-specified wait probability goal. They extended this model to an $M_t/M/c+G$ model with time-varying arrival rates and analyzed it asymptotically in steady state. By calculating the optimal staffing levels under a pre-specified wait probability, they found that the system performance strongly depends on the order

of over-dispersion (i.e., the arrival rate uncertainty), which is measured as λ^c, where λ denotes the mean Poisson arrival rate and $0.5 \leq c \leq 1$. However, the literature mentioned above did not incorporate the fluctuations due to patient arrivals, departures, and transfers, which might significantly impact the nursing demand [237, 255]. As we discussed earlier, the time-lag phenomenon—whereby the system congestion level lags behind patient arrivals—has been a major challenge to modeling systems with non-stationary arrivals. The direct outcome of such lagging is that hospitals cannot simply determine resource allocations at each staffing interval based on its corresponding average arrival rate [108, 84].

Several approaches have been proposed to deal with the time-lag phenomenon. Some are based on steady-state approximations, such as PSA, SIPP, and their lagged versions (as we discussed in Section 2.3.1.1). Assuming that the system reaches steady state quickly, one can compute the steady-state offered load (OL) for each interval; then, it is possible to apply traditional staffing strategies over that interval. When the service time is long, the modified offered load (MOL) approach [155] can be applied based on the steady-state or square-root approximation. For instance, in MOL, one can calculate or approximate the time-varying OL $R(t)$ via a corresponding system with ample servers; then use a time-varying adaptation of the square-root formula: $s(t) = R(t) + \beta\sqrt{(R(t)}$, where $R(t)$ is the OL, and β is a parameter characterizing the quality of service. Rounding s in the above formula up to the nearest integer provides a feasible staffing level [246]. de Vericourt and Jennings [50] recommended, as a remedy to fixed nurse-to-patient ratios, the use of policies that employ square-root staffing for large service systems. Aimed at

determining the minimal hourly staffing levels required to achieve the U.K. government's 4H target (i.e., 98% of patients to be treated within 4 hours of arrival), Izady and Worthington [108] derived an iterative algorithm that combines infinite server networks, square-root staffing, and simulation. After taking into consideration the factors such as time-dependent arrivals, various patient types, and resource sharing, they applied their algorithm to a real A&E department and greatly improved the success rate of achieving the 4H target. To reduce the proportion of patients who LWBS by a physician, Green et al. [79] studied a non-stationary queueing model to set ED physician levels. Using the $M/M/c$ queueing model as part of a lag-SIPP approach for time-varying demand, their scheduling policy has been implemented in practice and the proportion of patients who LWBS decreased significantly as a result.

Further, Zeltyn et al. [258] used QT-based simulation models to address the ED staffing problem with time-varying demand. They incorporated the OL technique and square-root safety staffing based on the $M/M/c$ queueing model. Their model helped ED staff with short-term (several hours or days ahead), mid-term (several weeks or months ahead), and long-term (several years ahead) physical ED relocation planning, as their ED was scheduled to move to a new location. The staffing recommendations they provided were implemented by a large Israeli hospital and they had satisfactory results. Yom-Tov and Mandelbaum [255] investigated a time-varying Erlang-R model with reentrant patients to determine required staffing levels to achieve predetermined service levels, for example, related to utilization and wait probability. The authors then used the model to develop a time-varying

square-root staffing policy based on the MOL. This model reflected the reality that patients occupied critical resources even when not being attended to by ED staff. They demonstrated that this model was useful in determining staffing levels, as it captured the complexities of the ED sufficiently well.

Queueing models can also be used to examine the impact of a more specific human resource in the ED. For example, Silberholz et al. [216] simulated an $M/G/c$ queue to evaluate how the residency teaching model affects operational efficiency in the ED at an academic hospital. Based on a natural experiment involving residents in the ED, they showed that—contrary to the popular belief that a residency program decreases ED efficiency—residents actually increase throughput and reduce service and wait times.

2.3.2.2 Non-human Resource Management

Beds

Ensuring sufficient bed capacity and maximizing resource utilization are two conflicting objectives for the ED system. Similar to the staffing problem, the allocation of non-human resources is also affected by the time-varying nature of patient arrivals. Steady-state allocation rules, such as the Rough Cut Capacity Planning (RCCP) and OL, are techniques for determining resource levels and are commonly used in manufacturing and service systems. These rules match offered capacity with the predicted demand using estimates of service times [236]. RCCP accounts for the variations in time spent at each resource and integrates demand predictions into its

plan for resource capacities, but it does not incorporate the lag between patient arrival and service times [258]. Patients often spend several hours in the ED on average; therefore, the effect of this lag cannot be ignored. OL, as a refinement of RCCP, calculates total workload in a more reasonable manner by using the average service rate to calculate the workload on the entire time horizon. The combination of OL with the corresponding steady-state Erlang model is a powerful tool for determining system resources. As an example, Zeltyn et al. [258] studied the optimal scheduling of X-ray resources under alternative operating hours and found out that the optimal operating hours were 12:00-18:00, instead of a 10-hour period as initially suggested.

The requirement on inpatient bed capacity is central to hospital management as it ultimately determines staffing level and costs [106]. QT has been widely utilized to analyze bed levels in various healthcare settings [78, 41, 213]. In the ED, Huang [106] extended the results by Pike et al. [190] by incorporating the day-of-week effect into the queueing model. Their results indicated that the daily occupancy level of the emergency bed follows a Poisson distribution. Gupta [87] applied an $M/M/1/k$ queue and concluded that under a fixed staffing level, increasing the number of ED beds would lead to longer patient wait times, but ambulance diversions would be reduced. de Bruin et al. [48] investigated a sequence of two-station queueing systems (FIFO for cardiac aid, then the coronary care unit) with blocking to study congestion in emergency care chains. Under the constraint of a performance target (e.g., maximum 5% refused admissions), they aimed to find a strategy for optimizing bed allocation. They demonstrated the impact of fluctuations in demand, and obtained

the optimal bed allocation strategy. Dealing with the same problem, de Bruin et al. [49] found that insufficient bed supply in the care chain led to refused admissions, and that large variations in workload were caused by variability in LOS and patient arrivals. Lin et al. [137] utilized two connected queues to determine the required number of ED and IU beds. Their results indicated that there is an optimal IU resource level for each performance target, and that increasing the capacity of the IU is the best option for managing the unpredictability in ED arrivals.

Room configuration and ED redesign

The growing number of patients has placed increased pressure on hospital administration boards for more healthcare facilities, outpatient services, and responsive treatment. One remedy is through ED redesign by optimizing space allocations, process flow, and operations [243]. Both space reallocations and process flow optimization are related to patient segmentation or new service areas in the ED. Zeltyn et al. [258] studied the effect of a newly designed, larger ED with longer walking distances. For the sake of infection control, infected or colonized patients are often separated from those who are susceptible (i.e., patient cohorting). Palvannan and Teow [179] studied how patient cohorting affects ED admission wait time. Using a $M/M/c$ model, they found that more beds are required to compensate for the longer wait times associated with partitioning the beds to serve these separated groups of patients. For example, an additional 5-7% bed capacity was required for cluster-level cohorting to restore the original two-hour wait time.

2.3.3 Summary

In this section, we classified ED QT research into two problem-specific subgroups: demand- and supply-oriented problems. We observed that on the demand side, the priority queue is the most frequently studied problem, whereas on the supply side, a variety of efforts have been devoted to staffing and scheduling problems. QT is a useful approach for these types of problems because there are readily available methods for priority queues, various staffing rules, as well as mechanisms to deal with time lagging.

2.4 Modeling-oriented Perspective

In this section, we review ED QT applications from the perspective of modeling techniques. Mathematical queueing models are used to gain closed-form or recursive formulae to calculate performance measures in steady state [87]. In an ED setting, however, the connections and routes between different sections can be quite complex (as shown in Figure 2.1); therefore, one has to make certain simplifying assumptions in order to adopt QT models. These assumptions typically involve the time distribution of arrivals and service, server types and capacities, room and bed capacities, queue disciplines, and rates of abandonment. In this section, we provide a summary of key applications of QT in the ED, by either viewing the ED as an independent queueing system or as a node in a larger queueing network. We list an overview of ED QT models in Table 2.4.

2.4.1 The Emergency Department as an Independent Queueing System

There are several ways to classify queueing systems. They can be classified into single- or multiple-station (i.e., network) models according to their structure. They can be classified into finite- and infinite-source models according to the size of their sources. And finally, they can be characterized as single or multiple customer class models [87]. In this section, we focus on the single-station models, in which the ED is modeled as an independent queueing system. In addition, we discuss the finite- and infinite-capacity model variants for this specific application. In the case of infinite-capacity models, the patient arrivals are independent of the number of patients in the ED. For the finite-capacity models, the arrival intensity depends on the state of the ED, as the system will block out patients exceeding the queue capacity. We list some specific ED QT applications and assumptions in Table 2.5.

2.4.1.1 Infinite-capacity Models

Queueing models with infinite queues and multiple servers (i.e., $G/G/c$) can be used to determine steady-state queue length and wait time statistics. The most common case of the $G/G/c$ model is the $M/M/c$ model. The popularity of the $M/M/c$ model is primarily due to its mathematical tractability and the fact that interarrival times are well approximated by the exponential distribution [87]. Many standard performance measures of $M/M/c$ queues such as the wait probability or

the mean wait time can be calculated either via the Erlang C formula or a Markov-type analysis. Even when arrival and service rates are not stationary, $M/M/c$ models can be applied to determine resource levels so as to prevent peak-period congestion [84, 87]. In an ED environment, arrival rates and service times can be estimated via averaging during a stationary period, and the $M/M/c$ model can be used to provide insight into system performance. However, a direct application of an $M/M/c$ model can underestimate the ED crowdedness. Yankovic and Green [254] found that ignoring the influence of nursing levels on bed dynamics led to negatively biased estimates of queue length and wait times, especially for scenarios with a high OL.

When the exponentially distributed arrival or service time assumptions no longer hold, one can use the $G/G/c$ model to study the finite server system. However, closed-form solutions for the $G/G/c$ model are available only when arrival and service rates follow some specific distributions. Therefore, one will need to either approximate GI or G by specific distributions (such as Erlang and phase-type distributions), or derive two-moment approximations for performance measures such as mean wait times and mean queue lengths [245, 75, 87]). Some examples of ED $G/G/c$ models are listed in Table 2.4.

2.4.1.2 Finite-capacity Models

Finite-capacity queueing systems can be used to model the overcrowding phenomenon in the ED. When ED waiting rooms are fully occupied, new arrivals can be

blocked until additional waiting space becomes available or current patients LWBS. The $M/M/c/k$ model can be applied to determine capacity levels. Staffing is one important aspect of capacity that determines the service rate. The number of ED beds is another important component of capacity, which affects the number of refused arrivals through AD. Both types of capacities influence patient wait times, and contribute to operating costs [87].

When the waiting room capacity equals the staffing level (i.e., $c = k$), we can apply the Erlang loss formula to calculate the overflow probability and the capacity requirement for the resultant $M/M/c/c$ system. For example, de Bruin et al. [49] used this model to examine bed allocation in an emergency cardiac ward. They first modeled the emergency care chain system as an $M/M/\infty$ queue, in which the bed capacity was infinite and the bed occupancy could be calculated for any time t. Later, they incorporated the phenomena of refused admissions using the $M/M/c/c$ model. They assumed when all c beds were occupied, a newly arriving patient would be blocked (i.e., refused admission).

It is worth noting that classical analysis of queues relies on a set of equations involving Markov steady-state transition probabilities. Using the normalization equation, one can estimate the number of patients or the level of utilized resources in steady-state. For instance, Almehdawe et al. [7] modeled the total number of ambulance patients in service (or waiting) in the kth ED at time t ($q_k(t)$) as a continuous-time Markov chain with finite-states, in order to compute the stationary distribution for the number of patients in the system. By partitioning the states into subclasses based on $q_k(t)$, they derived the infinitesimal generator of the Markov chain, and,

then, modeled a quasi-birth-and-death process with level-dependent rates.

2.4.2 The Emergency Department as a Node in a Queueing Network

In this section, we focus on ED QT articles that view the ED as a node within a larger queueing network model of the hospital. In the hospital, each department provides specialized services for many types of patients, which drives requirements for department resources [87]. Queueing networks have been studied extensively [122], and they are ideal for modeling the many interacting service components that operate within a hospital.

Hospital network

Armony et al. [15] modeled the ED as a node in the hospital queueing network. They developed a simple birth and death model in which the arrival and departure rates depend on the ED states, and found that such a model can characterize the distributions of ED occupancy and LOS reasonably well. Deo and Gurvich [52] modeled two EDs without ambulance diversion as independent $M/M/c$ queues. They integrated the two EDs using a continuous-time Markov chain model $X(t) = (X_1(t); X_2(t))$, where $X_i(t)$ is the number of patients in each ED at time t and examined the effect of AD. As mentioned previously, Almehdawe et al. [7] also studied the interaction between a regional EMS provider and multiple EDs (refer to Section 2.3.1.1).

Table 2.4: Overview of ED QT model applications.

Queueing Model		**Article**
Infinite Capacity $G_t/G/c_t$	$M/M/c$	Allon et al. [6], Broyles and Cochran [30], Deo and Gurvich [52], Green et al. [84], Haussmann [92], Palvannan and Teow [179], Sharif et al. [212], Vass and Szabo [235], Yankovic and Green [254], Zeltyn et al. [258].
	$M/M/c/\infty/n$	de Vericourt and Jennings [50].
	$M/M/1$	Madsen and Kofoed-Enevoldsen [147].
	$M/G/1$	Komashie et al. [120], Stanford et al. [223].
	$M/M/\infty$	de Bruin et al. [49], Hagtvedt et al. [89].
	$G/G/c$	Cochran and Roche [43], Lin et al. [137], Saghafian et al. [206], Silberholz et al. [216].
	$D/G/1$	Fiems et al. [65].
	$M_t/G/c_t$	Izady and Worthington [108].
	$G/GI/c/c$	Lin et al. [137].
	$GI/G/c_t$	Panayiotopoulos and Vassilacopoulos [180].
Finite Capacity $G/G/c/k$	$M/M/c/k$	Allon et al. [6].
	$M/M/1/k$	Cochran and Broyles [42], Gupta [87].
	$M/M/c/c$	de Bruin et al. [49].
	$M/G/c/c$	Cochran and Roche [43].
	$M/GI/c/c$	Lin et al. [137].
Queue With Abandonment	$M/M/c+G$	Maman [149].
	$M_t/M/c+G$	Maman [149].
	$M/GI/c/s+GI$	Wiler et al. [248].
Markov Process		Almehdawe et al. [7], Au et al. [17], Hagtvedt et al. [89], Saghafian et al. [206], Yankovic and Green [254], Zayas-Caban et al. [257].

Table 2.5: ED QT applications and assumptions.

Article / Problem	**QT Model**	**Assumptions**
Yankovic and Green [254] / determine ED staffing levels	Modified M/M/c model	1. Poisson arrival + service times; 2. Fixed inpatient number in the ward in a given time; 3. Independent nursing care requests with an exponentially distributed time interval; 4. Identical servers (nurses); 5. No blocking; 6. Infinite waiting room.
Cochran and Roche [43] / evaluate the performance of split flow	Multi-class queueing network; $M/G/c/c$	1. Capacity is decided by number of bed; 2. The acuity levels assigned to patients are accurate; 3. The general LOS data is correct.
Wiler et al. [248] / examine LWBS Rate	M/GI/c/s+GI approximated to M/M/c/s+M(n)	1. Stationarity of patient arrivals (validated for three 2-hour time periods); 2. Weibull distributed patient wait time tolerance.
Silberholz et al. [216] / residency teaching effect	$M/G/c$	1. Fixed Poisson arrival rate; 2. No abandonment; 3. FIFO queue discipline; 4. Each bed being treated as a server.

continued on next page

continued from previous page

Article / Problem	**QT Model**	**Assumptions**
Cochran and Broyles [42] / relationship between LWBS and business	$M/M/1/k$ with abandonment	1. Patients in a ED collectively behave as a group; 2. Approximate reneging ED queue with balking ED queue; 3. ED Service rate and capacity are not given.
Green et al. [79] / effect of staffing levels on LWBS rates	$M/M/c$ queue with Lag-SIPP	1. $M/M/c$ for every 2 hours; no triage; FIFO queue discipline; 2. Assume a delay standard (i.e., at least 80% of patients must be seen by a provider within one hour).
de Vericourt and Jennings [50] / nurse-to-patient ratio	$M/M/c/\infty/n$ closed queueing system	1. States for patients are: stable and needy; 2. Patients transit from stable to needy after an exponentially distributed time interval; 3. Exponential/ non-exponential service time; 4. FIFO queue discipline; 5. Identical nurses.
de Bruin et al. [49] / bed allocation	$M/M/\infty$, $M/M/c/c$	1. Finite number of beds and no waiting area; 2. An arriving patient will be blocked if all beds are occupied.
Stanford et al. [223], **Sharif et al.** [212] / priority queue	Priority queue modified from FIFO $M/G/1$, $M/M/c$	1. Stable queue; 2. Same arrival rates for both classes.

Table 2.6: Articles examining the ED-to-IU network.

Article	QT	Assumptions
Mandelbaum et al. [150]	Inverted-V-shaped queueing system	A single centralized queue and k heterogeneous wards; each ward contains N_i servers (beds). Upon arrival, each patient is either directed to an available ward or joins a centralized queue of infinite capacity.
Lin et al. [137]	$M/GI/c_1/\infty$ with priority and $G/GI/c_2/c_2$	ED queue: Five priority classes, with high priority patients receiving immediate service; patient is discharged or transferred from ED to IUs, depending on the availability of IU beds; IU queue: no priorities or buffer. Bed capacity is primary resource in the ED and IU.
Broyles and Cochran [30]	Two $M/M/c$ queues in series	Service rate for ED, IU is unknown and estimated by statistical methods. Bed capacity is primary resource for both queues.
Allon et al. [6]	$M/M/(N_1 - B)$ and $M/M/N_2/K$ queue (after approximation)	Two priority classes in separate queues; Poisson arrival rates to the ED and admission rates to the IU; each station has multiple servers (beds); hospital goes on AD if the number of boarded patients exceeds K.

Emergency department and inpatient unit network

Researchers have focused a lot of attention on studying the interaction between the ED and the IUs. There are several reasons for this focus:

1. There are many interactions between these departments.

2. The ED-IU subnetwork serves a large proportion of patients within the hospital. For example, among all the patients entering the hospital studied in [15], 53% of them stayed within this subnetwork.

3. This subnetwork has little interaction with the rest of the hospital [15].

In Table 2.6, we summarize the models and assumptions used to analyze the interaction between the ED and various IUs. All of these articles assumed stationary arrival rates, exponentially or generally distributed service times, and that the IU can accommodate all types of patients. Except for Mandelbaum et al. [150], all papers treated the ED and IU as separate queues. Mandelbaum et al. [150] studied various routing strategies that assign hospital patients from the ED to inpatient wards. They developed a queueing system based on [14] with a single centralized queue and k heterogeneous wards. Each of the wards contains N_i servers (beds). Depending on the availability of servers, a patient is either directed to an available ward or joins a centralized queue. Lin et al. [137] used two queues to model patient flow between the ED and the IU. The first queue was an $M/GI/c_1/\infty$ model with five priorities for patients based on their health conditions; it was used to calculate the wait time to access the ED. Then, the authors employed a $G/GI/c_2/c_2$

queue (where c_2 is both the number of servers and the capacity in the IU) without priorities or buffer (e.g., the waiting room in the ED) to model patient flow in the IU. They incorporated the coupling effect between the two units by estimating the probability of full capacity in the IU and the probability of blocked patients in the ED. Then, they proposed an iterative algorithm to derive the necessary and sufficient conditions (related to ED service rate), for which a steady state for both queues can be approximated.

Allon et al. [6] used a two-station queueing network to model patient flow in the ED and IU. They modeled each station with multiple servers where N_1 and N_2 denote the number of beds in the ED and IU, respectively. The priority streams of patients were modeled as two separate queues, with independent Poisson arrival rates to the ED and admission rates to the IU. They assumed that the service times in the ED and LOS at IU are both exponentially distributed, and the hospital diverts patients if more than K boarded patients are in the ED. In order to improve the analytical tractability, they approximated the ED with an $M/M/(N_1 - B)$ system and the IU by an $M/M/N_2/K$ system, where B represents the average number of beds occupied in the ED.

There are relatively few QT papers viewing the ED as a node in the overall hospital network. This is due to the complexity in system modeling and the limited tractability of QT models in these scenarios. As the system becomes more interactive (e.g., embedded system, chained system, multiple services with priorities), deriving analytical formulae of different measures may no longer be feasible. In the following section, we explore analysis that combines simulation with queueing

theory to address these issues.

2.5 Comparison of Queueing Theory and Simulation in the Emergency Department

Building an accurate queueing model for the ED system can be challenging; variations in clinical conditions, priority classes, and system resources are difficult to capture in an analytical formulation. Simulation, particularly discrete event simulation (DES), is an important methodology that has been used extensively in healthcare. DES models imitate system behavior using the sequential execution of events while exhibiting great flexibility in testing various interventions [185]. In the context of the ED, a patient's stay includes events such as arrival, triage, diagnosis, treatment, and departure, with waiting occurring at any point in the process when all resources are currently being utilized. Patients are usually modeled as passive entities who will consume resources such as physicians, nurses, and beds at different times during their stay. The greatest advantage of DES is that it captures the essence of human activity and operational details. As a result, many researchers have used DES to simulate systems in detail—rather than make a lot of simplifying assumptions—and obtain performance measures to compare to the observed system [15].

In this section, we examine the application of QT in combination with DES. We first examine ED QT articles that implement both methods for the purpose of validating results generated from each other (i.e., double validation). Then, we

explore research that combines both methods into a hybrid model. Next, we compare the data acquisition and challenges for each method. Finally, we identify conditions for which each method provides advantages over the other.

2.5.1 QT and Simulation for Double Validation

Queueing models are often simple approximations of actual ED systems that do not include all of the steps of the operational process; therefore, researchers compare these models with simulation models that better describe the dynamics between patients, staff, and other hospital resources [254]. In this subsection, we examine articles that attempt to validate queueing models using simulation. In Table 2.7, we compare results from QT and simulation models that are used in the same article for this purpose.

For some of the articles, results from both methods are quite similar. For instance, Lin et al. [137] found similar effects of the available IU capacity, ED patient arrival rate, and average wait time of different triage levels on the resources required to achieve performance targets. Sharif et al. [212] used simulation to verify their theoretical results for wait time distributions. Cochran and Roche [43] found that performance measures such as wait time and area overflow probability were consistent between the two methods. Xu and Chan [252] used simulation to explore potential reductions in patient delays when applying their proposed admission and diversion policies to the ED. They verified that the proactive policies based on QT were robust under the variation of the error rate of predicted arrivals, rate

of diversion, and rate of patient abandonment. The simulation results showed that their proposed policies consistently outperformed standing policies, and could reduce patient wait times by up to 15%. Similarly, Huang et al. [105] first simulated an ED having the same features as their queueing model to evaluate the performance of the patient selection policy they proposed. Then, they checked the robustness of their policy by adding more complex features to the simulation that were not incorporated into their QT model (e.g., time-varying arrivals, delays between visits, finite ED capacity, multiple servers, and patients who abandon the system). The results indicated that their queue-generated policy outperformed commonly used alternatives in all systems. They also showed that the more complex ED features would not degrade the performance of the queueing model.

Other articles observed mixed results when comparing the two types of models. Yankovic and Green [254] developed queueing and simulation models for the bed-staffing system to validate their results. The two models shared nearly the same assumptions on patient flow and parameter settings, except that the simulation model incorporated a specific nurse requirement and bed-cleaning time for the discharge process. The authors used the simulation to test the robustness of the exponential assumption for LOS, and found that the staffing estimates based on QT are very reliable under different arrival and service time distributions. They also compared the results between the two models in order to study the influence of average LOS on the staffing level and wait-time targets. The results indicated that when average LOS is short, the queueing model may underestimate delays and staffing levels.

Table 2.7: Comparison of QT and simulation applied in same article for double validation.

Paper	Simulation Purpose	QT/ Simulation Results Comparison
Yankovic and Green [254]	Test reliability and assumption validity of their QT model; Examine the impact of average LOS on staffing level and wait-time targets	The QT model's staffing estimates are very reliable under different input parameter distributions, with occasional underestimation of delays and staffing levels when average LOS is short
Cochran and Roche [43]	Validate patient wait times and area overflow probability	Performance measures are consistent
Lin et al. [137]	Validate the impact of several variables on required ED capacity	Results match very closely
Silberholz et al. [216]	Use QT model to validate simulation model	The wait times predicted by the QT model are lower than the simulation model; The two models point in the same direction for door-to-bed times
Yom-Tov and Mandelbaum [255]	Validate QT models in large and small systems to pinpoint unfitness; Compare staffing recommendations given by two QT ED models	In large system, the QT and simulation performance fit closely for some scenarios, but not necessarily for other scenarios
Allon et al. [6]	Validate the accuracy of their queueing approximations both with and without heavy ED traffic	Results match closely, and the accuracy of the queueing approximation increases as the traffic intensity of the ED increases

continued on next page

continued from previous page

Paper	Simulation Purpose	QT/ Simulation Results Comparison
Xu and Chan [252]	Verify the insights generated by the QT model on ED admission control and diversion	Simulation verified that the proactive policies based on the QT model are robust under various conditions, and reduce patient wait times by up to 15%
Armony et al. [15]	Test how the number of patients in the ED depends on the time and state of the system for different QT models	Discovered that a state dependent queueing model matches the behavior of the simulated and observed systems
Huang et al. [105]	Examine the proposed policy based on the queueing model; Test policy performance on relaxed conditions	Simulation verified that their queue-generated policy performs well, and the relaxed ED features do not lead to significant performance degradation
Saghafian et al. [205, 206]	Test conjectures made by their QT models; Test results under relaxed conditions	Simulation verified queue-based conjectures, and identified more general situations where the new policy can indeed improve patient flow
Sharif et al. [212]	Validate theoretical QT results	Results match with no discrepancy
Almehdawe et al. [7]	Validate theoretical QT model assumptions; Relax QT assumptions	Results are similar as long as the loss probability is small

Yom-Tov and Mandelbaum [255] used simulation to validate their MOL approach for the Erlang-R model within time-varying queueing networks. They validated their model using simulation from three perspectives: 1) within a large, general system that does reflect hospital operations, 2) a small system with patient arrival rates derived from hospital data, and 3) an actual emergency ward with more complexity. In the first case, the authors explored two operating regimes, a quality and efficiency driven system (QED), which is characterized by high resource-utilization and high service-quality (measured by queueing delays), and an efficiency driven system that focuses explicitly on the high levels of resource-utilization. They found that the results matched closely for the steady-state wait probability, average server utilization, and conditional distribution of the wait time given a delay in the QED regime, but not for the efficiency driven system because it violates the steady-state assumption. In the second case, they observed that there is a gap between the queueing and simulation results for the wait probability - service quality parameter relationship, which may be due to the rounding effect of using asymptotic approximations in small systems. In the third case, they applied their simulation model to a real hospital to determine the required staffing level. By comparing the simulated results to the Erlang-C and Erlang-R models, they demonstrated that the Erlang-R model yields better performance. From the insights generated by simulation, they also concluded that the queueing model (Erlang-R) implemented through MOL performs well for QED regime instead of the efficiency driven regime, and, the larger the system, the better the performance (for example, the nurse staffing recommendation performs better than the physician staffing recommendation, since

there are more nurses).

Allon et al. [6] utilized simulation to validate the accuracy of their queueing approximations with and without heavy ED traffic in predicting the fraction of time on AD and the wait probability. By fixing the arrival rates and ED size, both the simulation and the queueing models suggested that the fraction of time on diversion decreases as the number of inpatient beds increases. They also found that as the ED traffic gets more intense, the estimation accuracy of the QT model increases correspondingly. Similarly, Almehdawe et al. [7] applied simulation to validate their rigid QT model assumptions (i.e., zero transit time and exponential service times). By adding transit times to the QT network and using general service time distributions, they compared results from both the theoretical QT and simulation models, and found that the assumptions made in the QT analysis are valid as long as the ambulance utilization is low enough.

Saghafian et al. [205] and [206] both described detailed simulation models for testing conjectures suggested by their queueing models under more general assumptions. By incorporating more realistic features such as non-stationary arrivals, multistage service, inaccuracy in triage classifications, potential bed blocking in the hospital, and limits on physician-to-patient ratios, they confirmed their conjectures and identified situations for which the new patient flow design was better suited. Their results indicated that the new design was more robust to patient mix variation and triage errors. It is worth noting that the flexibility of simulation models allows for testing more complex scenarios that may be difficult to evaluate using QT models. For instance, in [206], the authors analyzed scenarios for which triage

classification errors are symmetric (i.e., equal probability of false positives and false negatives) or asymmetric. Silberholz et al. [216] used simulation and QT to analyze the impact of ED residents on wait times, throughput, and LOS. Their queueing model reported that there was a 59% reduction in wait time when residents were present relative to when they were absent, compared to 35% from the simulation model. They also observed that the queueing model underestimated the wait times due to the simplifying assumptions. They explained that there is less variability in the queueing model than the real system, hence less likelihood of high congestion and lower average wait times. Yet, with respect to the door-to-bed time, the queueing and simulation models both point in the same direction.

Simulation can also be used to compare multiple QT models. For example, Armony et al. [15] used a validated simulation model of a specific ED to measure the quality of their proposed queueing models. They conducted experiments comparing the proposed queueing and simulation models to the observed number of patients in the system. They fit stationary, time-varying ($M_t/M_t/\infty$), state-dependent ($M_i/M_i/\infty$), and time- and state-dependent ($M_t/M_i/\infty$) queueing models with parameters estimated from empirical data. Figure 8 illustrates comparisons between the empirical distribution of the number of patients in the ED and the distributions estimated by the aforementioned queueing and simulation models. Of the queueing models, they found that only the state-dependent model ($M_i/M_i/\infty$) fit the outcome well across the majority of the distribution.

We observe that QT and simulation can usually produce similar results when used to model the same system. Simulation can be used to test the quality and ro-

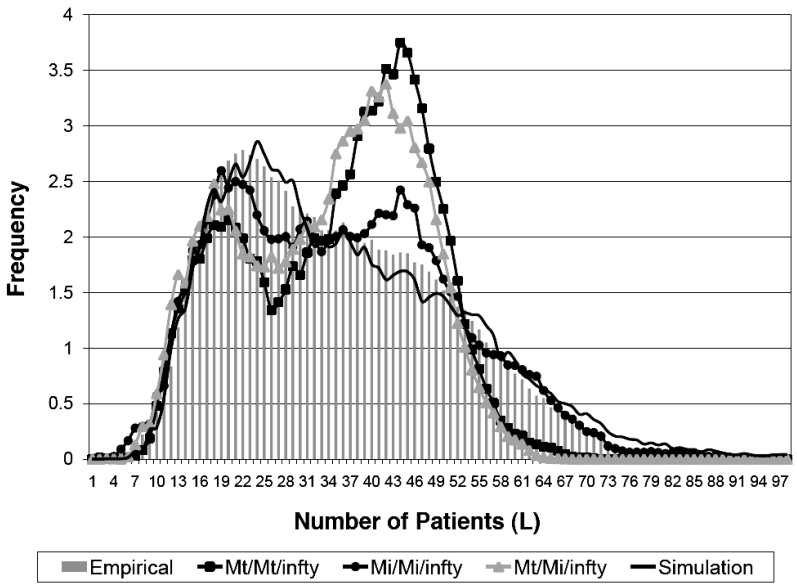

Figure 2.8: Comparison of empirical distribution of patient number against various queueing models and simulation (reproduced from [15] with permission).

bustness of queueing models, as well as validate or generalize any insights generated from them. However, as a result of simplified assumptions, with respect to arrival and service time distributions (e.g., stationarity and Poisson arrival assumption), patient heterogeneity (e.g., classes or priorities), or system boundaries (e.g., interactions within a larger network of queues), some QT models tend to underestimate wait times. In other cases, some assumptions (e.g., exponential service time distributions) may overestimate wait times. In general, QT appears to be more reliable when modeling larger, high-traffic systems, which often generate less variability than their smaller, less busy counterparts.

2.5.2 QT and Simulation as Complementary Modeling Approaches

Queueing models that attempt to capture many of the complexities of the ED are often analytically intractable, so consequently, researchers resort to the combination of simpler QT models and simulation as a modeling approach. The research on the OL concept is an example of this approach. Zeltyn et al. [258] combined analytical staffing formulae with simulation to develop a staff-scheduling algorithm. The authors extended the framework of a single-station system proposed by Feldman et al. [64] to a service network designed for the ED. Assuming a non-stationary Poisson arrival rate and resources with infinite capacity (e.g., physicians and nurses), they first calculated the number of busy resources for each hour to determine the time-dependent estimate for the OL for each resource via multiple simulation runs. The recommended staffing level for each hour was then calculated using square-root staffing formulae based on the steady-state approximation of the wait probability given by the $M/M/c$ queueing model. The method carefully balanced low wait time with high utilization of resources. In Table 2.8, we list articles that combined QT and simulation models as complementary modeling approaches.

In order to test the reliability of their queueing model, Yankovic and Green [254] studied the nurse staffing levels required to limit the probability of inpatient delay to an acceptable level. Experimenting on a set of parameters including unit size, nursing intensity, average LOS, and bed utilization, they used the queueing model to estimate the minimum required number of nurses for each scenario. Then, they used the simulation model to test whether the proposed nursing level resulted

in a probability of inpatient delay that satisfied the objective. They defined an experiment as unsuccessful if the simulated probability of inpatient delay was 10% greater than the target. They found that less than 2.5% of the experiments were unsuccessful, and that the queueing model was more likely to overestimate staffing levels in some cases. Similarly, Izady and Worthington [108] used simulation to estimate the percentage of patients discharged within 4 hours for a target wait probability. They searched over a range of service parameters (i.e., β in the square-root formula) until the target wait probability was achieved. They extended the staffing method proposed by Jennings et al. [109] to $M_t/G/c_t$ networks with K servers, and demonstrated how queueing models combined with simulation could reduce ED congestion by modifying staffing levels.

QT can also be used to generate analytical insights and provide direction to the development of large-scale simulation models and experiments. For instance, Hagtvedt et al. [89] first developed a small-scale queueing model containing only two beds and found that while on selective diversion, an increased arrival rate leads to prolonged time on full diversion. Then, they incorporated insights from this analytical model to build a large-scale simulation model of a hospital with 100 beds and a diversion policy for which the hospital will not accept new patients until 10 free beds are available. They tested a variety of occupancy levels for enforcing partial diversion and found a significant impact of this level on the time spent on diversion.

All of the models we have discussed in this section are separate queueing and simulation models that complement each other. Since the computational cost of

most queueing models is usually much less than that of a comparable simulation model (which must also be run for multiple replications to achieve a desired level of precision), it is often desirable to combine queueing and simulation into a more cost-efficient hybrid model. Shanthikumar and Sargent [211] summarized four types of hybrid models applied to general settings:

- Type I models operate over time by alternating between an analytic and simulation model through their interfaced solution procedure;
- Type II models have analytic and simulation models that operate over time through a joint solution procedure;
- Type III models use results from the solution procedure of simulation within the analytical approach, and Type IV models use the results from the analytical method as input to a simulation.

Note that in our review, [258], [108], and [254] applied the Type IV hybrid model to the ED settings.

Table 2.8: Articles that combined queueing theory and simulation methodologies.

Papers/ Problem	**Methodological Approach**
Zeltyn et al. [258] /Utilization improvement	Used the square-root-staffing principle based on input from simulation and the $M/M/c$ queue to determine recommended staffing levels
Izady and Worthington [108] /Congestion alleviation	Combined queueing network with a heuristic iterative algorithm, for which simulation was used to estimate the percentage discharged within 4 hours for a specific delay probability
Hagtvedt et al. [89] /Ambulance diversion policy	Derived a small scale QT model to generate the corresponding qualitative solution, then applied simulation to mimic the -scale network
Yankovic and Green [254] /Nurse staffing	Used the simulation model to test whether the nursing level generated from the queueing model results in a tolerable probability of delay under various parameter combination

Table 2.9: Data source used in the surveyed articles.

Data Source	**Literature**
Hospital/ED Electronic Medical Records	Cochran and Roche [43], de Bruin et al. [49], Green et al. [84], Hagtvedt et al. [89], Haussmann [92], Maman [149], Silberholz et al. [216], Wiler et al. [248], Zeltyn et al. [258].
Historical Operational Data	Allon et al. [6], Armony et al. [15], Huang et al. [105], Izady and Worthington [108], Lin et al. [137], Mandelbaum et al. [150], Palvannan and Teow [179], Saghafian et al. [205, 206], Siddharthan and Jones [215], Silberholz et al. [216], Yankovic and Green [254], Yom-Tov and Mandelbaum [255].
Field Measurement	Green et al. [84], Mayhew and Smith [156], Zeltyn et al. [258].
Expert Estimation	Cochran and Roche [43], Izady and Worthington [108], Komashie et al. [120]
Research Literature	Saghafian et al. [205, 206].

2.5.3 Data Sources and Challenges

Both QT and simulation studies employ various data sources for their models. Frequently used data sources include hospital- or ED-based electronic med-

ical records (including relevant event timing, medical and demographic information about patients, as in [38]), historical operational data (e.g., arrival and service rates, number of staff and beds in the ED or hospital), field measurements (i.e., direct observation), and expert estimation (via interviews with care providers and administrators). And in the absence of these direct sources, data is often referenced from the research literature. These data support modeling patient arrival patterns and flow as well as service-time distributions for various care-related activities. We summarize the data sources used in the selected ED QT literature in Table 2.9.

As an illustrative case of data acquisition and usage for a QT-based modeling study, Izady and Worthington [108] applied their approach for a typical accident & emergency (A&E) department in the U.K. using information collected from a 7-day survey among 12 hospitals. The authors estimated arrival profiles (i.e., hourly arrival rates) for patients with minor, major, and admittable conditions based on this survey data. The hourly arrival rates for each patient type were based on an annual attendance of 87,000 patients, which approximates an average sized A&E department in the U.K.. Local sources provided the percentage of each patient type for diagnostic tests, and A&E experts modeled the average service rate using the exponential distribution. Similarly, we observe that most of the referenced simulation articles also obtain input data from similar sources, except that they need more detailed and realistic inputs such as time-varying patient arrival and service rates, gap time between activities, and time-dependent staffing levels to achieve the reliability of the simulated process (readers can refer to [206] as an example).

QT and simulation models both require information from valid data sources.

As a result, there are several data-related challenges associated with modeling ED operations:

1. Incomplete records due to archiving or system-related errors. This challenge was common prior to the widespread implementation of electronic medical records.

2. Unavailable or censored data due to ED system complexity or lack of sophisticated data collection systems. Typically, electronic data systems in hospitals only contain some information on the current state of the ED, and information such as the number of patients waiting in the queue is usually unavailable [258]. However, due to the potential costs associated with collecting this information, some data is likely to remain unattainable, such as physician service rates [83] or patients' tolerance for long queues [258]. For example, wait times are observed for patients who are willing to complete their waits and unobserved for patients who leave without treatment, which has effectively censored this outcome for these patients [21]. Also, it is difficult to accurately estimate the mean service time and service capacity. Since ED beds are unavailable for patients during turnover periods, the average service time cannot be estimated based on a patient's LOS. Also, sometimes ED beds are underutilized because they are not fully staffed and are, therefore, not operational [30].

3. Inaccurate records due to hospital operations and policies rather than patients' real physical transactions. As an example, a patient's recorded time in the ED may be artificially extended if that patient is boarded as a result of an inpatient

bed shortage. Such a transfer delay can also inappropriately influence the wait times for subsequent patients [30, 42, 15].

The incorrectness or incompleteness in ED data may result in unrealistic data input, thus, influencing modeling accuracy and the validity of results. Therefore, estimating key model parameters under these circumstances is an important challenge.

To address the data scarcity issue, several simulation studies have attempted to fill the gap of data unavailability using prediction methods. For example, Zeltyn et al. [258] applied their simulation-based modeling approach to help ED administrators infer missing information about the current state of the ED. This approach contributed to the short-tem estimation of the ED state, and provided decision support on staff scheduling. As another example, Kuo, et al. [124] proposed meta-heuristic methods for estimating the parameters of a Weibull distribution to model several operational processes (e.g., the duration of a doctor's consultation). These parameters could not be estimated directly from data due to incomplete records.

By contrast, the abundance of data can lead to other challenges and opportunities. The advancement in big data and cloud storage has made possible the accumulation, management, analysis, and assimilation of large disparate data from healthcare systems [24]. The emergence of new data resources in healthcare, such as personal technology (e.g., mobile, wearable, and location-tracking devices) and social media (e.g., Twitter, Facebook, Google) has realized the monitoring of individual-based data in real time [153, 195]). A few hospitals have utilized these new data

sources and achieved noticeable results [258, 4, 181, 12]. These new data sources can potentially improve the amount and quality of usable data for more accurate estimation of patient wait times, LOS, and service times inside the ED, and provide valuable resources for system analytics, optimization, and validation [186, 100, 130, 12]. Furthermore, it enables researchers to develop personalized, real-time information systems that reflect the status of patients and the resources that serve them.

However, despite the benefits of big data, there are several challenges for integrating QT, simulation and big data to address complex questions in the ED [153].

1. Complete aggregation of various data sources is not always feasible. For example, providers and the payers may use distinct identifiers for patients in order to protect private information. Therefore, although the data is abundant, it is usually difficult to establish a full linkage between various data sources.

2. The derivation of key model parameters requires significant data cleaning and manipulation efforts [15]. With disparate data in various formats and structures, it can take significant amount of efforts to clean and normalize data.

3. Big data alone does not provide sufficient information to inform patient-centered care and improve healthcare delivery [153]. Modeling and analytics are still needed to enhance the value provided by big data.

From the above discussion, we observe that queueing models facilitate the development of analytical formula and theoretical insights with minimal data requirements. By comparison, detailed simulation models can capture more complex

behaviors through incorporating analytically intractable probability distributions or complex care pathways. Although simulation has the advantage of being more flexible, QT can provide analytical solutions that offer more generalizable insights that are less sensitive to parametric changes. The combination of QT and simulation, either via validation, comparison, generalization of each other, or complementary or hybrid models, provides theoretical insights and practical foundations for ED optimization [112, 233, 5].

2.6 Conclusions

In this chapter, we presented a review of 48 articles published from 1970 to 2015. We acknowledge that this review may not be entirely exhaustive, but it reflects the contemporary research on the application of QT to modeling ED processes and the comparison of QT vs. simulation approaches within the same context. From this review, we observed that the number of ED-related queueing studies has increased tremendously in the past five years and researchers in the healthcare management and operations research fields have become the key contributors of these publications.

We observed that queueing models are invaluable tools for ED design and management. With minimal data requirements and efficient computation costs, queueing models offer theoretical insights to an ED system and provide directions and predictions to the large-scale operational process. For this reason, it is often desirable to use QT to inform the development of a detailed simulation model.

However, the highly generalized queueing models cannot capture the complexity of the actual ED dynamics and may predict less variability than the real system. In particular, we found that QT tends to simplify the system and underestimate delays and congestion and, thus, obtain less accurate results than simulation. These issues are less prevalent for larger and busier systems.

By comparison, simulation models have the advantage of incorporating more detailed behavior and generating more actionable results. We found that simulation is often more suitable when modeling hospital systems with more variability, such as small EDs under tight resource constraints, as small systems tend to be more sensitive to variation in parameters. However, the insights derived from these models can be specific to individual ED settings and, thus, less generalizable.

Therefore, the combination of queueing and simulation methods leads to a powerful approach to better ED modeling. There is a growing trend of interaction between QT and simulation, as hybrids of queueing and simulation models can be more cost-efficient. We observed that simulation can be used to test the quality and robustness of queueing models, as well as validate, refine, or complement insights generated from them. Simulation can also be used to estimate missing parameters in queueing models, if necessary. Meanwhile, queueing models can provide analytical formulae that facilitate the development of various performance measures from simulation, while offering generalizable insights that are less sensitive to parametric changes.

In order to mitigate ED congestion, we should attempt to optimize processes from within, and promote parallel efforts to improve operations for connected sys-

tems (e.g., patients arriving via ambulance, patients being discharged to IUs). In addition, careful consideration should be given to the employment of priority queues for patients. Certain interventions, such as fast tracking, can make it more challenging for the ED to provide timely service for patients in urgent needs [137].

QT is a powerful tool for the analysis of healthcare systems; however, in practice, applications of QT to real ED systems are still limited despite the abundance of established QT work. Given the trend in this research area, together with the accelerating rate of computing and big data technologies, we expect to see more interaction between traditional queueing models and other techniques in the ED. Below we list several potential future research directions:

1) **More realistic QT modeling**. Future QT research related to the ED may consider incorporating state dependence into modeling, because the patient arrival rate and service rate not only depend on time, but also depend on occupancy levels. Other possible directions include modeling for interrupted service/treatment time and parallel tasks for care providers, as physicians may serve multiple patients during the same time window and, thus, have discontinuous service durations for each patient. Campello et al. [32] provide related results for ED case-managers with multiple patient assignments and repeated interactions with each customer. As their results are generated for systems with stationary arrivals and homogeneous servers, more study on discontinuous service in the ED is open for exploration.

2) **Combination of QT with other methods**. Hybrid models involving QT,

simulation, optimization, statistics, and machine learning may help to capture more realistic ED behavior with fewer simplifying assumptions and lower computational costs. Lee et al. [131] described an ED decision support system that combined machine learning, simulation, and optimization to improve patient flow, while incorporating the variability in patient conditions and their requisite care. The predictive modeling can provide better estimation of patient clinical outcomes and support for automatic decision-making in healthcare systems [227]. Specifically, with the availability of more personal data through mobile-devices, Saghafian et al. [206] proposed finding data-driven rules that correlate patient characteristics, symptoms, and evaluations with treatment times and resource requirements, which can potentially lead to more effective prioritization and streaming policies. Therefore, an integrated system with QT and other modeling methods can be a possible direction for better ED system modeling.

3) **Real-time control and personalized operational planning**. Historical data does not always have accurate predictive power [130]. The popularity of healthcare applications and smartphone-based sensors has enabled real-time monitoring of disease progression and patient location within the ED through accurate measurements [153]. Such data has been used for pre-triage, remote consultation, emergency care consultation (e.g., within the ED), real-time disease outbreak prediction, as well as real-time prediction of ED during pre-hospital, in-hospital, and post-hospital stages [70, 100, 195, 240]. How do

these changes affect ED operations and modeling? How do we take advantage of them to better manage patient flow inside and outside the ED? How will the integration of QT with real-time data inform decision-making and transform health service delivery?

2.A Appendix – ED Performance Measures

Defining a set of key performance measures is important before making any operational decisions. First, we need to select a measure or set of measures that best represents the primary problem, and then use the models to identify an intervention that improves operations regarding these measures. There are numerous choices of appropriate ED performance measures [244]. We list some of the most commonly selected ED performance measures below:

a. Wait probability: The probability that a patient cannot be served immediately upon arrival due to the unavailability of providers. The advantages of such a measure include its insensitivity to model details, interpretation independent of scale [90], ease of computation [83], and convenience of goal setting [246]. We refer the reader to [82] for several alternative measures based on the wait probability.

b. Expected wait time: The expected time between arrival and being seen by a care provider (or being assigned to a bed). An important advantage of this performance measure is that little information is needed on the distribution of the wait time, and it is easy to set target goals relative to such a measure

[80]. This first advantage is directly linked to the main drawback: When using expected wait time, the remainder of the wait time distribution is neglected. Whereas the length of the wait might be acceptable on average, wait times might be intolerably high for some patients.

c. Length of stay (LOS): The time between arrival and departure. Sometimes policy makers set a maximum limit on LOS in order to 'satisfy' the social demand for rapid service. The most well-known example is the 4-hour target in the U.K., which states that 98% of patients must be served, discharged, or admitted to an inpatient unit within 4 hours [156]. The main drawback of setting completion time targets is that they might cause some inconvenience in the treatment procedure or reduce the quality of service [176]. Additionally, it seems that a very small percentage of extreme cases can have a severe impact on the mean LOS observed in hospitals [125], making it important to consider medians or other percentiles in the analysis of this performance measure [55].

d. Leave without being seen (LWBS) rate (also known as reneging rate or abandonment probability in QT): The proportion of patients who leave the ED before receiving care from a provider, usually because the anticipated wait time is too long. The benefit of this measure is that it incorporates the patient's perception of service, as it is the patient's decision to stay or leave. However, since reneging is a continuous-time and state-dependent decision process, the LWBS rate depends not only on time, but the queue length. The unavailability of patient departure times causes these measures (i.e., time before LWBS)

to be difficult to estimate for the real system [42].

e. Boarding time (also known as inpatient delay): Defined as the prolonged time interval between the hospital admission decision and the departure time from the ED [244]. ED boarding results in lower ED service capacity, longer ED wait times, higher rates of patients who LWBS, and more ambulance diversion [30].

f. Resource utilization: Defined as proportion of utilized resources to available capacity. Higher utilization has been shown to be correlated with longer length of stay and higher acuity [244].

For additional measures, please refer to [73], [122], and [244].

2.B Appendix – Notation and Terminology

In this review, we employ Kendall's notation [116] to describe queueing models. A queueing system is denoted by $A/B/s/K/n/D$, where A and B refer to the probability distributions of the inter-arrival and service times, respectively; s denotes the number of servers; $K \geq s$ is the capacity of the system including patients in service. n is the size of the source population and D refers to the queueing discipline (e.g., first in, first out). Typical distributions for A and B include: M for exponential, D for deterministic, G for general, GI for general independent (i.e., independent and identically distributed), and PH for phase-type distribution. When the last three parameters are not specified, it is assumed that K and N are infinite and D is first-in, first-out. We use $+G$ to indicate a system for which abandonment

is allowed with an arbitrary distribution to model abandonment times, and $+GI$ to indicate that the customer abandonment times are independent of the arrival process and service times and identically distributed. The subscript t (e.g., M_t) indicates that such distribution is time-dependent. The arrival rate into the queue (λ) multiplied by the average service time (τ) gives the offered load (OL), which measures the long-term average demand placed on the system resources.

Chapter 3: Early Detection of Bioterrorism

3.1 Introduction

Bioterrorism, namely the intentional release of viruses, bacteria, or other toxic biological agents, is considered a significant threat to the United States. Early detection of a potential bioterrorism incident is vital for controlling diseases and limiting the damage. However, as some candidate bioterrorism diseases (e.g., anthrax and viral hemorrhagic fever) present symptoms in humans similar to those of common illnesses (e.g., the flu), it is difficult to quickly distinguish between a bioterrorism outbreak and a natural disease. The existing techniques for early detection have failed to differentiate between epidemic diseases and bioterrorism attacks, required too many assumptions, or seemed too complex.

Wagner *et al.* [239] summarized mathematical foundations of early detection and reviewed previous work concerning the measurement of detection timelines. Using signal detection theory and decision theory, the authors identified strategies to improve the timeliness of detection and position ongoing detection system development within that framework. Lober *et al.* [141] discussed six existing public health surveillance systems, which were designed to enhance early detection of bioterrorism events. However, their method cannot distinguish between a bioterrorism outbreak

and a natural disease.

Yahav *et al.* [253] proposed a conceptual framework for differentiating between bioterrorism and epidemic scenarios. They constructed a multilayered network that included social and spatial components, and incorporated functional principal components analysis (fPCA) to characterize disease transmission. Despite the accurate results from their method, its shortcoming is that the model contains many arbitrary assumptions about the structure of the social, location, and human-location networks.

Our goal is to propose a method that is both accurate and easy to implement in practice. In this chapter, we develop a simple-structured, agent-based model to capture the transmission patterns of diseases caused by bioterrorism attacks or epidemic outbreaks. Based on the aggregated regional infection trends or the individual infection curves at each city, our research seeks to detect an attack when only a small proportion of the population is infected.

In Section 3.2, we discuss our methodology and simulation models for the bioterrorism and epidemic outbreaks, where human behaviors are based upon travel patterns during the day and at night. In Section 3.3, we present results for both models under various local working probabilities and validate our simulation results with a two-phase mathematical model for the epidemic case. In Section 3.4, we include some conclusions and a mention of future work. We are particularly interested in distinguishing between these two disease scenarios when the first outbreak time is uncertain and in identifying appropriate infection control measures.

3.2 Methodology

We conduct experiments in computer-generated households and work places at three cities within a region, and simulate the spread of disease via bioterrorism and epidemic scenarios, respectively. The experiments are implemented using NetLogo (v. 5.0.4), an agent-based programming language and integrated modeling environment [247]. Our primary assumption is that the epidemic disease is transmitted only via human interactions [40], whereas bioterrorism is transmitted only through a person's proximity to the source of the attack [35].

3.2.1 General Model Description

Our preliminary results are based on 900 households in three connected cities—A, B, and C. We constrain the environment to be closed: no immigrants from the outside and no newborns from within the three cities. The models for the epidemic disease and the bioterrorism attack share a few common assumptions. We describe their shared properties in this section, and discuss their unique assumptions in Sections 3.2.2 and 3.2.3.

There are two phases to each model—SETUP and GO. In the SETUP phase, three square-shaped cities are generated. Our initial experiments suggest that city size does not matter as long as we analyze the proportion infected and dead, rather than the absolute counts. Therefore, we can assume all the cities are the same size, and generate 300 households per city. We also assume that the number of residents in each household is drawn randomly from the following distribution: [1, 2, 2, 3,

3, 4, 4, 4, 5, 5, 6]. In the GO phase, the model will oscillate between a daytime state (6 a.m.– 6 p.m.) and a nighttime state (6 p.m.– 6 a.m.). We assume that only public interactions take place during the daytime state and only within-household interactions take place during the nighttime state. Figure 3.1 demonstrates our model description in an activity-flow diagram.

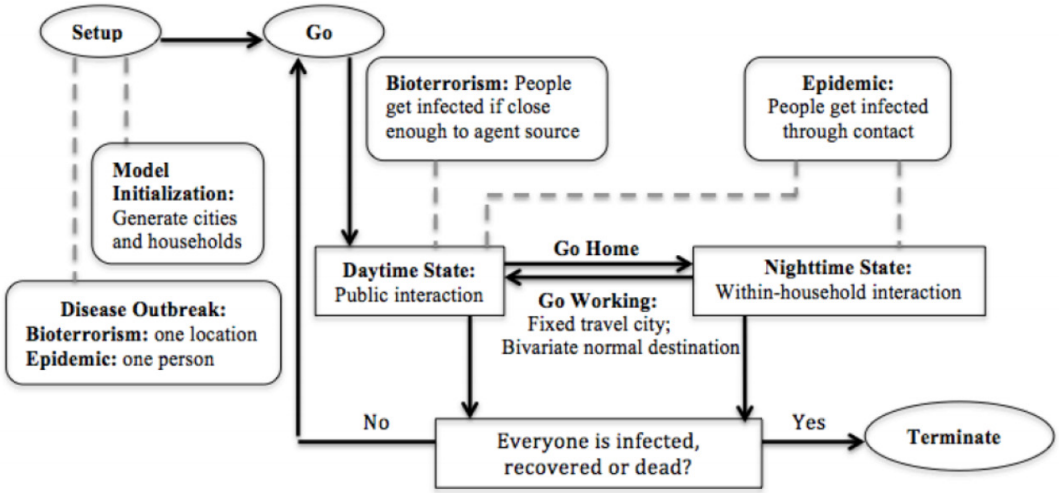

Figure 3.1: Activity-flow diagram, where ovals represent the model phase and arrows indicate the flow of progress. Rounded rectangles connected to the dashed lines explain the process in ovals and rectangles.

We suppose that each person's work city is fixed throughout the simulation. During the daytime state, some people work from their home city with a predefined local working probability p_L, whereas the remaining people travel to the other two cities for work with an equal chance. We make this assumption because for most people, their work or school cities are usually fixed. In addition, we assume that each person's destination inside the travel city follows a bivariate normal distribution, which is not fixed from day to day. This is based on the general observation that

most cities have a central area where many people work, and a person might appear in various locations within the city for different purposes. Figure 3.2 demonstrates the initial uniformly distributed spread of households and the bivariate normally distributed travel destinations.

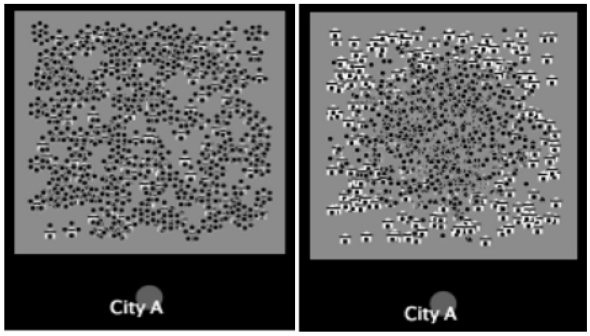

Figure 3.2: Initial setup for uniformly distributed households and bivariate normally distributed work locations in city A. The black and white house-shapes stand for households, and the grey circles with a black dot in the center represent people.

In each model, we assign one of four states to each person: healthy, infected, recovered (and immune), and dead. We assume everyone in the model must be in one of these states, and initially no one is immune to the disease. This is because if we assume a percentage of residents are immune to the disease at the initial state, then the infection curves will shrink by approximately that percentage in height (see Appendix 3.A for a more detailed justification). We assign an infection radius R for each disease, indicating the immediate area within which the disease is transmissible.

In our model, each infected individual has a predetermined chance of death or recovery. We assume those who will die pass away after a specific amount of

time (i.e., death survival time) and those who will recover will do so by the end of some stochastic recovery time t_R. The above assumption is based on the observation that if an infected individual dies, he tends to die quickly, and if he survives, the recovery usually takes much longer. If we denote the average recovery time for a disease to be $\overline{\tau}$, then we assume the recovery time t_R is drawn from a truncated normal distribution N($\overline{\tau}$,$\overline{\tau}$/6), where the truncation follows from 3.1 and 3.2 for the epidemic and bioterrorism cases, respectively [16].

$$t_R = \begin{cases} 2\overline{\tau} & t_R \geq 2\overline{\tau} \\ 3 & t_R < 0 \end{cases} \tag{3.1}$$

$$t_R = \begin{cases} 3\overline{\tau} & t_R \geq \overline{\tau} \\ 8 & t_R < 0 \end{cases} \tag{3.2}$$

There are a few variables in our models; however, most of their values are predetermined and can be drawn from historical data (e.g., recovery probability, average recovery time). For parameters whose values are uncertain (e.g., home/work infection probability, infection radius), we try to assign values that are as reasonable as possible or we deliberately control the infection dynamics of the epidemic and bioterrorism cases to be similar, so as to ensure that we can distinguish these two cases even for the most difficult situations.

There are three diseases in our experiment: normal epidemic, extreme epidemic, and bioterrorism disease, where the normal and extreme epidemic share the same simulation model yet differ in their recovery probability. An example of an extreme epidemic would be a super influenza pandemic, such as the catastrophic 1918 pandemic, which had an extremely high death rate [184, 224].

3.2.2 Bioterrorism Model

We propose a conceptual model for a bioterrorism disease. We assume the disease breaks out in one city (e.g., city C in our case). It only transmits to people who are within a certain distance from the source, and does not transmit among humans. During the daytime, a person may get infected (if near the source), recover, or die. During the nighttime, people go home, where no disease transmission occurs. We utilize a maximum location infection probability p_M to represent the probability of a person being infected near the bioterrorism source. For a healthy person who is within the infection radius R of a bioterrorism source, the probability of getting infected in the source city is inversely proportional to the square of the distance from the source location.

We assume the bioterrorism source is located at the city center, where the population density is highest. This is because the goal of bioterrorists is to create as much chaos as possible; thus, the city center would be an ideal attack location. Because bioterrorism has a high mortality rate—ranging from 0.2 to 0.9 if untreated, and 0.01 to 0.45 if treated in time, we set the death probability to be 0.7 without distinguishing whether the patient is being treated or not [96, 34]. We also assume p_M is high (0.6). This is based on the properties of most biological agents [63]. Finally, we assign survival times and and average recovery times drawn from practice [86, 16]. A complete listing of parameter settings of the bioterrorism model is shown in Table 3.1. Later in Section 3.3.3, we will consider many combinations of parameter settings and develop the maximum, minimum, and median dynamic curves.

3.2.3 Epidemic Model

We assume the epidemic initiates with a single person, and the disease propagates to other individuals within the region through human interaction only. During the daytime, people within the influence radius R of an infected person will acquire the flu with a work infection probability p_W. During the nighttime, only people who live with an infected family member will get infected, with a home infection probability p_H. In both cases, a patient may recover or die from the disease. We assume the work infection probability p_W is low (0.03), and the home infection probability p_H is relatively high (0.4), due to the intimacy among family members. We apply 0.001 and 0.3 as the death probability for normal flu and super influenza, respectively [184, 224, 134]. The parameter settings are presented in Table 3.2.

3.3 Results

Our results are based on the interaction between three cities, because this configuration is representative of many metropolitan areas. In this experiment, we only consider the initial 15 days. This is partially due to the uninhibited environment of the initial spread of a disease (i.e., no residential protection like mask-wearing and improved hygiene, and no human intervention like quarantine, travel restrictions or vaccination). In addition, we are only interested in quickly distinguishing bioterrorism from an epidemic in an early stage, so the study of the spread during the initial days is reasonable.

In the following experiments, we vary the local working probability p_L from

Table 3.1: Parameter setup—bioterrorism.

Parameter	Value(s)
Average Recovery Time $\overline{\tau}$	15 days
Death Survival time t_D	3 days
Local Working Probability p_L	0.33, 0.6, 0.9
Recovery Probability p_R	0.3
Death Probability	0.7
Maximum Location Infection Probability p_M	0.6
Infection Radius R	10
Outbreak Location	Centre of city C

0.9 to 0.33, due to the belief that many residents will work locally, but there is still interaction among cities. Experiments indicate that our model has low variance across many replications, therefore we base our results on the mean outcome under the same parameter settings. In Figure 3.3, we see the aggregated infection curves over three cities for our three scenarios when the local working probability $p_L = 0.6$. We observe that the normal and extreme epidemics share the similar "S" curve, whereas the bioterrorism curve has a curve with a strictly decreasing slope. However, in practice, when the initial outbreak time is unknown, it is difficult to tell whether an unknown disease is a bioterrorism outbreak from day 2 to day 6 or an epidemic disease from day 5 to day 10.

Table 3.2: Parameter setup—normal & extreme epidemic.

Parameter	Value(s)
Average Recovery Time $\overline{\tau}$	10 days
Death Survival time t_D	5 days
Local Working Probability p_L	0.33, 0.6, 0.9
Recovery Probability p_R	0.999 for normal flu, 0.7 for extreme flu
Death Probability	0.001 for normal flu, 0.3 for extreme flu
Infection Radius R	3
Work Infection Probability p_W	0.03
Home Infection Probability p_H	0.4
Initial Infection Number	1

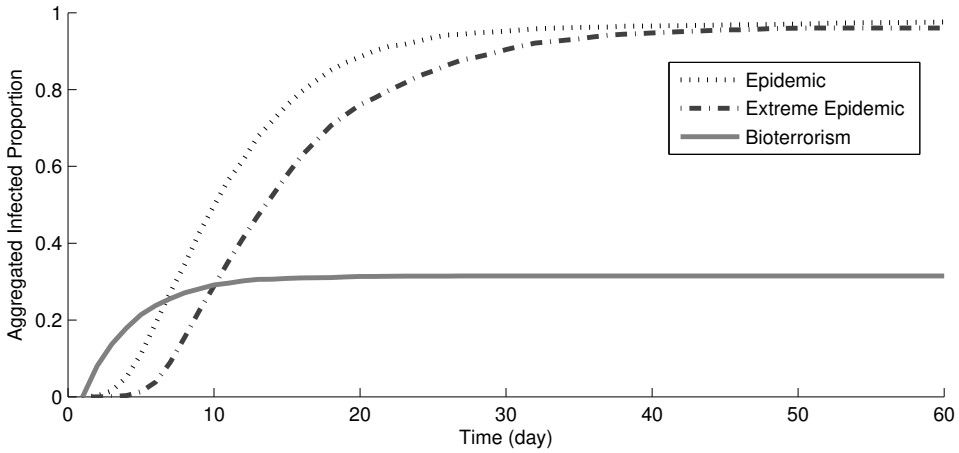

Figure 3.3: Comparison between the aggregated infection curves of three diseases among three cities when local working probability $p_L = 0.6$.

3.3.1 Bioterrorism Scenario

In the following analysis, we gradually decrease p_L from 0.9 to 0.33 while keeping the other parameters fixed. Experiments indicate that a high p_L is not always associated with high proportions of aggregated infection and death. The bioterrorism curves share an increasing trend with a decreasing slope, which is due to the shrinking susceptible pool that travel to the source location.

Figures 3.4, 3.5, and 3.6 present aggregated and individual city infection and death curves for local working probability p_L equal to 0.9, 0.6, and 0.33. As p_L decreases from 0.9 to 0.33, the overall mobility of the population increases, yet the aggregated infection and death proportion among the three cities does not vary much. However, the infection curves for individual cities present distinct characteristics. When p_L is 0.9, the infection curve for the bioterrorism source city C dominates the others, with a city infection proportion of 0.9, whereas the infection proportion for the other two cities is less than 0.1. As p_L decreases to 0.6, city C's infection proportion decreases to 0.6, while the infection proportion of the other two cities increases to 0.2. When p_L is 0.33, the three cities have a similar infection proportion at 0.33, and the difference among the three city curves is indistinguishable. The decrease in infection dominance of city C is a result of our fixed work city assumption. When p_L is high, few people from the other two cities have the chance to approach the bioterrorism source in city C, thus a very small portion of people from cities A and B can get infected. Therefore, p_L has little to do with the overall infection or death proportion among the three cities. However, as p_L decreases and,

thus, the mobility of the population increases, the three cities have similar infection and death curves.

3.3.2 Epidemic Scenario

We apply a similar method to study epidemic disease transmission. A normal flu epidemic differs from an extreme epidemic only in death rates, so here we only present the extreme epidemic curve, for it demonstrates a more severe death curve. Experiments indicate that the infection and death proportion curves in the extreme epidemic model for various p_L share a similar "S" shape, which is due to the increasing infection agents at the beginning stage and a shrinking susceptible pool later on.

Figures 3.7, 3.8, and 3.9 present aggregated and individual city death and infection curves for $p_L = 0.9$, 0.6, and 0.33, respectively. In contrast to the bioterrorism case, the behavior of individual cities is relatively close, but the transmission curves present a "time-lag" pattern, especially when p_L is large. For example, when $p_L = 0.9$, the infection curve of the source city (C) starts to grow sooner. Yet after about two days, the infection curves of city A and city B begin to grow in the same manner. Such a phenomenon is the result of the epidemic transmission property each person can be considered as a "disease source", thus once a non-initial-outbreak city has an infected person, this city will reproduce the disease dynamics of the initial outbreak city, given that they have similar transmission characteristics. As p_L decreases, the difference in disease dynamics between the initial outbreak city and the other two

cities becomes less visible.

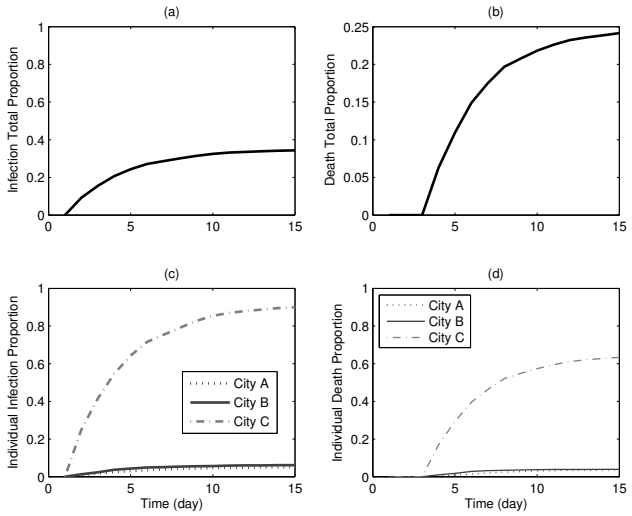

Figure 3.4: Aggregated infection (a) and death (b) curves and individual city infection (c) and death (d) curves in bioterrorism model for $p_L = 0.9$.

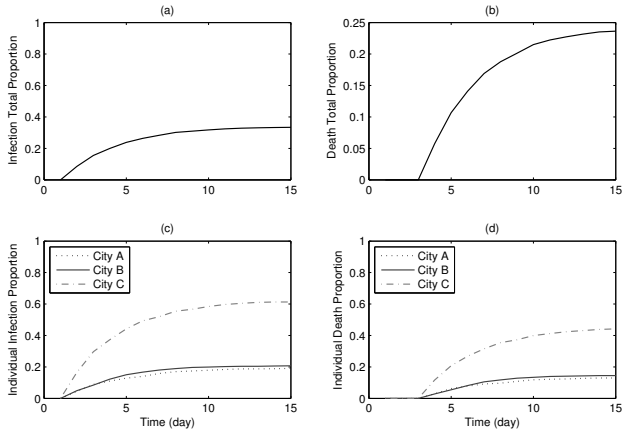

Figure 3.5: Aggregated infection (a) and death (b) curves and individual city infection (c) and death (d) curves in bioterrorism model for $p_L = 0.6$.

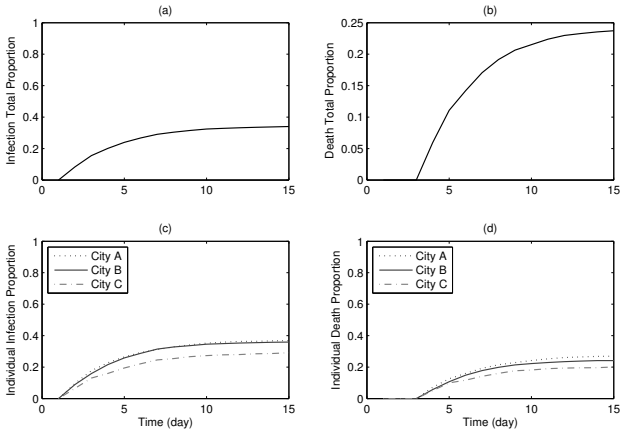

Figure 3.6: Aggregated infection (a) and death (b) curves and individual city infection (c) and death (d) curves in bioterrorism model for $p_L = 0.33$.

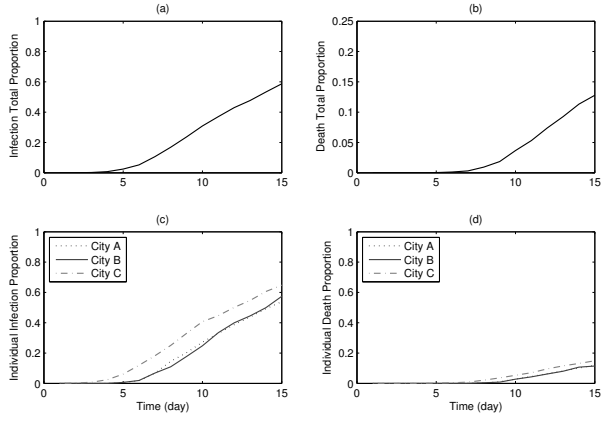

Figure 3.7: Aggregated infection (a) and death (b) curves and individual city infection (c) and death (d) curves in the extreme epidemic model for $p_L = 0.9$.

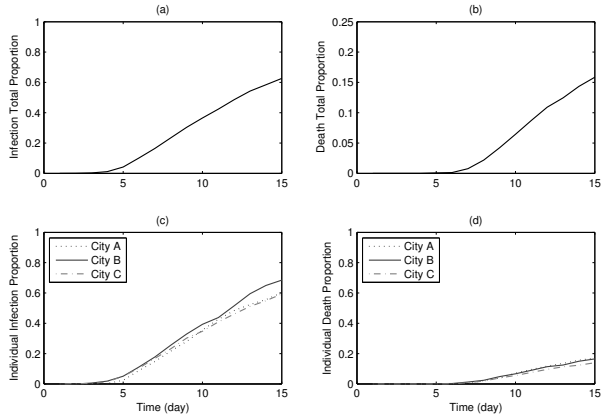

Figure 3.8: Aggregated infection (a) and death (b) curves and individual infection (c) and death (d) curves in the extreme epidemic model for $p_L = 0.6$.

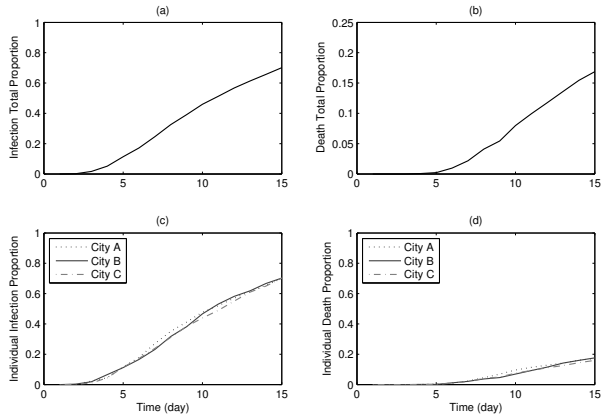

Figure 3.9: Aggregated infection (a) and death (b) curves and individual city infection (c) and death (d) curves in the extreme epidemic model for $p_L = 0.33$.

3.3.3 Model Validation

In this section, we propose a two-phase equation-based model to validate our epidemic simulation results. We retain the simulation model assumptions such as the population details, immunity levels and travel patterns, and let $S(t)$, $I(t)$ and $R(t)$ represent the susceptible, infected, and recovered population in the three cities at time t. For the epidemic scenario, we adopt the discrete version of the Kermack-McKendrick SIR model [148]:

$$S_{t+1} = S_t - \frac{\theta S_t I_t}{N},\ I_{t+1} = I_t + \frac{\theta S_t I_t}{N} - \gamma I_t,\ R_{t+1} = R_t + \gamma I_t \tag{3.3}$$

where $\theta = \alpha\beta$,with α representing the average number of contacts per individual per unit time and β representing the transmission probability during each contact, and the recovery rate $\gamma = [\overline{\gamma} + (1 - p_R)t_D]^{-1}$. We assume the model oscillates between the day and night phases. Accordingly, N stands for the population at a household in the night phase (N_H) and the population of a city in the day phase (N_C).

We apply the home and work infection formula: $\theta_H = (N_H - 1)p_H$ and $\theta_W = p_W R/rN_C$, where r stands for the city radius. Figure 3.10 shows the aggregated infection curves for this two-phase equation-based model.

In comparison, transmission curves generated from our simulation model (Figure 3.11, left) exhibit a similar shape as the two-phase SIR model for the epidemic case, and with a much shorter model running time. In addition, our agent-based model can keep track of each individual's behavior, whereas the mathematical model fails to distinguish between individuals and their travel patterns. Since analytical

bioterrorism models that characterize infection dynamics, location, and transportation are not available yet [200], we are not able to validate our bioterrorism model using a similar approach. However, our simulation results for both epidemic and bioterrorism diseases resemble the transmission curve shapes produced by the model of Yahav et al., which are generated over a complex social and location network.

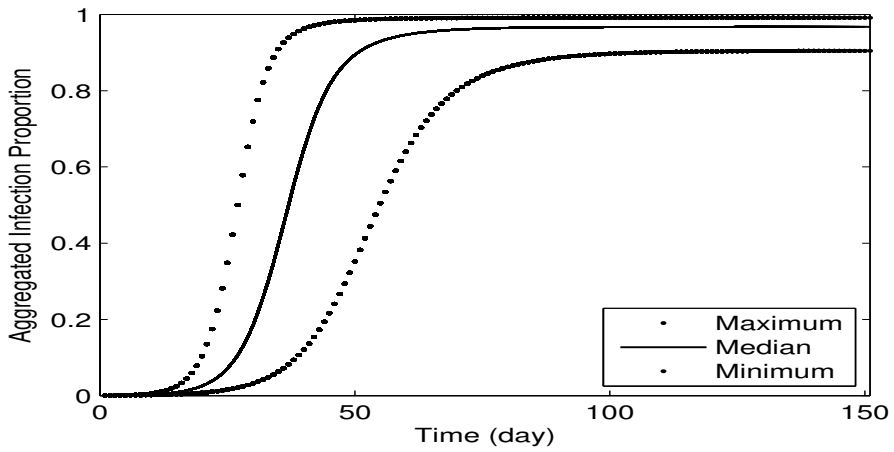

Figure 3.10: Aggregated infection curves for the epidemic disease from an equation-based model. The maximum, median, and minimum curves are based on experiments with the home infection probability p_H as it ranges from 0.1 to 0.7 and the work infection probability p_W as it ranges from 0.01 to 0.08.

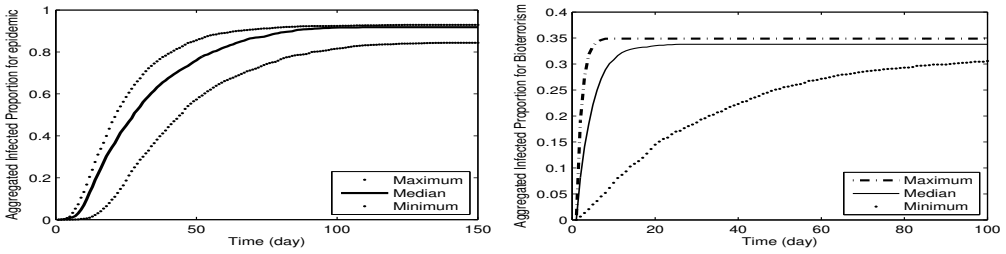

Figure 3.11: Aggregated infection curves for the epidemic (left) and bioterrorism (right) scenarios from our simulation model, with p_L ranging from 0.1 to 0.9. In the epidemic case, the home infection probability p_H ranges from 0.1 to 0.7 and the work infection probability p_W ranges from 0.01 to 0.08. In the bioterrorism case, the infection range R changes from 3 to 15, while the maximum location infection probability p_M changes from 0.5 to 0.9.

3.4 Conclusions and Future Directions

We proposed two models with simple structures and relatively few assumptions to capture the essence of bioterrorism and epidemic transmission. Our models indicate that the aggregated infection and death curves for a region can serve as an indicators in differentiating the two scenarios: the slope of the epidemic infection curve will increase initially and decrease afterwards, whereas the slope of the bioterrorism infection curve will strictly decrease. Our results also show that the local working probability p_L has little to do with the aggregated infection proportion for both scenarios; yet for the bioterrorism outbreak, as p_L increases, the bioterrorism source city exhibits a more dominant infection proportion. In contrast, for the epidemic model, the behavior of individual cities presents a "time-lag" pattern,

especially when p_L is large. As p_L decreases, the three dynamic curves (one for each city) converge as time progresses.

We are particularly interested in distinguishing between the two scenarios when the first outbreak time is not known. In practice, due to the time delay of reporting and seasonal noise, we may not be aware of the disease outbreak until after numerous cases or deaths, and it would be difficult to trace back and determine the first occurrence time. Because bioterrorism and epidemic diseases present similar transmission dynamics curves in their initial occurrence, the question becomes how to discriminate the two scenarios given an unknown first outbreak time. For example, in Figure 3.2, it is obvious that the dynamics of the bioterrorism outbreak and epidemic disease are different; however, if we simply observe an overall infection curve with an increasing shape, how can we tell whether a disease is at the initial stage of a bioterrorism outbreak or in an ongoing process of an epidemic disease? We want to develop tests to quantify such differences.

In future work, we would like to generate a database for simulations of various parameter settings. In this way, by comparing a disease curve from the real world with our database, we can tell which scenario the real world disease most closely resembles. We hope to explore the following research topics using our agent-based model:

1. Filter background noise

   Due to the influence of seasonal influenza, there will be background noise that affects the total number of infections reported by hospitals and clinics. Our

goal will be to filter this background noise and distinguish between an epidemic and a bioterrorism disease.

2. Impact of multiple bioterrorism attacks

   We want to investigate how the execution of the bioterrorism attack will impact our detection ability. The biological agents can be released in a single or multiple locations in the same city, or multiple locations in different cities, so it is important to understand how the release pattern impacts disease recognition.

3. Impact of cities' geographic and demographic traits

   By incorporating diverse geographic and demographic properties into the model (e.g., different city sizes, infection rates, travel patterns, etc.), we can examine how these differences impact our ability to determine whether a disease outbreak is due to a bioterrorism attack.

4. Early detection based on reports from individual cities

   Instead of examining the aggregate number of infected individuals in a region, we can investigate the transmission curve for each city. Is it possible to diagnose a bioterrorism threat in its early stage by looking at the infection number in each city?

5. Impact of the bioterrorism transition pattern

   Depending on the spreading pattern of the agent—whether it is transmitted by direct contact, air, contaminated water or food sources, or via vectors such as

mosquitoes—we can investigate how each pattern will affect the transmission trend and further influence disease detection.

6. Impact of human behavior

   The public response may influence disease detection as well. On the one hand, patients may choose to stay home or go to a hospital, or to restrict social interaction. On the other hand, susceptible people may take medical prophylaxis such as antibiotics or vaccinations, or adopt physical protection such as gauze masks. All of the these behaviors will impact the disease infection curve and challenge our ability to distinguish the bioterrorism scenario from an epidemic.

7. Impact of government policies

   We can investigate the influence of possible government policies for disease control such as quarantine or school closures.

8. Suggestions to the Centers for Disease Control and Prevention and local health agencies

   By evaluating the effectiveness of various disease control approaches, we might provide suggestions on strategies or policies that should be adopted when facing a bioterrorism threat.

3.A Appendix – Aggregated Infection Proportion under Various Immunity Levels

If we assume a percentage of residents to be immune to the disease initially, then the infection curves will shrink by approximately that percentage in height. Figure 3.12 and Figure 3.13 demonstrate the aggregated infection proportion under various immunity levels for bioterrorism and an extreme epidemic, respectively. In fact, if y represents the steady state aggregated infection percentage and x represents the immunity percentage, then the linear regression models for bioterrorism and an extreme epidemic disease are $y = 0.33x + 0.3316$ ($R_2 = 0.9609$) and $y = 1.0512x + 0.94814$ ($R_2 = 0.9965$). In both cases, increasing immunity by 1% reduces the height of the curve by 1% of the initial height (Since the initial height for the two diseases are 0.33 and 0.95, respectively).

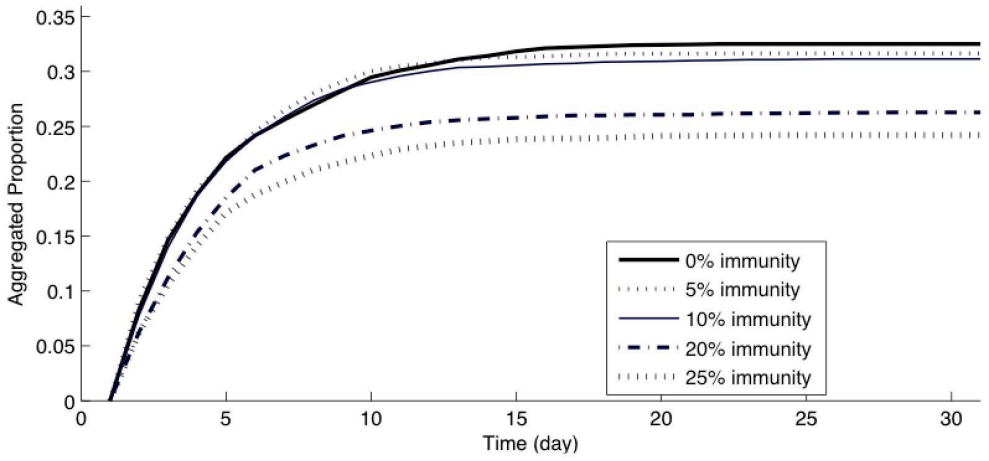

Figure 3.12: Aggregated infection proportion under various immunity levels for a bioterrorism.

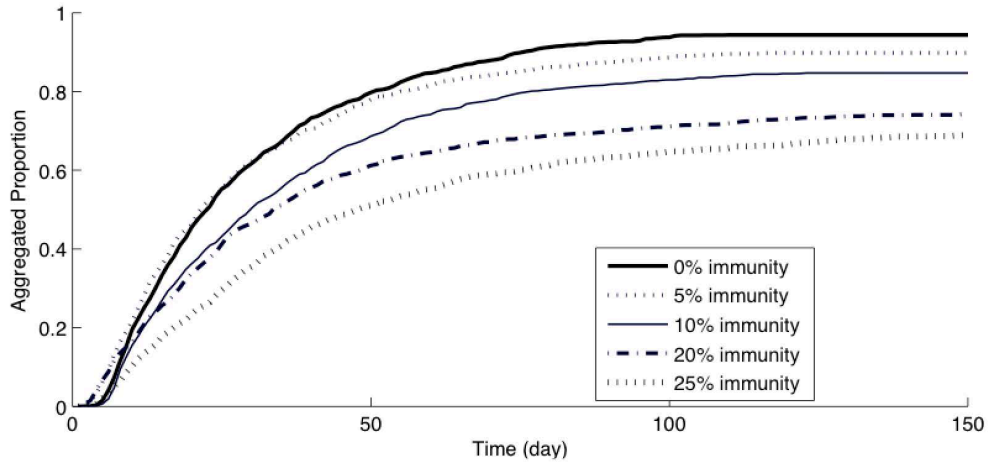

Figure 3.13: Aggregated infection proportion under various immunity levels for an extreme epidemic.

Chapter 4: Intelligent Selection of Case Management Enrollees

4.1 Introduction

Emergency departments (ED) are central to providing patients with acute access to medical care. Frequent ED users, that is, individuals visiting an ED multiple times per year, impose a significant burden on the healthcare system [29, 198]. Studies reveal that patients visiting an ED four or more times per year account for only 4.5-8% of all ED patients but 21-28% of all ED visits [125]. Such disproportionate usage of the ED not only causes overcrowding but these frequent users often have more complex conditions that are difficult to treat [165, 99, 8]. The care of frequent ED users is also more costly for insurance payers and health plans, as ED treatments are generally much more expensive than those provided at other healthcare facilities that do not provide 24-hour care [160].

Several interventions have been implemented to reduce the number of ED visits by frequent ED users, with case management (CM) being the most popular and effective method [173, 29]. Based on interdisciplinary cooperation, CM is a comprehensive interventional program to plan, customize and guide individuals' health services in order to promote patients' well-being and an effective and efficient healthcare system [132]. CM often focuses on outreach, for example, utilizing nurses or

skilled employees in call centers to create individually tailored care plans and maintain patient engagement [11]. In the case of CM for frequent ED users, a single point of contact (e.g., case manager, ED consultant, or social worker) is assigned to a CM enrollee to provide personalized care guidance and social support, which may extend beyond the ED and into the community [85, 8]. While some CM programs provide evidence of reducing frequent ED usage and improving clinical outcomes, findings on efficacy and cost-effectiveness remain mixed [214, 123, 220, 173, 8]. Despite the significant costs and resources required for a successful implementation, CM programs have not consistently lowered overall healthcare costs, and it is still unclear whether the resultant clinical and social benefits will balance the additional costs for managing frequent ED users [221, 8, 220].

One challenge for traditional CM interventions lies in the inappropriate enrollment of candidates. Oftentimes, intervention programs only target current frequent ED users [85, 220, 29]. Such an enrollment strategy can lead to a waste of system resources, as even without intervention, current frequent users may not continue to use the ED frequently in the future [8], or they may not be impacted by CM. Mandelberg et al. [151] found that a current frequent user has only a 28-38% chance of remaining a frequent user the following year. Applying CM to non-repeating frequent users may not improve the enrollee's health, and the resources could have been used for other patients. As the cost of CM is usually high (with reported costs ranging from $1,833 to $5,599 per patient in the United States) [173, 214, 47], programs need to carefully determine who to enroll, especially with limited resources. Ideal enrollees are those who will achieve improved health outcomes from their enrollment

and, if possible, generate cost savings for the organization [11].

One way to measure the effectiveness of a case management program is by its ability to reduce the incurred healthcare costs. So the impact of a CM program is determined jointly by: 1) the detection accuracy of identifying future frequent utilizers, and 2) the efficacy of the CM program in impacting the enrollees' future ED usage. Researchers have suggested that breaking existing habits requires more effort than starting new habits under behavior change interventions [72, 127]. Therefore, the efficacy of CM in reducing future ED usage of current frequent users (who have formed the habit of heavily utilizing the ED) may be low. In other words, even if we can accurately predict future ED usage levels of current frequent users, targeting and enrolling these members exclusively may not improve outcomes or save costs. As a result, it may be more effective to enroll current infrequent users whose ED usage may increase significantly in the future, as they have not yet formed the habit of heavy ED usage, and, thus, could benefit from CM.

Motivated by this reasoning, we present a novel machine learning framework—using claims data—for effectively selecting enrollees for CM, with the objective of maximizing the intervention's future savings. Specifically, instead of targeting current frequent users (exclusively), who may or may not repeat their ED usage behavior, our approach seeks to identify a mix of future frequent users, which includes both current frequent and infrequent users, for enrollment. We divide the future frequent users into two categories: 1) "jumpers" whose usage of the ED increases from infrequent to frequent from one year to the next, and 2) "repeaters" whose usage of the ED remains consistently high. Jumpers represent a small proportion of the

population (12% in our data), but they are an important segment when considering their healthcare costs will increase significantly in the future. Jumpers are usually not at the top of the list for CM enrollment, as no single factor (such as high number of prior ED visits) indicates that they will become high utilizers of the ED. In addition, compared with repeaters who have formed the habit of repeatedly visiting the ED, jumpers may benefit more from CM, based on the aforementioned discussion on behavior change interventions. Depending on the CM program size and the relative efficacy of impacting future jumpers and repeaters, our results demonstrate that by allocating some resources to high-risk future jumpers, early intervention could improve health outcomes and minimize costs.

Our machine learning framework is implemented on a set of insurance claims data for Medicaid patients. This population comes with multiple challenges; often these members do not have established relationships with primary care physicians, have no access to alternative treatments, or have no incentive to utilize less expensive healthcare services than the ED [119]. Studies have shown that Medicaid enrollment leads to increased ED visits [225, 66]; therefore, it is important to select enrollees who are likely to visit the ED often without intervention and can benefit from a CM program.

The remainder of this chapter is organized in the following manner: After reviewing related work in Section 4.2, we introduce our data and data processing in Section 4.3. In Section 4.4, we first describe the predictive models for identifying potential jumpers and repeaters, present prediction results, and then propose optimized selection strategies based on estimated savings and insights from cost

effectiveness analysis. We conclude in Section 4.5 with insights derived from our results, contributions, and directions for future research.

4.2 Related Work

Characteristics of frequent users

Depending on the goals of the study, the definition of frequent ED usage varies from 2 visits per year to 16 visits per year [59, 251]. Previous studies have indicated that demographics of frequent users such as age (older) and gender (female), chief complaints (pain, injury, skin disorders, cardiovascular, gastrointestinal, urinary tract, complications and exacerbations of chronic illness), health conditions (drug and alcohol abuse and mental illness), and usage of the overall healthcare system (outpatient visits, mental health visits) may be associated with frequent ED usage [165, 71, 125, 58, 29, 242]. Compared with infrequent users, frequent users tend to present a higher rate of morbidity, mortality, and complications of chronic conditions [29, 165, 71]. The majority of frequent ED users are in fair-to-poor health [107], and are more likely to be socially disadvantaged and homeless compared to infrequent users [165, 71]. Studies also show that frequent access to the ED also suggests inadequate use of primary, specialty, dental, and outpatient mental health care [160].

Predictions of frequent users

Several studies have explored predictive models for frequent ED users, utilizing traditional binary classification methods such as logistic regression. For example,

Wu et al. [251] predicted future frequent users based on ED registration data. They achieved an area under receiver operating characteristic curve (AUC) of 0.83 and 0.92 for predicting frequent users (defined as either with $\geq$ 8 or $\geq$ 16 ED visits), respectively. One of the strongest predictors in their study is the distance between home and the ED. Neufeld et al. [171] studied frequent ED users among rural older adults receiving home care services. They found that frequent ED users are associated with certain sociodemographic and clinical characteristics (such as age and the number of medications). They only reported adjusted odds ratios and confidence intervals for their features, making it impossible to compare their predictive performance with other models. Additional work has focused on the application of more advanced machine learning algorithms. For instance, Pereira et al. [187] studied the 3-class classification problem of "bucketing" patients into low, medium, and high frequency users, for which they applied various classification models based on discharge records. They showed that it is easier to predict low ($\leq$ 1) and high frequent users ($\geq$ 5) than medium frequency users (2-4 ED visits). All of the above studies are implemented on a relatively small set of features (i.e., less than 40), and none of them aims specifically to select the riskiest population for the purpose of CM.

Member Segmentation

Member segmentation has been widely applied in marketing to efficiently target heterogeneous populations [37]. In healthcare, population segmentation (sometimes via clustering) has been applied to plan for group-specific services and care arrangements [144], identify Medicare beneficiaries to foster informed healthcare

decisions [250], and understand patient demographic characteristics and patient preferences with respective to healthcare attributes (such as care efficiency, clinical reputation, and hospital environment) [167, 139]. In Vuik et al. [238] a general population is divided into utilization-based groups via k-means clustering, and the low-utilization group is targeted for preventive interventions. To capture highly distinct characteristics among a heterogeneous population, Dong and Taslimitehrani [56] propose Contrast Pattern Aided Classification to match group-specific classifiers with population segments that exhibit certain patterns.

Jumpers are a group of people who switch between member segments, yet the problem of identifying ED jumpers appears to be understudied. To our knowledge, no paper has focused on predicting jumpers in terms of their ED usage. Two papers predicted jumpers for other contexts [60, 11]. Both studies utilized claims data and tried to identify jumpers based on healthcare costs, that is, predicting low-cost individuals whose medical expenses were likely to increase significantly in the future. Anderson and Bjarnadóttir [11] show that CM will not reduce overall costs unless it can prevent over 7.5% of cost increases. They also demonstrate that predicting jumpers is far more challenging than identifying general future high-cost members.

4.3 Data and Descriptive Analysis

4.3.1 Data Sources and Preprocessing

For the development of prediction models for frequent ED users, we utilize insurance claims data generated when healthcare providers send information to re-

ceive reimbursement for their services. The data includes information on 190,009 members who were insured by a Medicaid plan from May 2008 through April 2013. Due to privacy concerns, the data is stripped of all dates (except the service year) and location information; therefore, we analyze the data on an annual basis from 2009 to 2012. To ensure completeness of the data, we include only the members who are enrolled for at least 350 days per year for two consecutive years (for details on eligible enrollment, please refer to Appendix 4.A). We call the first year the observation year, and use the information in the observation year to predict the individual level of ED utilization (i.e., frequent or infrequent) in the second year, hereby referred to as the outcome year. For instance, 2009, 2010, and 2011 are observation years for outcome years 2010, 2011, and 2012, respectively. Such a one-year prediction setup is motivated by the low enrollment retention rate of the Medicaid plan, as a member's enrollment is renewed annually based on both the member's continuing qualification and the current enrollment policy (such as a lottery-based enrollment) [225, 229]. In our Medicaid claims dataset, only 18% of all members who were enrolled in any year between 2009 and 2012 were continuously enrolled for 4 years, and only 45% of them were enrolled for 2 consecutive years. Therefore, we focus on a one-year prediction window in order to make our model applicable to a larger proportion of the population.

Our data set consists of members' enrollment records and claims for dental, pharmacy, mental health, lab tests, and medical services. The diagnosis data is coded using the International Classification of Diseases, Ninth Revision, Clinical Modification system (ICD-9-CM). The prescription drugs are coded using the Na-

tional Drug Code system [170]. The data contains masked patient identification numbers, which allows us to link individual patient records across different service categories. In this study, we define a visit to include all services coded under the same claim number.

We conduct significant data processing and feature engineering in order to bring the data into a suitable format for analysis. We aggregate all claims to create a medical profile for each member. To reduce the dimensionality of the data, we group individual diagnosis codes utilizing a general diagnosis category [169] and the Clinical Classifications Software (CCS) [94]. We also identify chronic medical visits using the Chronic Condition Indicator [93]. Further, we utilize the NYU ED Visit Severity Algorithm to label ED visits into a probability distribution over 9 categories including non-emergent, emergent, and substance abuse categories [26]. Then, for each individual, we calculate the average ED usage probability distribution over the aggregated ED visits. For more information on the data processing, please refer to Appendix 4.A. The resulting dataset has 164,402 records and contains 465 features on member demographics and annual usage of health services such as primary care visits, ED visits, dental visits, and prescriptions. All features are based on previous studies of frequent users (see Section 4.2) and our preliminary analysis. Table 4.1 summarizes the extracted features (refer to Appendix 4.A for definition details).

4.3.2 Descriptive Analysis

Figure 4.1 shows the distribution of the number of ED visits in the outcome year. We observe that most members have very few ED visits, and the distribution of the number of ED visits is extremely skewed, with a mean of 2.15 but a median of 0. The number of individual ED visits reaches as high as 287 in the outcome year. Overall, 83 patients (0.05% of total population) each paid over 50 visits to the ED in the outcome year, and 1,150 patients (0.7% of the total population) had more than 20 ED visits.

We define frequent ED usage as four or more ED visits per year, and infrequent ED usage as three or fewer annual ED visits. This is the threshold most commonly used in the literature [125]. The jumpers, therefore, refer to members with less than four ED visits in the observation year and four or more ED visits in the outcome year. We define a repeater as a member who has four or more ED visits in two consecutive years. In the aggregated data from 2009 to 2012, frequent users constitute 21% of the population; yet they account for 78% of ED visits. Jumpers account for 11% of the records, with the increase in ED visits ranging from 2 to 85; the median increase is 5 and the mean is 5.91.

We further analyze the relationship between each patient's number of ED visits in the observation year and outcome year (see Figure 4.2). We observe a lack of a strong linear relationship between ED usage in two consecutive years ($r = 0.506$). In fact, as illustrated in Table 4.2, 57% (12% / (9% + 12%)) of the frequent users in the observation year are no longer frequent users in the outcome year.

Table 4.1: Features extracted for analysis.

Category	**Feature (based on the observation year)**	**Notes**
Profile	Member masked ID	
	Sex	
	Age	Varies with year
	Birth year	
	Year of service	2009, 2010, 2011, 2012
	Years of consecutive enrollment	1, 2, 3, 4
Dental	Number of dental visits	
	Indicator of any dental visits	Binary variable
	Total number of unique dental providers	
ED	Number of ED visits	
	ED intensity group	Frequent ED, Infrequent ED
	Number of different ED complaints	Based on the CCS Categories
	Number of different ED vendors	
	Number of ED visits divided by number of ED vendors	
	Indicator of any mental health ED visits	Binary variable
	Number of mental health ED visits	
	Number of ED visits per general diagnosis group	19 variables, based on the general categories for ICD-9-CM
	Number of ED visits per CCS diagnostic group	287 variables
	Distribution of NYU ED usage probability over members aggregated ED visits	9 variables

continued on next page

continued from previous page

Category	**Feature (based on the observation year)**	**Notes**
Mental Health	Number of MH visits	
	Total number of unique MH providers	
	Number of visits broken down by MH disease	20 Variables; top 20 CCS-based MH diseases included.
	Number of MH visits divided by number of unique MH providers	
	Indicator of any mental visits	Binary variable
Medical Data	Number of different chronic diseases	Based on the Chronic Condition Indicator for ICD-9-CM
	Number of unique chronic visits	Based on the Chronic Condition Indicator for ICD-9-CM
	Number of visits per chronic disease	100 variables (corresponding to the 100 most frequent chronic diseases) based on the CCS Categories
	Number of outpatient visits	
	Number of inpatient visits	
	Number of primary care visits	
Pharmacy	Number of different pharmacies	
	Number of unique medications	
	Total days of medication supply	
	Total days of opioid medication supply	
Lab Tests	Number of lab tests	

Table 4.2: Distribution summary of frequent and infrequent users.

Observation year	**Outcome year**		
	Frequent users	**Infrequent users**	**Total**
Frequent users	9% (Repeaters)	12%	21%
Infrequent users	12% (Jumpers)	67%	79%
Total	21%	79%	

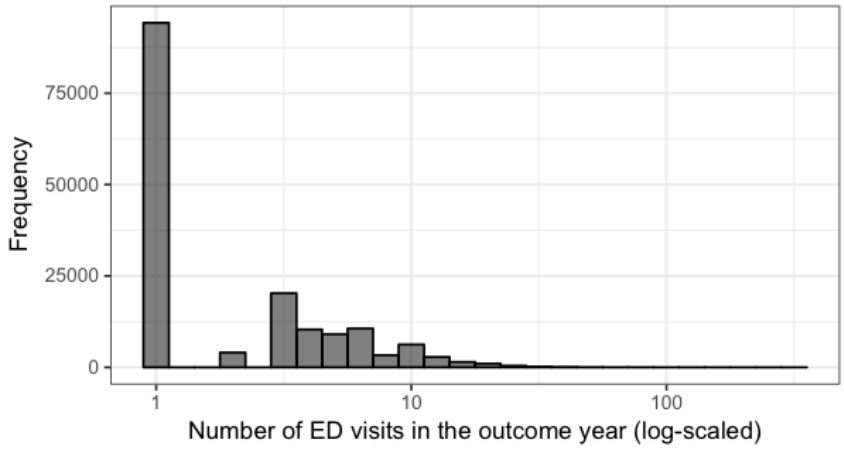

Figure 4.1: Distribution of ED visits number ($\log(x+1)$ transformed) in the outcome year.

4.4 Prediction Modeling and Optimized Selection Strategy

We now describe the methodology used to select candidates for CM, based on the pre-processed claims data with yearly-aggregated features. Due to the scale of CM programs and the constraints on CM resources (e.g., staffing), only a very limited number of patients can be selected into any CM program. Our goal is to select a group of individuals who can maximize the total benefit from the CM

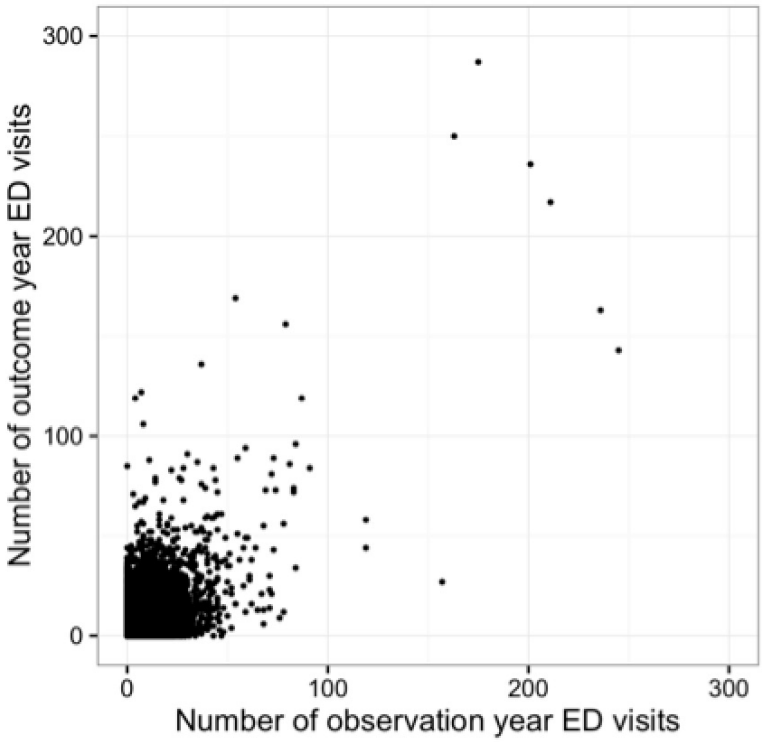

Figure 4.2: Relationship between the number of ED visits in two consecutive years.

program. We partition the candidate population into two groups based on their observation year ED usage level: those who are frequent users in the observation year (thus are eligible to become potential repeaters) and those who are not (thus are eligible to become potential jumpers). We will then combine the prediction models in order to find the optimal combination of potential jumpers and repeaters as CM enrollees.

4.4.1 Prediction of Potential Frequent Users

We divide the problem of predicting future frequent users into two parallel sub-problems—predicting potential jumpers and potential repeaters. The two

groups differ in both their utilization of healthcare services and disease burden; therefore, it is important to model each sub-population independently. Further, this segmentation leads to a simplified analysis of the cost effectiveness of CM, as many parameters are group-based (e.g., the expected benefits for potential repeaters and potential jumpers are different). In the subsections that follow, we introduce the performance measures and setup the binary classification problem, and then present two prediction methods—a baseline model and a machine learning model. Each model will make predictions with respect to the segmented population groups independently.

Data setup and performance measures

Based on members' ED usage in the observation year, we partition the dataset into training (with the observation years being 2009, 2010 and the outcome years being 2010, 2011 respectively) and testing sets (with the observation year being 2011 and the outcome year being 2012) for frequent users and infrequent users, respectively. The testing set has a population size of 62,982, which is the number of enrolled Medicaid population for two consecutive years during 2011 and 2012 in our data. For each prediction problem, we examine the performance of different models by analyzing how accurately each model can select a small number of future frequent users. We use detection accuracy as our performance measure, which is defined by

$$\text{Detection Accuracy} = \frac{TP^x}{TP^x + FP^x} \tag{4.1}$$

where TP^x and FP^x stand for the number of true positive and false positive predic-

tions when selecting the top $x\%$ of the population, respectively. Detection accuracy is related to precision or positive predictive value [175], but we only calculate it for a subset of the population. For instance, suppose that there are 10,000 Medicaid patients, and our goal is to select 1% of the population (i.e., 100 members) for CM, then if out of the top 100 highest risk members suggested by our model, 40 indeed become frequent ED users, then the model detection accuracy is $40/100 = 40\%$. Notice that this rate does not indicate the model's prediction power is less than chance; in fact, a random guess would assign each of the 10,000 members an equal chance of getting selected, resulting in a detection accuracy equal to the prevalence of frequent ED use (which in our dataset is 21%). We do not use more traditional (and global) performance measures such as the AUC, F_1-score, or accuracy, because the overall performance of the model is of less importance for our application. Our interest is in detecting a small set of candidates with high accuracy, motivated by both the high cost of CM and the scale of CM implementations in practice.

Baseline models

In practice, the most commonly used CM candidate selection approach is to use the current ED usage level as the enrollment criteria, assuming individuals will repeat their previous behavior with respect to ED usage. Members are typically selected for CM if their ED usage level exceeded a certain threshold in the past (for example, five or more ED visits in the previous month or year) [29, 173], if the hospital staff or the prescription monitoring program identify issues with a patient's medical usage [85], or through referrals if health care workers believe a patient would benefit from enrollment. Therefore, we define a baseline model that uses the number

of ED visits from the observation year as the prediction of patients' ED usage in the outcome year, and then selects the patients with the highest predicted ED usage as CM enrollees (the current high utilizers). To break ties among members with an equal number of ED visits, we also rank members based on their total number of medical visits. The baseline prediction models are, therefore, based solely on prior rankings of an individual's ED and medical visits. In order to identify both potential jumpers and repeaters, we apply the baseline approach to the current infrequent and frequent users, respectively. For instance, if the number of ED visits and the number of total medical visits in the observation year for members A, B, C, D, E, and F are 250, 10, 10, 3, 3, and 0 and 300, 300, 50, 30, 20, and 20, respectively, the baseline model assumes the members will preserve their relative ED and medical usage level in the result year. As the ED visits from members A, B, and C are at least 4 and 250 is the largest among the three, member A will be ranked first in the potential repeater CM enrollee list, and members B and C will be ranked as second and third based on their total medical visits (because they have the same number of ED visits). Similarly, members D, E, and F will be ranked first, second, and third in the potential jumper enrollee list, based on their ED and medical visits in the observation year. Notice that this enables us to label infrequent users in the observation year as frequent users in the outcome year if both their ED and medical visit numbers are relatively high among the current infrequent users group.

Machine learning models

We develop supervised machine learning models to generate risk scores of future jumpers and repeaters for all individuals based on the training sets among

infrequent users and frequent users, respectively. By prioritizing based on the predicted probabilities of being frequent users, we select the top members from each group as CM enrollees.

Our classification problems are challenging due to severe multicollinearity among features, relatively low frequency of the dependent class (amongst infrequent ED users) and the curse of high-dimensionality. Of the numerous binary classifiers that we trained (see Appendix 4.B for the performance of a selected subset of classifiers), the Extreme Gradient Boosting (implemented with XGBoost) achieved the best predictive accuracy, followed by a boosted tree algorithm (implemented with C5.0) and Linear Discriminant Analysis (LDA). In addition, we build ensemble models using the predicted outcome probabilities produced by the different base classifiers as features. We tested numerous ensemble models utilizing different combinations of features and algorithms (such as logistic regression, support vector machine, and boosted trees). For each targeted population percentage in the training set, we select the ensemble model that maximize its detection. All parameter tuning and model selection for these models was performed on the training set via cross validation. For the prediction of jumpers, the performance of a logistic-regression-based ensemble—which uses predicted outcome probabilities from logistic regression, LDA, C5.0, and XGBoost models as features—was frequently selected as the final ensemble model. A similar approach was applied to select ensemble models to predict the repeaters. The overall most successful ensemble model used the C5.0 and XGBoost predictions as features. The ensemble models enhance the predictive performance, and are, therefore, applied to the testing set as our final

prediction models (referred to as the "ensemble model" in both the jumper and repeater prediction settings).

4.4.2 Prediction Results

We compare the ensemble model performance with the baseline model on the independent test set (comprised of the last observation year of data). Figure 4.3 summarizes the results of the prediction models for jumpers and repeaters in the test sets. We observe that the machine learning model significantly improves the detection accuracy of potential jumpers, especially when selecting only a small proportion of the population. For example, when targeting the top 0.3% of the jumper population (approximately 189 members), the ensemble model outperforms the baseline model with a detection accuracy of 50% compared to a detection accuracy of 30% for the baseline. However, the models perform similarly when predicting future repeaters, with less than 5% absolute increase in detection accuracy over the baseline model. These results indicate that the baseline model—using only the number of observation year ED visits and counts of medical claims—achieves a very high detection accuracy when predicting future frequent users from a group of current frequent users. In addition, current frequent users with the highest number of ED visits tend to remain frequent users in the future.

Based on the feature importance of the base classifiers XGBoost, C5.0, and LDA on the training set, we observe that age, gender, the number of ED visits, the number of outpatient and primary care visits, the number of primary care treatable

ED visits, and presentation of certain types of chronic diseases (such as asthma) and mental health issues (such as attention deficit and conduct disorder) in the observation year are key indicators in predicating potential jumpers and repeaters. In addition, both jumpers and repeaters tend to present large numbers of visits for chronic care, large numbers of different medications, and higher numbers of dental visits and dental providers. Repeaters are also more likely to have large numbers of different chronic disease related with ED visits during the observation year. By comparison, hyperlipidema (an abnormally high concentration of fats or lipids in the blood) is associated with predicting future jumpers.

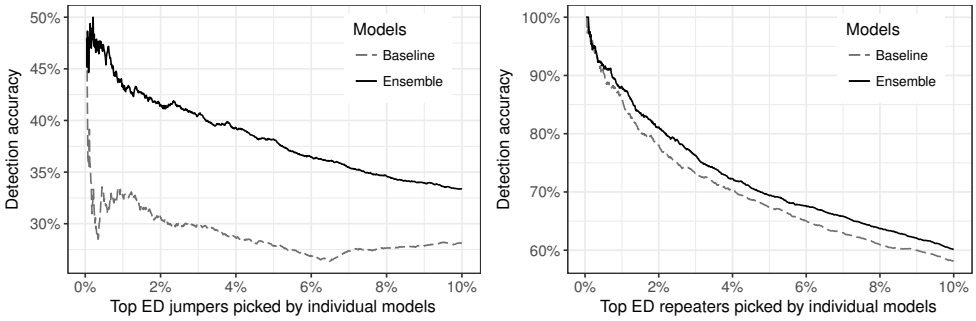

Figure 4.3: Detection accuracy of potential ED jumpers (left) and repeaters (right) on the test set.

4.4.3 Using Prediction Models to Maximize the Impact of Case Management

As previously stated, research suggests that breaking existing habits requires more effort than starting new habits under behavior change interventions. Therefore,

compared with repeaters who have formed the habit of heavily utilizing the ED, jumpers may be more easily influenced. In other words, the average efficacy of case managing jumpers is likely higher than that for repeaters. Therefore, though repeaters can be more accurately identified from our prediction models, the overall effectiveness of a CM program that exclusively enrolls repeaters may be limited due to the low efficacy for this particular group.

Based on this reasoning, we formulate the enrollee selection problem as follows. Suppose our objective is to enroll $x\%$ of a population of size N (which in our case is the size of the test set), who will benefit the most from participation in CM in the outcome year. Given an ordered list of individuals based on their likelihood of becoming jumpers or repeaters using our prediction models, intuitively, we can select the top $[\lambda \cdot x\% \cdot N]$ potential jumpers and the top $[x\% \cdot N] - [\lambda \cdot x\% \cdot N]$ (or approximately $[(1-\lambda) \cdot x\% \cdot N]$ potential repeaters, where $[\ \ ]$ is the integer after rounding down. Note that selecting the top $[x\% \cdot N]$ potential jumpers only, or the top $[x\% \cdot N]$ potential repeaters only, are special cases with $\lambda = 1$ and $\lambda = 0$, respectively (which are discussed in Appendix 4.D). One of the main motivations behind such combined strategy is that the performance of classification models is uneven over the population; in particular, as the detection accuracy is usually significantly higher for the members identified as being the highest risk, the overall impact of CM may be improved by enrolling high-risk jumpers as well as high-risk repeaters.

To characterize the different efficacy levels of CM for jumpers and repeaters, we introduce the efficacy level e as the fraction of prevented ED visits due to enrollment in a CM program. For instance, if on average, the CM program can reduce

the expected ED visits by 30%, then $e = 0.3$. We assume that for the same CM program (i.e., the same healthcare organization), e is invariant with regard to different selection strategies, and is only affected by the types of the enrollees (i.e., potential jumpers or repeaters) that the CM program is targeting. Intuitively, e measures the effectiveness of intervening with respect to different groups of patients under a particular CM program. The higher the efficacy, the more successful is the CM program, and the easier it is to influence members' healthcare utilization. As preventing unnecessary ED visits is easier for jumpers than repeaters, we assume $e_J > e_R$.

As a proxy for benefit, we model the cost savings from the members if selected as CM enrollees. When targeting the top $x\%$ of the population, the total savings (or impact) can be calculated and is denoted by $Saving_S^x$. Intuitively, $Saving_S^x$ measures the balance between the savings from successfully preventing unnecessary ED visits and the costs incurred for CM, and is driven by the accuracy of the prediction models for identifying potential frequent users, CM efficacy levels for potential jumpers and repeaters, the population characteristics including the relative proportion of jumpers and repeaters, as well as various costs associated with CM and the utilization of other healthcare resources (for calculation details, please refer to Appendix 4.C). Our goal is to find the optimal combination of potential repeaters and jumpers to enroll in CM in order to maximize the total savings, which can be expressed as:

$$Saving_{S_{mixed}}^x = \lambda \cdot Saving_{S_J}^{\lambda x} + (1-\lambda) \cdot Saving_{S_R}^{(1-\lambda)x} \tag{4.2}$$

where S_J, S_R, and S_{Mixed}, refer to the policies of enrolling jumpers exclusively, en-

rolling repeaters exclusively, and enrolling both jumpers and repeaters, respectively.

Based on the predictive results, we can search for the optimal value of λ to maximize the total savings for a given set of $x\%$, e_J, e_R, and costs. For the convenience of demonstrating our results, we assume the cost for CM per year per individual is \$2000, the cost for a single ED visit is \$2000, and the cost for the alternative treatment is \$200. These numbers are based on estimates from recent literature [220, 110, 2]. We have also experimented with various combinations of cost parameters and observe similar trends as in the analysis below. The largest uncertainty is in the estimate for the efficacy of the CM program. As a result, we investigate a range of values for e_J and e_R and find the corresponding optimal value of λ and the resulting savings for each scenario.

Enrollment strategy

We aim to find an optimal mixed strategy that results in the largest savings by combining jumpers and repeaters as enrollees under various efficacy levels. We fix the overall percentage of the population to be selected, and find the optimal λ (denoted by λ^*). Figure 4.4 summarizes the numerical results of these experiments. For instance, suppose a CM program has efficacy levels $e_J = 0.5$ and $e_R = 0.1$ for jumpers and repeaters, respectively, and the objective is to select 1% of the members as CM enrollees. Then, since $\lambda^* = 0.69$ (as indicated by P1 on Figure 4.5), including 69% jumpers and 31% repeaters will lead to the largest savings. From Figure 4.6, we can also tell the corresponding savings is \$1.23 million (as indicated by P2 on Figure 4.6).

More generally, from Figure 4.4 (left), we observe that, as the targeted size of

the CM program increases (left to right), λ^* increases. This means that an increased proportion of potential jumpers should be included, as the size of the enrolled population grows. Meanwhile, Figure 4.4 (right) illustrates that in order for a CM program to result in positive savings, the required level of e_J and e_R increases as $x\%$ increases. As the selected enrollee population increases, the false-positive prediction rates increase, requiring higher efficacy in order to generate positive savings. As long as e_R and e_J are sufficiently large (for example, $e_J > 0.4$ and $e_R > 0.2$), increasing the number of CM enrollees will increase the proportion of jumpers included and generate positive savings. We also observe that when e_R is relatively small, one should always include jumpers as a part of the enrollee pool. For instance, when targeting 0.1% of the population (Figure 4.4a(1)), as long as $e_R < 0.35e_J$, the optimal saving strategy always includes enrolling jumpers. Likewise, when targeting 1% (Figure 4.4b(1)) and 10% (Figure 4.4c(1)) of the population, the CM program should include jumpers as long as $e_R < 0.94e_J$ and $e_R < e_J$, respectively. The benefit from including more jumpers decreases as e_R increases, until $e_J = e_R$, in which case λ^* is always 0, meaning that we should only enroll repeaters in this scenario. When targeting 0.1%, 1%, and 10% of the population if $e_R < 0.04e_J$, $e_R < 0.08e_J$, and $e_R < 0.15e_J$, respectively, the program will achieve its maximal savings by enrolling jumpers only. In most cases, a mix enrollment of repeaters and jumpers will be optimal for CM programs.

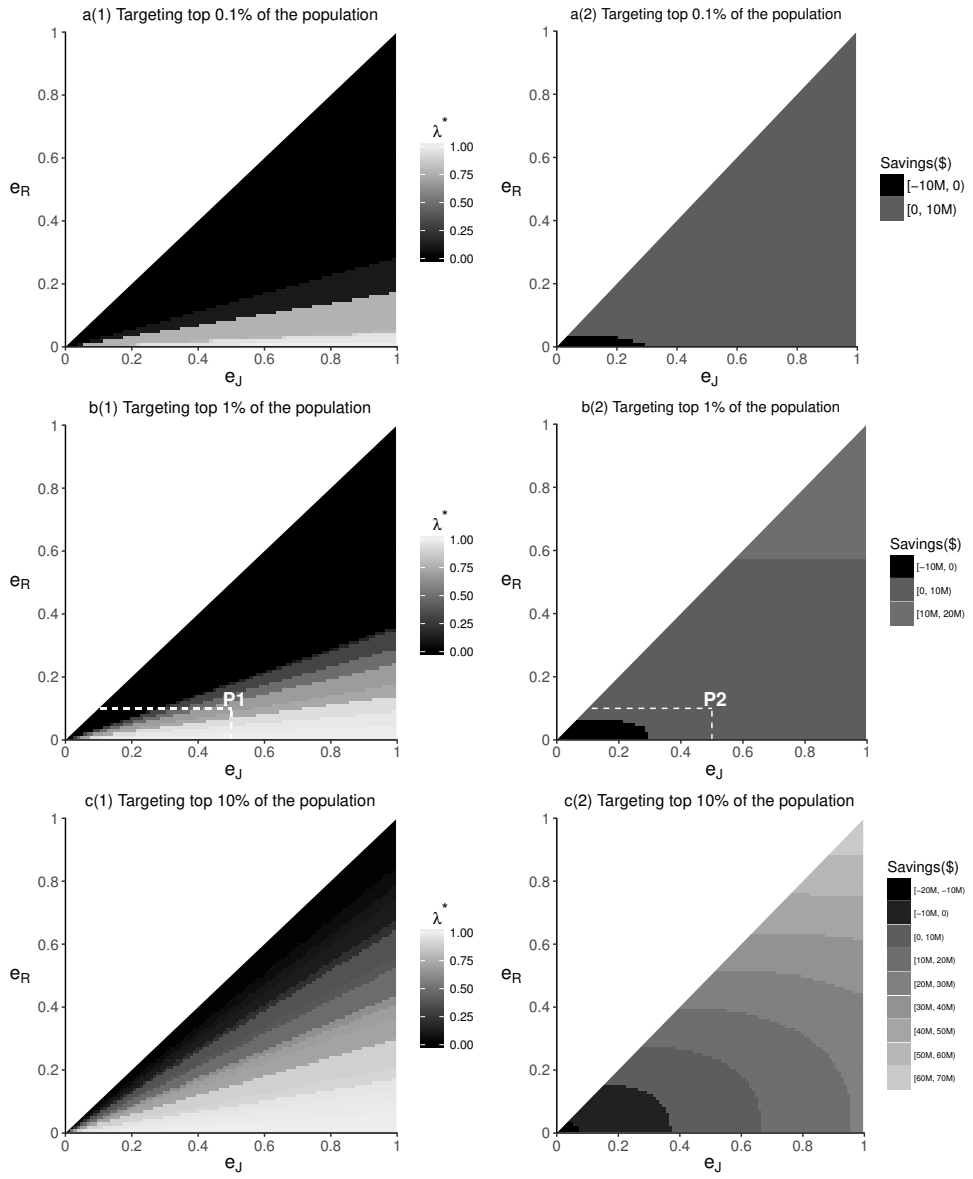

Figure 4.4: The optimal combination (captured by λ^*) of potential jumpers and repeaters (left) and the resulting maximized CM savings (right) when enrolling 0.1% (upper), 1% (middle), and 10% (bottom) of the population under any pair of (e_J, e_R). As e_J is always greater than e_R, therefore the upper triangles of all plots are empty.

4.5 Conclusions and Future Directions

In this chapter, we propose a framework for intelligently selecting candidates for enrollment in CM programs using prediction models for frequent ED usage, and evaluate the effectiveness of several selection strategies. Motivated by the difference in managing members with different prior usage patterns, we introduce the notion of jumpers and repeaters within the context of ED usage, and develop prediction models using historical claims data. Healthcare resources are limited; therefore, it is critical to include individuals who are likely to benefit the most from interventions. We were able to show that a traditional selection strategy works well for targeting potential repeaters; however, this strategy will not result in positive savings when the CM efficacy level is low. By comparing our approach with traditional methods, we demonstrate how our strategy can potentially improve the benefit of CM programs to potential members and the associated savings. We also show that as the number of selected enrollees increases, a larger proportion of potential jumpers should be included to maximize the savings. Also, under a fixed efficacy level, as the program size increases, the respective savings or losses usually increase. Therefore, financially, it is important for the CM programs to estimate their efficacy (or at least the ratio of efficacy between managing repeaters vs. jumpers)—and costs and potential benefits—before expanding their program, and to optimize their current enrollment strategies.

To our knowledge, this is the first study exploring the cost effectiveness of CM enrollee selection strategies based on machine learning predictions. Our framework

offers increased cost effectiveness over traditional enrollment strategies that solely target frequent users for CM, as we also aim to prevent infrequent users from becoming frequent users. Our framework is adaptable to various prediction algorithms, patient populations, and efficacy levels for various interventions, which would ideally be estimated from historical data. Based on the number of staff and amount of resources at hand, CM programs can easily determine how many enrollees to target, and the associated savings to expect. Furthermore, healthcare organizations can estimate their costs of the program, individual ED visits, and of other relevant services. Healthcare practitioners and decision-makers can, therefore, pursue the most cost effective policies to optimize the quality of care and healthcare costs at the same time, noting that in its current form, our framework does not incorporate other tangible benefits of CM programs (e.g., patient wellbeing).

An alternative approach to identifying candidates for case management is to predict the number of ED visits directly (using regression) and select those members with the largest expected number of ED visits as CM enrollees. In our experiments, regression-based models (e.g., log-linear, Poisson, and tree-based regressions) did not perform well. In fact, even after taking into account the over-dispersion and large number of zero outcomes (i.e., using Zero-augmented models such as Zero inflated negative binomial regression and Hurdle models) [128, 168], prediction of a count outcome did not improve our results. Both of the two Zero-augmented models achieved comparable prediction performance (both in terms of detection accuracy and number of detected potential ED numbers) with the ensemble model for predicting the jumpers, yet they underperform the ensemble model for predicting

the repeaters, as there are a smaller proportion of zero outcomes.

Our study has several limitations. Some features that have been previously associated with frequent ED usage are not available in our data. For example, previous studies have suggested that socioeconomic factors such as income [97, 135, 28] and race and ethnicity [129] can all influence ED usage. Individual health behavior and habits such as exercise, diet, and smoking can also impact members' health outcomes. In addition, our data is aggregated at the yearly level due to privacy concerns. Incorporating more granular temporal patterns (e.g., daily, monthly activity) of repeaters and jumpers has the potential to improve our predictions and subsequent enrollee selection, and is an interesting future direction.

For future research, we would like to explore CM efficacy in more detail, as it may depend on an individual's historical ED usage levels. One can also incorporate the "stickiness" (i.e., the consistency in maintaining certain ED usage frequency) of frequent users or infrequent users into the efficacy function, as past behavioral frequency provides an adequate proxy for habit resistance [232]. Intuitively, frequent users who have extensively used the ED for many years would have a lower efficacy level than those who transition between the frequent and infrequent usage levels. It may also be worth investigating the role of preventable ED visits, and primary-care treatable ED visits on efficacy. We could also study the impact of individual-based costs for use of the ED and alternative treatments, such as distribution-based costs with parameters specific to each individual, as well as the influence of the threshold (e.g., 4 or higher) that defines frequent ED users. How does the prediction performance and λ^* change as a function of the threshold? How robust

are the models for different proportions of selected enrollees? Finally, we would like to extend the framework for multi-year or partial-year prediction, in which multiple years of claims history or partially-enrolled yearly records are used to predict future ED usage levels. As the cost of CM programs is very high and the research on cost effectiveness is lacking (especially from the perspective of efficacy), we expect to see more studies in this area. We would also expect to see the application of the ideas of jumpers and repeaters in other settings, such as in studying online usage or purchasing behavior.

4.A Appendix – Data Processing Procedures and Feature Generation

In the main text we provide an overview of the features used for the prediction models in Table 1. Below we provide additional details on the data processing. The raw datasets are processed into a format suitable for analysis using the following steps:

a. Calculate days enrolled for each year;

b. Age and gender adjustments;

c. Claim aggregation;

d. ED visit categorization based on ICD-9-CM diagnosis codes;

e. Extract yearly information for individual members from five datasets.

Details follow below.

a. Calculate days enrolled for each year

Based on the individual insurance enrollment information (with year and date of enrollment), we calculate how many days an individual is enrolled in a specific year. We set the allowable gap (AG) to 15 days, which means that any yearly record for a member enrolled for less than 350 days is excluded from the analysis. Records with at least 350 days of enrollment are eligible for analysis. The number of eligible yearly records is 255,922, corresponding to 184,929 individuals. Note that AG = 30 days only increases the sample size by 0.47% and AG = 60 days increases the sample size by 10.3%. In addition, 1% of the enrolled members have no claims and are excluded from our analysis.

b. Age and gender adjustments

By combining all gender and age information extracted from the five datasets for each individual, we find that 71.85% of the members in our datasets have an inconsistent birth year (mostly within a 2-year variation) and 0.46% of the individuals have an inconsistent gender. We, therefore, select for each member the most frequent birth year and gender as their birth year and gender for our analysis. Based on the birth year, we generate age for each individual in a specific year.

c. Claim aggregation

From the raw datasets, we extract the number of unique ED visits, inpatient visits, total outpatient visits, mental health visits, and dental visits based on the claim ID. Each visit is associated with a unique claim ID, and multiple

visits and treatments under the same claim ID are considered to be the same visit. All mental health outpatient claims (independent of provider type) are combined into a total count of mental health visits outside the ED.

d. Add diagnoses categories to ED and medical visits based on their ICD-9-CM diagnosis codes

We used three methods to group the 6114 diagnoses codes in our ED data:

I. General Diagnosis Category

Based on National Center for Health Statistics (1980), 19 disease-based diagnosis groups are generated.

II. CCS Diagnosis Category

The Clinical Classifications Software (CCS) maps ICD-9-CM diagnosis codes into 285 single-level CCS diagnosis categories [94]. We also identify chronic medical visits using the Chronic Condition Indicator [93].

III. NYU ED Visit Severity Algorithm

This algorithm maps the diagnosis code associated with each ED claim (ICD-9-CM based) into a probability distribution over the following 9 categories: Emergent and not Preventable/Avoidable; Emergent but Preventable/Avoidable; Emergent but Primary Care Treatable; Non-emergent; Alcohol related; Drug related; Injury; Mental Health; Others [26]. In this way, one can obtain the historical distribution of usage behind each ED visit. For instance, an ED visit with the diagnosis code

005.9 (i.e., "food poisoning") corresponds to the following probability distribution: $[0.17, 0, 0.46, 0.37, 0, 0, 0, 0, 0]$, which means that based on substantial historical data, with probability 0.17 such a visit is emergent and not preventable, with probability 0.46 such a visit is emergent but primary care treatable, and with probability 0.37 it is non-emergent. The algorithm differentiates ED visits based on the need for hospitalization and/or mortality risk. The algorithm is useful in studying ED utilization and evaluating policies that aim to reduce non-emergent and avoidable visits to the ED (as defined in Table A1). We apply this algorithm to each ED claim in our dataset, and obtain the average ED usage probability distribution over members' aggregated ED visits. For instance, if in one year an individual has 2 ED visits with NYU ED distributions $[0, 0.4, 0.4, 0, 0, 0, 0, 0, 0.2]$ and $[0.7, 0.2, 0.1, 0, 0, 0, 0, 0, 0]$ respectively, we can calculate his average ED usage distribution as $[0.35, 0.3, 0.25, 0, 0, 0, 0, 0, 0.1]$.

Table 4.3: ED visit group based on NYU ED algorithm.

ED Visit Group	NYU ED Algorithm Category
Preventable ED Visits	Non-Emergent Primary care treatable ED care needed; preventable/avoidable
Non-Preventable ED Visits	ED care needed; not preventable/avoidable

e. Extract yearly information for individual members from five datasets

For each member and calendar year, we extract a total of 464 yearly features, in addition to the outcome of whether or not they were a frequent user in the next year.

4.B Appendix – Comparison of Different Classifiers

During the model building phase, we implemented a variety of classifiers in order to predict future jumpers and repeaters. We applied appropriate resampling methods (such as down sampling and SMOTE sampling) to correct the data imbalance when necessary (e.g., for algorithms that do not cope well with imbalanced samples, such as boosted trees). Below we report the predictive performance of some of the base classifiers as well as the ensemble models (as detailed in the main text). Figures 4.5 and 4.5 show the performance of a selected subset of classifiers on the test set: C5.0 boosting tree (C50), Logistic Regression (LR), Linear Discriminant Analysis (LDA), Feed-Forward Neural Networks (NNet), Random Forest (RF), Extreme Gradient Boosting (XGBoost), and the final ensemble model. In Table B1 below, we summarize the performance of the various algorithms as functions of the enrolled population size.

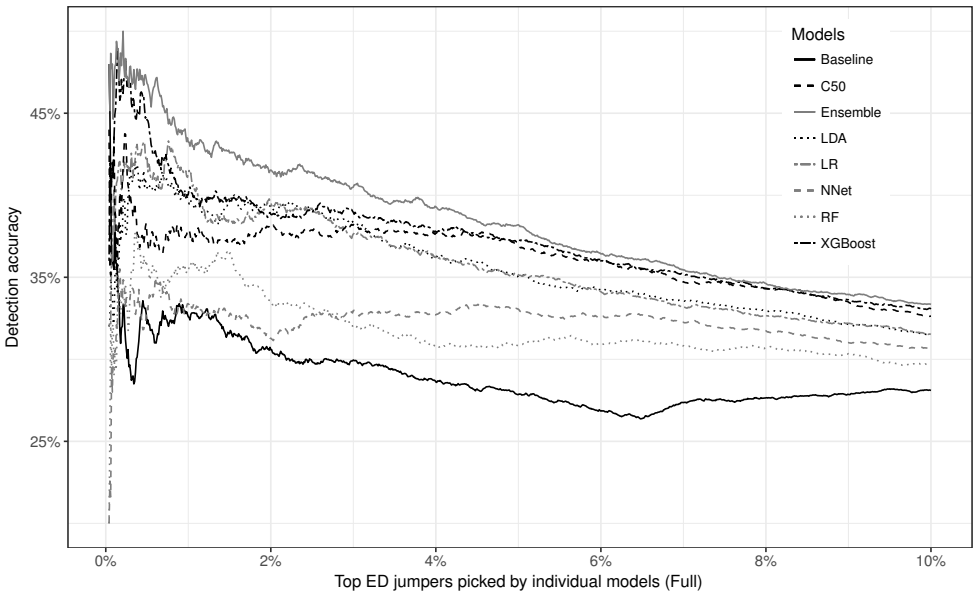

Figure 4.5: Detection accuracy of several classifiers of potential ED jumpers on the test set.

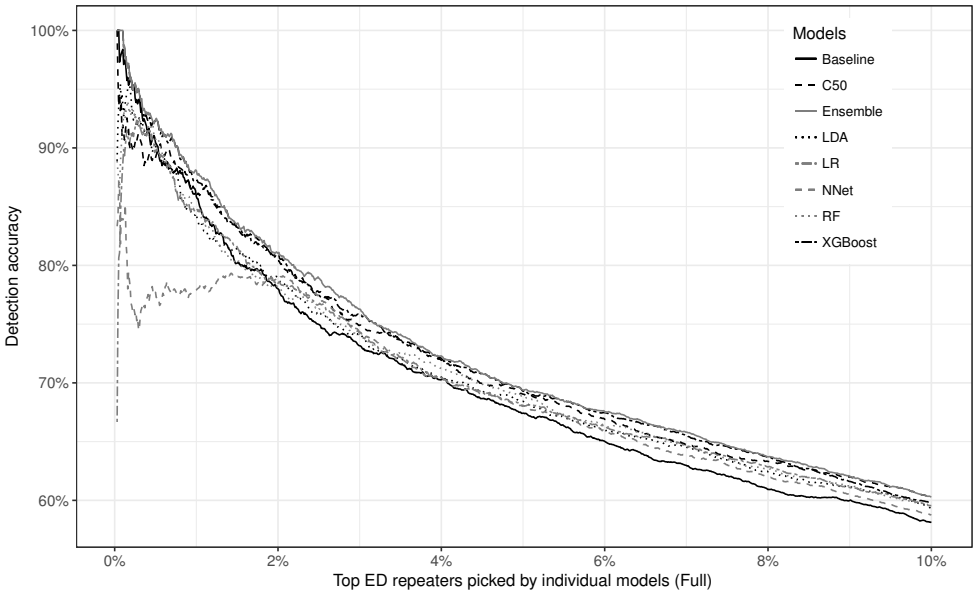

Figure 4.6: Detection accuracy of several classifiers of potential ED repeaters on the test set.

Table 4.4: Performance of various algorithms.

	Detection Accuracy (on a test set of 62982 people)					
Classifier	Jumper Prediction			Repeater Prediction		
Target Percentage	0.1% (63 members)	1% (630 members)	5% (3149 members)	0.1% (63 members)	1% (630 members)	5% (3149 members)
Baseline	0.3548	0.3259	0.2785	0.9839	0.8585	0.6742
C5.0	0.4355	0.3784	0.3687	0.9194	0.8585	0.6904
Ensemble	**0.4677**	**0.4324**	**0.382**	**1**	**0.8812**	**0.6941**
LDA	0.3226	0.3975	0.3515	0.9355	0.8442	0.6843
LR	0.4032	0.4134	0.3509	0.8871	0.8458	0.6809
Nnet	0.3065	0.3339	0.3306	0.8387	0.779	0.6802
RF	0.2903	0.3561	0.309	0.9194	0.8458	0.6882
XGBoost	0.4355	0.4006	0.3722	1	0.8712	0.6929

4.C Appendix – Formulas for Expected Savings

Suppose the size of the population of interest is N, which in our cases is the size of the test set. Our objective is to select $x\%$ of the population who will benefit the most from enrollment in CM. As a proxy for benefit, we model the cost savings from the members if selected as CM enrollees. In order to formulate the selection strategy problem, we first introduce some definitions:

- c_{CM}^{i}: Annual cost of CM for enrolled patient i.
- $c_{ED}^{i,j}$: Cost of an ED visit for patient i during his or her j-th ED visit in the outcome year.
- $c_{ALT}^{i,j}$: Cost of alternative treatment for patient i's prevented attempt of j-th ED visit (if patient i is under CM program) in the outcome year.
- n_{ED}^{i}: Number of ED visits for patient i (or attempts, without CM interventions) in the outcome year.
- I_S^x: Set of indices for individuals selected for CM under selection strategy S. $I_S^x \subseteq \{1, 2, \ldots, N\}$.
- K_S^x: Number of the top $x\%$ riskiest population selected for CM under selection strategy S. $K_S^x = [N \cdot x\%]$. The cardinality of the set I_S^x is K_S^x.
- e_J, e_R: Efficacy of CM for jumpers and repeaters, respectively.
- $Saving_S^x$: Total amount of money saved by CM using selection strategy S when targeting the top $x\%$ of the population.

To compute the expected savings under a certain CM enrollee selection strategy S, the total cost from ED visits to be paid without any CM strategy in the outcome year is:

$$C_{CM\_none} = \sum_{i=1}^{N} \sum_{j=0}^{n_{ED}^i} c_{ED}^{i,j}. \tag{4.3}$$

Selecting the top $x\%$ riskiest population as enrollees, the total cost to be paid under CM enrollee selection strategy S is:

$$\begin{aligned} C_S^x &= \sum_{i \in I_S^x} c_{CM}^i + e \cdot \sum_{i \in I_S^x} \sum_{j=0}^{n_{ED}^i} c_{ALT}^{i,j} + (1-e) \sum_{i \in I_S^x} \sum_{j=0}^{n_{ED}^i} c_{ED}^{i,j} + \sum_{i \notin I_S^x} \sum_{j=0}^{n_{ED}^i} c_{ED}^{i,j} \\ &= \sum_{i \in I_S^x} c_{CM}^i + e \cdot \sum_{i \in I_S^x} \sum_{j=0}^{n_{ED}^i} c_{ALT}^{i,j} + \sum_{i=1}^{N} \sum_{j=0}^{n_{ED}^i} c_{ED}^{i,j} - e \cdot \sum_{i \in I_S^x} \sum_{j=0}^{n_{ED}^i} c_{ED}^{i,j}, \end{aligned} \tag{4.4}$$

which is the sum of the cost for CM on the subset of the population under CM, the cost spent on alternative treatment other than ED visits for these patients (as the direct result of CM diverting patients away from the ED), and the cost spent on ED visits for the remaining ED visits (including those who are not prevented by CM and those who are not under CM). Therefore, the savings (or impact) resulting from CM enrollee selection strategy S is:

$$Saving_S^x = C_{CM\_none} - c_S^x = e \cdot \sum_{i \in I_S^x} \sum_{j=0}^{n_{ED}^i} \left(c_{ED}^{i,j} - c_{ALT}^{i,j} \right) - \sum_{i \in I_S^x} c_{CM}^i. \tag{4.5}$$

This intuitively represents the balance between the money saved from diverting patients away from the ED under CM and the cost of the CM program.

To simplify the numerical analysis (and since the cost for individual claims is not available in our dataset), we further assume the cost for CM, cost for each ED visit, and cost for each alternative treatment are the same across the full population,

that is, $c^i_{CM} = c_{CM}$, $c^{i,j}_{ED} = c_{ED}$ and $c^{i,j}_{ALT} = c_{ALT}$ for all $1 \leq i \leq K^x_S$. Then,

$$C_{CM\_none} = c_{ED} \sum_{i=1}^{N} n^i_{ED}. \tag{4.6}$$

$$\begin{aligned} C^x_S &= c_{CM} \cdot K^x_S + e \cdot c_{ALT} \sum_{i \in I^x_S} n^i_{ED} + (1-e) c_{ED} \sum_{i \in I^x_S} n^i_{ED} + \sum_{i \notin I^x_S} n^i_{ED} \\ &= c_{CM} \cdot K^x_S + e \cdot c_{ALT} \sum_{i \in I^x_S} n^i_{ED} + c_{ED} \sum_{i=1}^{N} n^i_{ED} - e \cdot c_{ED} \sum_{i \in I^x_S} n^i_{ED}. \end{aligned} \tag{4.7}$$

Therefore, (4.5) simplifies to:

$$Saving^x_S = C_{CM\_none} - c^x_S = e\,(c_{ED} - c_{ALT}) \sum_{i \in I^x_S} n^i_{ED} - c_{CM} \cdot K^x_S. \tag{4.8}$$

Based on this formula, the savings resulting from exclusively targeting potential jumpers ($Saving^x_{S_J}$) and repeaters ($Saving^x_{S_R}$) can be calculated. As our goal is to find the optimal combination of potential repeaters and jumpers to enroll for CM in order to maximize the total savings, the total savings from partitioning K^x_S into two sub-candidate populations is:

$$Saving^x_{S_{mixed}} = \lambda \cdot Saving^{\lambda x}_{S_J} + (1-\lambda) \cdot Saving^{(1-\lambda)x}_{S_R}. \tag{4.2}$$

4.D Appendix – Cost Effectiveness Analysis—Exclusive Strategies

Strategy 1: Jumpers only

Below we consider two simple strategies and contrast them to the mixed strategy from the main chapter. In the first scenario, we assuming only jumpers are targeted as CM enrollees, and calculate the corresponding savings from a CM program using the baseline and ensemble models. In Figure 4.7, we observe that the

ensemble model results in much larger savings than the baseline model (reflective of the improvement in predictive performance). This holds true over a broad range of efficacy levels. Given our numerical assumptions on cost of medical care and running the CM program, in order for the CM program to be cost effective, the efficacy level has to be at least 0.3. The minimum efficacy level for cost effectiveness increases as the number of included members increases, which reflects the decreasing detection accuracy as the size of the included population grows. For instance, when targeting the top 1% of the population, the strategy based on the ensemble model requires an efficacy level of 0.3 in order for the CM program to result in positive savings (as indicated by P1 on the graph), compared to 0.4 using the baseline model (as indicated by P2 on the graph). When targeting the top 5% of the population with $e_j = 0.5$, the expected savings using ensemble models for enrollment is much higher than under the baseline prediction (in our case, \$3.26 million versus \$1.02 million). We also observe that as the program size increases, the respective savings or losses increase. For instance, when the efficacy level is below 0.25, increasing the number of CM enrollees (i.e., moving to the right on the x-axis in Figure 4.7) simply results in more losses. On the other hand, if the efficacy level is 0.6, increasing the number of CM enrollees (i.e., increase x%) results in more savings.

Strategy 2: Repeaters only

In this scenario, only repeaters are targeted as CM enrollees, and again we calculate the corresponding savings from a CM program under various efficacy and costs assumptions. In contrast to the jumpers strategy, from Figure 4.8, we observe that using the ensemble model offers no significant improvement over the baseline

model, reflective of the similar predictive performance of the baseline and ensemble models. Comparing Figure 4.8 with Figure 4.7, we also observe that in order to obtain savings, the requirement on CM efficacy for potential repeaters is much lower than for potential jumpers, ranging from 0.01 to 0.14, which is not surprising when considering the higher utilization rates of the repeaters. For example, when targeting the top 1% of the population, the strategy based on ensemble and baseline models both requires an efficacy of 0.06 in order for the CM program to result in positive savings (as indicated by P3 and P4 on the graph). As with Figure 4.7, we observe that as the percentage of individuals targeted increases, the respective savings or losses increase.

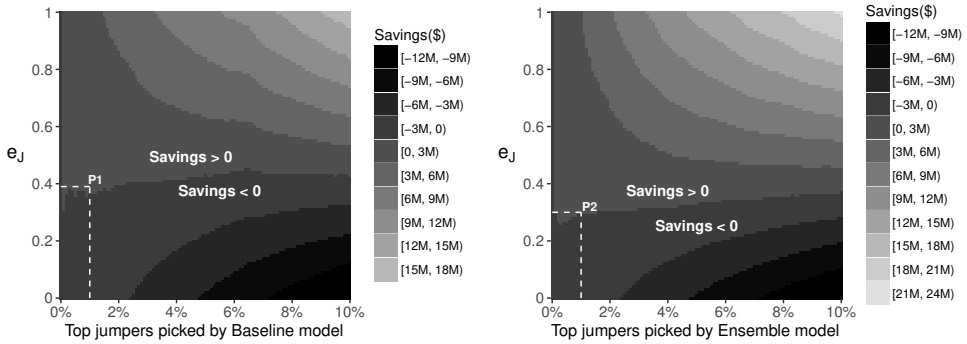

Figure 4.7: Expected savings from targeting potential jumpers exclusively based on predictions from Baseline model (left) and ensemble model (right). The targeted number of CM enrollees (x-axis) and the efficacy level of the CM program (y-axis) determine the corresponding savings based on prediction models. The lighter the shading, more savings are generated.

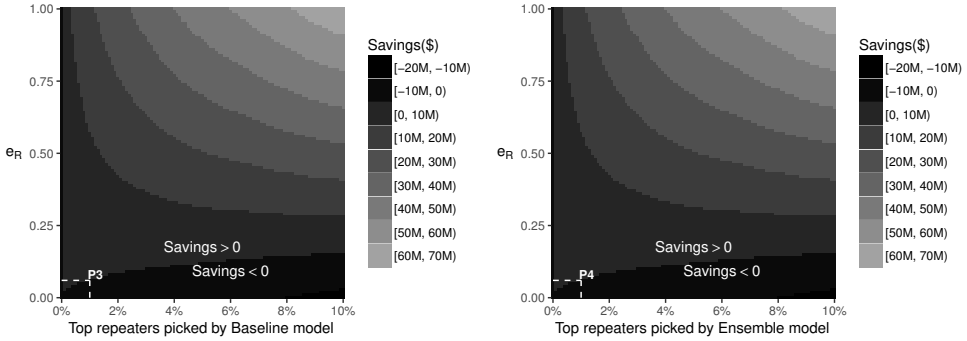

Figure 4.8: Expected savings from targeting potential repeaters exclusively based on predictions from Baseline model (left) and ensemble model (right). The targeted number of CM enrollees (x-axis) and the efficacy level of the CM program (y-axis) determine the corresponding savings based on prediction models. The lighter the shading, more savings are generated.

Chapter 5: Temporal Data in Risk Predictions

5.1 Introduction

Temporal data is of increasing importance in healthcare analytics. The increasing amount of available data from hospitals and medical practices via electronic health records (EHR) and claims data has facilitated detailed exploration of patient histories. Claims data is structured data, which is generated when healthcare providers submit electronic claims to insurance payers to justify payments and recive reimbursements. It captures patients medical conditions and health care activities from healthcare providers, and stores information such as dates of service for laboratory tests, prescriptions, emergency visits, and inpatient admissions and discharges. The patterns within those temporal sequences are indicative of disease progression and patient activities over time, and are important for disease diagnosis and surveillance, patient health management, and policy planning [39, 234, 138].

Efforts have been made to extract and apply temporal information from a sequence of medical events for better diagnostic summary and prediction [45, 20, 234]. Oftentimes, the very first step in temporal modeling requires aggregating time-stamped events in granular units (i.e., by day, week, month) or time intervals (e.g., hospitalization duration) from the raw data [23], and constructing the correspond-

ing temporal sequence in the form of $(X_{t_1}, X_{t_2}, \ldots, X_{t_n})$. The most common way of building temporal sequences is by summing, averaging, or describing the attributes (e.g., increasing trend, convex shape, maximum) of the underlying data over segmented intervals [202, 98]. Other sequence representation methods include representing temporal sequences as graphs [138], event matrixes [241], and symbolic languages and grammars that characterize the relationship among time-dependent events [166]. There are multiple challenges associated with the sequence construction process:

1. The level to which the claims and health records should be aggregated. Preferably, the temporal information should be aggregated in a high-level, meaningful way. Such aggregation is usually facilitated by appropriate domain-specific knowledge and the research goal [203].

2. The information represented by the time sequence. There may be multiple dimensions of information associated with one activity in the claims data, but such information may be hard to be fully characterized within one temporal sequence. For example, to construct a sequence describing a patient's claim frequency, the simplest way is to form the sequence by recording the count of claims per day. However, the information of claim diagnosis and the severity (which can sometimes be obtained by transforming the claim diagnosis) are lost. Besides, the input value may be from different abstraction levels and data types (e.g., cost may be a numerical variable, whereas severity may be a categorical variable) [210]. Therefore, one needs to map claims into a time

series as real numbers in a meaningful and ordered way.

3. Irregularly sampled time series. Unlike standard time series from finance or climate studies with numerical records at uniform points in time, temporal sequences describing claims activities are usually sparse and uneven. As the claims data are recorded only when a medical event happens, the event of interest is usually of a low frequency and the occurrence is non-periodic. The excessive number of zeros in a sequence would make the direct application of traditional time series models (e.g., autoregressive moving average (ARMA), autoregressive integrated moving average (ARIMA)) and similarity measures (such as dynamic time warping [197]) inapplicable; therefore, temporal aggregation or extraction is needed.

4. The complex interaction between different activities. There are sequential or linked hospital activities. For example, during one emergency visit, the care procedures may involve triage, lab tests, image exams, intake, treatment, and the back and forth steps between each of them [87].

5. The integration of static and dynamic features. As claims records are multivariate [20], researchers need to integrate non-temporal features (such as administrative and demographic data) with temporal features into the same model by finding a shared method of representation.

6. Missing data, unrecorded activities, and invalid records may exist, due to mistakes or patients' change of health plans [234] or multiple health plans.

After transforming the raw data into a structured time series, one can proceed with temporal data mining, using various modeling techniques that we will describe in Section 5.3.2.2.

As it usually requires effort and domain knowledge to convert claims data to temporal sequences and to select appropriate modeling techniques, it is important to explore the role of temporal features in predicting health outcomes, and the performance of different temporal representation schemes (i.e., based on granularity, such as day or week) related to different diagnostic prediction problems. As the claims date information is sometimes stripped due to patient privacy concerns as defined by Health Insurance Portability and Accountability Act of 1996 (HIPAA) [3], it is worthwhile to analyze the improvement that precise date information brings, in comparison to the yearly aggregated claims data.

As a case study on the relevance of detailed temporal features, we focus on the risk prediction of inpatient visits related to diabetes or its complications among diabetic patients. Diabetes is a chronic disease impacting 9.5% of the adult population worldwide in 2008 [68]. It is especially costly because of its devastating complications like blindness, kidney disease, amputations, and heart diseases [68]. The national cost of diabetes in the U.S. in 2012 amounted to $245 billion and each diabetic patient annually spends 2.3 times more on healthcare annually than non-diabetes [10]. As inpatient visits are more expensive than other types of medical services (such as primary care visits) and diabetes complications develop simultaneously or along with diabetes [196], it is important to detect the diabetic patients who are at risk of hospitalization due to diabetes or its complications, and if possible

mitigate their conditions ahead of time. In our study, by aggregating temporal information at certain periodic levels (i.e., daily, weekly, monthly, yearly) and extracting relevant information from the corresponding sequences, we evaluate various ways of incorporating the detailed temporal information related to diabetic treatments and monitoring, and analyze the effectiveness of those methods in predicting diabetic-related (i.e., with diabetes or diabetic complications as the chief complaints) inpatient visits within a specific timeframe. Our goal is to provide empirical evidence to quantify the influence of different aggregation schemes.

The remainder of this chapter is organized in the following manner. Section 5.2 presents a brief literature review of related temporal modeling and diagnostic prediction models. Section 5.3 describes our dataset and methodology. Section 5.4 contains the results of the empirical analysis. In Section 5.5, we summarize implications from our findings and provide directions for future research.

5.2 Related Work

Temporal aggregation and abstraction is of great importance in healthcare. High-level aggregation of time-stamped data can provide physicians with concise, context-sensitive summaries, and is useful for decision support [210]. The basic temporal abstraction can be represented by state abstraction (e.g., High, Normal, Low levels in specific time intervals) and trend abstraction (e.g., decrease or increase) [203]. Complex temporal abstraction can be achieved by aggregating information on multiple intervals as new features or applying functional fit to the data.

Several researchers have studied the level of temporal aggregation from the perspective of economics. Most of their studies are within the context of time series modeling, such as ARIMA and ARMA models [217]. For example, Rossana and Seater [202] studied the effect of temporal aggregation in time series prediction. They find that aggregation causes a loss of information, and quarterly aggregated data is optimal for econometric analysis for modeling unemployment rates as it preserves both variation and robustness in comparison to monthly and yearly data. However, to our knowledge, there is no article focusing on the level and influence of temporal aggregation in health claims data or for diagnostic prediction.

Various data mining methods have been introduced to extract temporal patterns in time stamps or time intervals for clinic and diagnostic study. Among them are unsupervised methods such as temporal association rules, which extend the traditional association rules by incorporating temporal factors to find the frequent associations between intervals in a state sequence [98]. Concaro et al. [45] studied health administrative data among diabetic patients and a control group. They extend temporal association rules to find frequent associations concerning sequences of hybrid events (i.e., point-like events such as primary care visits and interval-like events such as drug usage), and detect patterns specifically related to the diabetic population in comparison with a control group. Batal et al.[20] proposed a recent temporal pattern-mining framework based on association rules and applied it to detect adverse diabetic conditions. They assume that the patterns closer in time to the event of interest to be more predictive than less recent temporal patterns. By mining from the most recent frequent temporal patterns to the least recent ones,

patterns that occur earlier in time would receive less weight and, thus, are less likely to be considered as input features for classification. Their models are applied to diabetic patients, and are able to assign disease labels (in terms of 8 categories) to patients at any time point with relatively high accuracy. However, they fail to take into consideration the severity level of the disease—as having a high risk of cardiovascular disease is not equivalent to having severe cardiovascular disease that requires hospitalization or emergency treatment—and therefore, their model is not applicable for hospitalization risk prediction.

Bayesian network models have also been widely implemented to describe time series data due to its interpretability in classifying and predicting future states [22]. Long [142] used Bayesian networks and time intervals for heart disease prediction. Van der Heijden et al. [234] extended the framework for chronic obstructive pulmonary disease prediction. By adding temporal features that characterize the progression of clinical conditions, their models iteratively search for a network structure that best explains the data and achieves high prediction accuracy (e.g., improving AUC from a baseline of 0.84 to 0.90). However, as Bayesian models require posterior probability estimation through intensive empirical computation, the computational cost grows dramatically as the models get more complex (e.g., the number of treatments increase), making the model less scalable. Besides, since the construction of the network is subjective, a successful implementation of a Bayesian network requires deep understanding of the disease progression, as it is easy to falsely assume every state is connected.

Other machine learning models have also been employed for diagnostic codes

mining, such as temporal graph models for predicting heart failure [138], deep learning models that utilize patients yearly diagnosis codes [161], and hidden Markov methods that model sequential activities [57].

In the following sections, we focus on studying the influence of temporal aggregation levels on predicting inpatient risks over different time horizons. We employ several ways to build temporal sequences to explore the most meaningful temporal representation in claims data. We also mine temporal sequences by utilizing Discrete Fourier Transformations and the K-nearest neighbors algorithm.

5.3 Data and Methods

5.3.1 Data

In this study, we utilize insurance claims data, which are collected when health services providers receive reimbursements. Our datasets contain members' basic demographics, insurance plan type, pharmacy and medical usage, and all types of outpatient, inpatient, and follow-up visits. The diagnosis data is coded according to the International Classification of Diseases, Ninth Revision, Clinical Modification system (ICD-9-CM). We include 29,472 diabetic patients as our study subjects, who had diabetes as their chief or subsequent complaint (note there can be up to four diagnosis codes per claim, with one diagnosis being the primary one) in 2007 and who were continuously enrolled from 2007 to 2013.

To make our data suitable for analysis, we construct member-based profiles by aggregating claims and other information for each member. We map individual diag-

nosis codes into 287 categories using the Clinical Classifications Software (CCS) [93]. To indicate diabetic complications and to quantify the degree of diabetic progression, we utilize the Diabetes Complications Severity Index (DCSI), which is designed to improve the prediction of adverse diabetes outcomes [256]. DCSI associates ICD-9-CM codes with 7 categories that are related to diabetes complications—specifically, cardiovascular disease, nephropathy, retinopathy, peripheral vascular disease, stroke, neuropathy, and metabolic—with a number ranging from 0 (not severe) to 2 (severe). The total sum of DCSI for all claims in one year measures a person's diabetic health condition. In 2007, 39.4% of the members in our data exhibit diabetes complications in their claims (either as chief or subsequent complaints).

Further, we group the 25 claim types (such as optometry evaluation, psychiatric evaluation, and psychiatric hospitalizations) into five major claim categories: diagnostics (lab, tests, and image examination), evaluation and prevention (including primary care), treatment and therapy, ED/urgent care and observation care, and finally inpatient hospitalization. For additional details on the categorization, please refer to Appendix 5.A.

As a preliminary analysis, we utilize patient claims records from the previous year to predict whether a diabetic patient will have a diabetic-related inpatient stay in the following periods, including 1) the first quarter of the following year; 2) the following year; 3) the following two years; and 4) the following three years. We train our machine learning models using claims data from 2007 with true prediction outcomes obtained from 2008-2010 data, and test the model performance on 2008, with true outcomes obtained from 2009-2011.

5.3.2 Methods

As risk prediction can be formulated as a binary classification problem, we start by constructing the learning features that we develop in the first stage of our modeling:

a. Administrative features: gender, age, and insurance type (e.g., Medicare, Medicaid, Commercial)

b. Yearly aggregated non-temporal features: number of claims per CCS diagnostic group (287 features), number of unique claims, number of claims per claim category (5 features based on Appendix 5.A), number of diabetic-related claims per claim category (5 features based on Appendix 5.A), DCSI score, and yearly total cost (2 features, including total cost and reimbursable cost).

c. Yearly aggregated temporal features: number of days with medical visits per claim category (5 features based on Appendix 5.A), number of days with diabetic-related medical visits per claim category (5 features based on Appendix 5.A), and summary statistics related to diabetic-related hospital stays (e.g., maximum, median, and mean length of stay).

d. Sequential-based temporal features: based on each patients daily, weekly, monthly, or quarterly aggregated diabetic-related hospital utilization level, we first generate temporal feature vectors for each diabetic patient to reflect the number or type of claims, and 2) apply various methods to extract temporal information from these feature vectors. In Section 5.3.2.1 and 5.3.2.2, we

explain in detail our methods for temporal sequence generation and temporal learning.

5.3.2.1 Sequence Construction

As the goal of the study is to predict diabetes-related inpatient risk, the ideal temporal sequence should capture the patients diabetes condition, and his/her diabetic-related healthcare utilization. In the following analysis, we assume there is a linkage between patients temporal care patterns and subsequent diabetic-related inpatient visits. In order to represent a patients claims utilization, we propose 11 ways of constructing temporal sequences $\tau^k = (X^k_{t_1}, X^k_{t_2}, \ldots, X^k_{t_{366}})$, $k = 1, \ldots, 11$, each with length 366 (corresponding to 366 days, and if a year has 365 days, the 366^{th} element of the sequence is 0). The construction rules are introduced in Table 5.1. The default value of $X^k_{t_i}$ is 0, meaning that the patient did not utilize medical services on day t_i.

In addition, based on the daily sequence constructed above, we aggregate the daily sequence to weekly, monthly, and quarterly levels by summing the appropriate elements for the defined frequency. The corresponding sequences have lengths 52, 12, and 4, respectively.

5.3.2.2 Sequence Learning

In order to reduce the data dimension and extract the most useful temporal information, we employ several methods to learn patterns from the constructed tem-

poral sequence. For a temporal sequence $\tau^k = (X^k_{t_1}, X^k_{t_2}, \ldots, X^k_{t_{366}})$, $k = 1, \ldots, 11$, where t_i represents the i^{th} arbitrary time unit (such as day, week, month, quarter, year), $X^k_{t_i}$ is the sum of daily values collected between $X^k_{t_{i-1}}$ and $X^k_{t_i}$. We utilize the following methods to extract temporal information:

Summary Statistics

Summary statistics such as mean and standard deviation describe the temporal sequence. In our data, we apply summary statistics (including minimum, first quartile, median, mean, third quartile, maximum, and standard deviation) to each temporal sequence τ^k and use them as inputs to our classification algorithm.

Discrete Fourier Transformation

To further extract the information from the temporal sequence, we apply a Discrete Fourier Transformation (DFT), which represents a time series by its complex-valued spectral distribution [163]. The DFT coefficients F_h of a time series $\tau = \{x_0, x_1, \ldots, x_{n-1}\}$ are complex numbers given by

$$F_h = \sum_{m=0}^{n-1} x_i e^{i2\pi hm/n} \tag{5.1}$$

where $h = 0, 1, \ldots, n-1$.

Table 5.1: Construction of daily sequences.

Sequence	**Description**	**Additional Description**
1	Same as that of Sequence 3	
2	Same as that of Sequence 3	$X_{t_i}^2 = 2$ if the claim is diabetic-related
3	All medical claims are considered; $X_{t_i}^k = 1$ $(k = 1, 2, 3)$ if the person has claim(s) on day t_i	$X_{t_i}^3 = 2$ if the claim is for diabetes, $X_{t_i}^3 = 3$ if the claim is for diabetic complications of DCSI level 1, $X_{t_i}^3 = 4$ if for complications of DCSI level 2
4	Same as that of Sequence 5	
5	Diabetic-related claims only; $X_{t_i}^k = 1$ $(k = 4, 5)$ if the person has diabetic claim(s) on day t_i	$X_{t_i}^5 = 2$ if the claim is for diabetic complications of DCSI level 1, $X_{t_i}^5 = 3$ if the claim is for diabetic complications of DCSI level 2
6	Same as that of Sequence 8	
7	Same as that of Sequence 8	$X_{t_i}^7 = 2$ if the inpatient claim is diabetic-related
8	Inpatient claims only; $X_{t_i}^k = 1$ $(k = 6, 7, 8)$ if the person has inpatient claim(s) on day t_i	$X_{t_i}^8 = 2$ if the inpatient claim is for diabetes, $X_{t_i}^8 = 3$ if the inpatient claim is for complications of DCSI level 1, $X_{t_i}^8 = 4$ if the inpatient claim is for complications of DCSI level 2
9	Same as that of Sequence 10	
10	Inpatient diabetic-related claims only; $X_{t_i}^k = 1$ if it exists	$X_{t_i}^1 0 = 2$ if the inpatient claim is for complications of DCSI level 1, $X_{t_i}^1 0 = 3$ if the inpatient claim is for complications of DCSI level 2
11	Only diabetic-related claims are considered	$X_{t_i}^k$ = claim category (ranging from 0 to 4)

DFT decomposes a signal into all possible frequencies that make it up, and is widely used in signal processing. As the first few DFT coefficients can preserve most information from time series, they are often used for dimensionality reduction. In addition, the transformation is invariant to horizontal shifting, thus is an ideal representation method for our problem. For instance, two patients with temporal sequences $(0, 0, 0, 1, 1, 0)$ and $(1, 1, 0, 0, 0, 0)$ will have identical frequency presentation under DFT, which is consistent with our intuition, as the two sequences should indicate similar behavior patterns. DFT allows us to measure the similarity between two individuals (i.e., the distance) by computing the difference between the real and imaginary parts of their DFT coefficients.

K-nearest Neighbor

We represent each temporal sequence by its Fourier Transformation coefficients, and apply this new feature vector to classification using the K-nearest Neighbors (KNN) algorithm for risk prediction [9]. Specifically, for some predetermined K (which can be selected via cross validation), we find the K closest (defined by Euclidean distance) DFT samples around each DFT sample f, and record the number of positive outcomes among the K samples. The predicted risk score of a new sample x (with τ as its time series and f as its corresponding DFT) is, thus,

$$\text{Predicted Risk of } x = \frac{P^f}{K} \tag{5.2}$$

where P^f stands for the number of positive outcomes among the K neighbors for f, and the predicted risk corresponds to the percentage of nearby positive cases. It is worth noting that in this approach, the prediction is based on time series τ only,

therefore, it directly reflects the information in our temporal sequences.

In Table 5.2, we summarize the models proposed in this section and their utilization in our study. Since the goal of this research is to find effective methods for generating temporal features and assessing the value of these methods, we utilize least absolute shrinkage and selection operator (LASSO) regression [231] for training and feature selection for all classification models (except KNN) in order to give a fair comparison between different methods for capturing the temporal information. In order to find the most descriptive sequence and quantify the contribution of individual features, we first train each type of model separately, and then combine features from different models to include more information. For instance, when combining the baseline and the KNN model, the inpatient risk produced by the KNN model is fed into the final prediction model as a single feature.

5.4 Prediction Results

In Table 5.3, we show the prediction results from various individual models with respect to risk prediction on various time horizons. The area under the receiver operating characteristic curve (AUC) is used to evaluate the performance of the different models. We use the temporal sequence that has the best predictive power during sequential learning (i.e., Sequence 2) for the KNN algorithm.

Table 5.2: Summary of the training model and features.

Model	**Description/ Feature Used**
Baseline	Registration features + Yearly aggregated non-temporal features, as described in Section 5.3.2 a) and b). There are 304 features.
Model based on yearly aggregated temporal features only	Yearly aggregated temporal features only, as described in Section 5.3.2 c). There are 13 features.
Models based on sequential temporal features (summary statistics + DFT)	Sequential-based temporal features only, as described in Section 5.3.2 d). Sequences are constructed according to methods in Section 5.3.2.1. There are 499 features, including: a. Daily sequence (366 features) + summary statistics (7 features) + top DFT sequences (20 features) b. Weekly sequence (52 features) + summary statistics (7 features) + top DFT sequences (10 features) c. Monthly sequence (12 features) + summary statistics (7 features) + top DFT sequences (5 features) d. Quarterly sequence (aggregated by quarter) (4 features) + summary statistics (7 features) + top DFT sequences (2 features)
KNN based on sequential temporal features only	DFT features only. DFT is constructed on temporal sequences of different aggregation levels: daily, weekly, monthly, and quarterly. KNN is used for training and risk prediction.

It can be seen that the baseline model with yearly aggregated features and zero temporal information outperforms the rest of the models that utilize temporal features only. This is not surprising, since the features of the baseline model characterize patients from various dimensions. We also observe that temporal features indeed provide evidence about patients' future inpatient risk, especially when appropriately summarized (as demonstrated by summary statistics and DFT models based on Sequence 2 and the KNN model based on weekly aggregated data). Among all the sequential temporal features that utilize summary statistics and DFT, Sequence 2—which describes patients' daily medical usage simply as "diabetic-related", "non-diabetic related", and "no claim"—has the most predictive power. The KNN model built upon Sequence 2 achieves a satisfactory AUC (considering that only the one-dimensional temporal information is used). Out of all aggregation levels, the weekly level aggregation performs best, followed by monthly and daily, and then quarterly. This is because weekly aggregation not only smooths noise from the daily specification, but also preserves enough variation for pattern discrimination. In terms of the prediction horizon, shorter and closer time horizons present better prediction outcomes, indicating that recent events are more predictive than less recent events (as indicated by [20]), yet patients' claims records in the previous several years are still very telling of their future health condition.

In Table 5.4, we combine the sequential models with the baseline information, and observe that temporal features do boost the performance over the baseline model model by 1-2%. These results indicate that although there is value in temporal features, the improvement they bring is not substantial (at least for this application

and our modeling techniques).

5.5 Conclusions

In this chapter, we explore the role of temporal features in predicting future health outcomes among diabetic patients. By examining various methods for feature aggregation and abstraction, we show the relatively strong predictive power of the temporal sequences, but relatively small improvement in inpatient predictions when combined with other non-temporal features. Our results indicate that yearly aggregated features can provide some information in predicting the risk of patient hospitalization, but granular temporal features improve the prediction by incorporating more detailed, dynamic information. In particular, we learned that aggregation at the weekly level is ideal for our claims data analysis, as it preserves a substantial amount of granular information while reducing dimensionality. We also show that traditional summary statistics cannot fully capture detailed temporal information; therefore, advanced data mining models can provide some additional benefit. In addition, more recent temporal information is more predictive than less recent information.

For future research directions, one can extend our framework to multiple years and construct longer sequences that contain more historical information. In our study, we build the training sequence based on a single year of patients claims records. Sequences of shorter length (such as quarter-level aggregated sequences), therefore, may not exhibit the full power of aggregation due to a lack of training

data. In addition, it would be useful to assign heavier weights to more recent temporal segments during the training stage [20]. Furthermore, as different modeling techniques may influence the prediction outcome, one can explore different methods of time series representation and abstraction. Besides DFT, discrete wavelet transformation is also commonly used for time series processing. Also, variations of the KNN algorithm (e.g., assigning a weight to each sample based on its distance from the unknown sample) may provide some benefit.

Table 5.3: AUC for individual models (Numbers in bold highlight the Baseline and the best-performed sequences in each model category)

Model	**Prediction Period AUC**			
	1 quarter	1 year	2 year	3 year
Baseline	**0.769**	**0.750**	**0.754**	**0.753**
Model based on yearly aggregated temporal features only	0.426	0.454	0.479	0.498
Models based on sequential temporal features (Summary statistics + DFT)				
Sequence 1	0.396	0.438	0.454	0.466
Sequence 2	**0.607**	**0.591**	**0.580**	**0.583**
Sequence 3	0.526	0.545	0.541	0.545
Sequence 4	0.605	0.584	0.564	0.560
Sequence 5	0.254	0.375	0.384	0.397
Sequence 6	0.371	0.409	0.422	0.430
Sequence 7	0.369	0.409	0.424	0.434
Sequence 8	0.368	0.412	0.423	0.432
Sequence 9	0.402	0.431	0.445	0.453
Sequence 10	0.399	0.430	0.444	0.450
Sequence 11	0.301	0.351	0.437	0.446
KNN based on sequential temporal features only (DFT implemented on Sequence 2)				
KNN (based on Sequence 2-daily)	0.661	0.664	0.647	0.659
KNN (based on Sequence 2-weekly)	**0.682**	**0.677**	**0.661**	**0.661**
KNN (based on Sequence 2-monthly)	0.670	0.663	0.646	0.641
KNN (based on Sequence 2-quarterly)	0.627	0.624	0.613	0.606

Table 5.4: AUC of combined models (Numbers in bold highlight the best-performed model)

Model	**Prediction Period AUC**			
	1 quarter	1 year	2 year	3 year
Baseline	0.769	0.750	0.754	0.753
Baseline + yearly aggregated temporal features	0.770	0.750	0.755	0.754
Baseline + Sequence 2	0.779	0.758	0.757	0.756
Baseline + KNN (based on Sequence 2-weekly)	0.781	0.765	0.763	0.764
Baseline + yearly aggregated temporal features + KNN (based on Sequence 2-weekly)	**0.782**	**0.767**	**0.764**	**0.764**

5.A Appendix - Claim Category

We group the 25 claim types into the following five major claim categories:

- Claim Type 0: Diagnostics (lab, tests, and image examination)
- Claim Type 1: Evaluation and prevention (including primary care)
- Claim Type 2: Treatment and therapy
- Claim Type 3: ED/urgent care and observation care
- Claim Type 4: Inpatient hospitalization

Table 5.5: Claim Category for Various Claim Types.

Type	**Sub-type**	**Category**	**Description**
DXTEST	(blank)	0	SPSD* diagnostic / testing services, which are not part of other encounters. For example, electrocardiograms, allergy tests, and audiology tests are included in this encounter type.
IMAGING	(blank)	0	SPSD Radiology/imaging services.
LAB_PATH	(blank)	0	Laboratory and pathology services that are not included in any of the above encounter categories (ACS1PD*).

continued on next page

continued from previous page

Type	Sub-type	Category	Description
EVALMGMT	MEDICAL	1	
EVALMGMT	PREVENT	1	Preventative visits, for example, physician visits for screening, checkups, or other preventative services. SPSD, same-provider services are grouped into an encounter.
EVALMGMT	PSYCH	1	Psychiatric evaluation and management services. SPSD, same-provider services are grouped into an encounter.
EVALMGMT	OPT	1	Optometry evaluation and management services, for example, eye exams. SPSD, same-provider services are grouped into an encounter.
EVALMGMT	OTHER	1	Other evaluation and management visits. SPSD, same-provider services are grouped into an encounter.
OFC_SVC	(blank)	1	Office-based services that do not meet the definitions of other encounters, for example, dermatology procedures and vaccine administration (ACS1PD).

continued on next page

continued from previous page

Type	**Sub-type**	**Category**	**Description**
CHIRO	(blank)	2	Chiropractic or osteopathic manipulation visits (ACS1PD).
REHAB_TH	CARDIAC	2	Cardiac rehabilitation services (ACS1PD).
REHAB_TH	RESP_PUL	2	Respiratory and/or pulmonary rehabilitation services (ACS1PD).
REHAB_TH	HEAR_SPC	2	Hearing and/or speech rehabilitation and therapy (ACS1PD).
REHAB_TH	OCCUPAT	2	Occupational therapy (ACS1PD).
REHAB_TH	PHYSICAL	2	Physical therapy (ACS1PD).
AMB_SURG	(blank)	3	Ambulatory surgeries and procedures (all SPSD services).
EDTR	(blank)	3	Emergency department treat-and-release visits. These are visits to a hospital emergency room that do not result in an inpatient admission or observation care visit (short-term admit). SPSD services not included in the above encounters are summarized.

continued on next page

continued from previous page

Type	**Sub-type**	**Category**	**Description**
OBS	(blank)	3	Observation care visits. Observation care visits are short-term hospital visits, which are considered outpatient services by most payers. Same-patient services that occur within a 3-day time frame are grouped into encounters.
TRANSPRT	(blank)	3	Medical transport (ambulance) services. SPSD, medical transport type services are grouped into one encounter.
URGCARE	(blank)	3	Urgent care visits, including all SPSD services.
INPT	REHAB	4	Inpatient rehabilitation visit. These visits typically occur in skilled nursing facilities or rehab units within a hospital. All medical and inpatient/outpatient services for a given patient that occur during the time frame of the inpatient stay, and which are not included in the above encounters, are grouped into the encounter.

continued on next page

continued from previous page

Type	**Sub-type**	**Category**	**Description**
INPT	PSYCH-ED	4	Psychiatric hospitalizations, including substance abuse hospitalizations, which begin with an emergency room visit. All medical and inpatient/outpatient services for a given patient that occur during the time frame of the inpatient stay, and which are not included in the above encounters, are grouped into the encounter.
INPT	ACUTE-ED	4	Acute hospital admissions that occur via the emergency room. All medical and inpatient/outpatient services for a given patient that occur during the time frame of the inpatient stay, and which are not included in the above encounters, are grouped into the encounter.

continued on next page

continued from previous page

Type	**Sub-type**	**Category**	**Description**
INPT	PSYCH	4	Psychiatric hospitalizations, including substance abuse hospitalizations, which do not occur via the emergency room. All medical and inpatient/outpatient services for a given patient that occur during the time frame of the inpatient stay, and which are not included in the above encounters, are grouped into the encounter.
INPT	ACUTE	4	Acute hospital admissions that do not occur via emergency room visits. Surgeries, radiology visits, etc., will be included in these encounters. All medical and inpatient/outpatient services for a given patient that occur during the time frame of the inpatient stay, and which are not included in the above encounters, are grouped into the encounter.

**SPSD: same-patient, same-day.*

**ACS1PD: All category services are for one patient/day.*

Chapter 6: Conclusions

In this dissertation, we examine a few mathematical and computational models to mitigate overcrowding in the healthcare system. As discussed in Section 2.1, the healthcare system differs from other general systems in its mission and complexity. Therefore, it is important to adapt existing models and adjust them to deliver actionable insights for healthcare practices.

Our research contributes to the current literature in healthcare operations in two ways. Firstly, we compare simulation models with analytical models, and identify their advantages and shortcomings in healthcare modeling. We observe that in comparison with detailed simulation models, analytical models (e.g., queueing models in Chapter 2, and differential equation models in Chapter 3) can usually capture the essence of system dynamics with minimal data requirements, and are less sensitive to parametric changes. However, queueing models may miss important details of the actual system due to their simplicity. For instance, queueing models are more likely to underestimate system delays and congestion, and smaller systems with tight resource constraints tend to suffer the most from such model limitations. In addition, we observe that analytical models are limited because they are sometimes unavailable (e.g., the case for bioterrorism transmission in Chapter

3), or too costly to solve (e.g., when closed-form solutions are not available and numerical methods have to be utilized). In comparison, simulation models tend to incorporate more detailed behavior and capture the sensitive variations in smaller systems, yet the results from them are often less generalizable. Due to the reasons above, it is advisable to use analytical models (especially those with closed-form solutions) for initial estimation to obtain generalizable insights for later more detailed research. Then, one can resort to detailed simulation models or hybrid models for more environment-specific solutions.

Secondly, we explore the effective usage of claims data for risk prediction. As healthcare resources are limited, it is desirable to identify individuals who are most likely to consume large amount of resources and implement corresponding preventative strategies (e.g., case management). In Chapter 4, we propose a comprehensive framework that selects enrollees for CM programs and evaluate the effectiveness of several selection strategies. By introducing the notion of jumpers and repeaters within the context of ED usage, we demonstrate the importance of enrolling current low-cost members. We show that our proposed-framework can improve the benefit of CM programs to potential members and increase the associated savings. Our results indicate that, as the number of selected enrollees increases, a larger proportion of potential jumpers should be included in order to maximize the savings. Also, it is important for the CM programs to estimate their efficacy, since under a fixed efficacy level, the savings or losses usually increase as the program size increases. In Chapter 5, we examine various methods for incorporating temporal information from claims data for risk prediction. We demonstrate that yearly aggregated fea-

tures can provide some information in predicting risk of patient hospitalization, but granular temporal features improve the prediction by capture more detailed, dynamic information. Our results indicate that weekly aggregation is ideal for claims data analysis, as it preserves most of the granular information while reducing the high dimensionality of the data.

Moving forward, with the emergence of big data, we expect to see more hybrid models for real-time decision-making and operational planning. This is particularly useful in healthcare units where overcrowding is prevalent. Hybrid models, which can combine aspects of simulation, optimization, machine learning, and QT, can be applied for quick diagnosis (or triage) and optimized patient-flow [131]. We also expect that researchers will devote more effort to the effective integration of different data sources (such as mobile and personal tracking devices, social media, electronic medical records). New algorithms need to be developed for automatic data cleaning and information extraction from these different platforms for better health prediction, decision-making, and management. Big data alone does not provide sufficient information to inform patient-centered care and improve healthcare delivery. Modeling and analytics techniques are still needed to enhance the value provided by big data.

Bibliography

[1] Abraham, G., Byrnes, G., and Bain, C. (2009). Short-term forecasting of emergency inpatient flow. *IEEE Transactions on Information Technology in Biomedicine*, 13(3):380–388.

[2] Abrams, L. (2013). How much does it cost to go to the ER? The Atlantic. Available at http://www.theatlantic.com/health/archive/2013/02/how-much-does-it-cost-to-go-to-the-er/273599/ (accessed September, 2016).

[3] Act, A. (1996). Health insurance portability and accountability act of 1996. *Public law*, 104:191.

[4] Ajami, S. and Rajabzadeh, A. (2013). Radio Frequency Identification (RFID) technology and patient safety. *Journal of research in medical sciences: the official journal of Isfahan University of Medical Sciences*, 18(9):809.

[5] Albin, S., Barrett, J., Ito, D., and Mueller, J. (1990). A queueing network analysis of a health center. *Queueing Systems*, 7(1):51–61.

[6] Allon, G., Deo, S., and Lin, W. (2013). The impact of size and occupancy of hospital on the extent of ambulance diversion: Theory and evidence. *Operations Research*, 61(3):544–562.

[7] Almehdawe, E., Jewkes, B., and He, Q. (2013). A markovian queueing model for ambulance offload delays. *European Journal of Operational Research*, 226(3):602–614.

[8] Althaus, F., Paroz, S., Hugli, O., Ghali, W. ., Daeppen, J., Peytremann-Bridevaux, I., and Bodenmann, P. (2011). Effectiveness of interventions targeting frequent users of emergency departments: a systematic review. *Annals of emergency medicine*, 58(1):41–52.

[9] Altman, N. (1992). An introduction to kernel and nearest-neighbor nonparametric regression. *The American Statistician*, 46(3):175–185.

[10] American Diabetes Association and others (2013). Economic costs of diabetes in the US in 2012. *Diabetes care*, 36(4):1033–1046.

[11] Anderson, D. and Bjarnadóttir, M. (2016). When is an ounce of prevention worth a pound of cure? Identifying high-risk candidates for case management. *IIE Transactions on Healthcare Systems Engineering*, 6(1):22–32.

[12] Ang, E., Kwasnick, S., Bayati, M., Plambeck, E., and Aratow, M. (2015). Accurate emergency department wait time prediction. *Manufacturing & Service Operations Management*, 18(1):141–156.

[13] Argon, N. and Ziya, S. (2009). Priority assignment under imperfect information on customer type identities. *Manufacturing & Service Operations Management*, 11(4):674–693.

[14] Armony, M. (2005). Dynamic routing in large-scale service systems with heterogeneous servers. *Queueing Systems*, 51(3):287–329.

[15] Armony, M., Israelit, S., Mandelbaum, A., Marmor, Y., Tseytlin, Y., and Yom-Tov, G. (2015). On patient flow in hospitals: A data-based queueing-science perspective. An extended version (EV). Working paper. Available at http://ie.technion.ac.il/serveng/References/Patient%20flow%20main.pdf (accessed March 30, 2015). DOI: 10.1214/14SSY153.

[16] Atkinson, W., Wolfe, S., and Hamborsky, J. (2005). *Epidemiology and prevention of vaccine-preventable diseases*. Public Health Foundation, Washington DC.

[17] Au, L., Byrnes, G., Bain, C., Fackrell, M., Brand, C., Campbell, D., and Taylor, P. (2008). Predicting overflow in an emergency department. *IMA Journal of Management Mathematics*, 20(1):39–49.

[18] Au-Yeung, S., Harrison, P., and Knottenbelt, W. (2006). A queueing network model of patient flow in an accident and emergency. In *Proceedings of 2006 European Simulation and Modelling Conference, August*, pages 60–67, Bonn, Germany.

[19] Bagust, A., Place, M., and Posnett, J. (1999). Dynamics of bed use in accommodating emergency admissions: stochastic simulation model. *Bmj*, 319(7203):155–158.

[20] Batal, I., Fradkin, D., Harrison, J., Moerchen, F., and Hauskrecht, M. (2012). Mining recent temporal patterns for event detection in multivariate time series data. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 280–288. ACM.

[21] Batt, R. and Terwiesch, C. (2015). Waiting patiently: An empirical study of queue abandonment in an emergency department. *Management Science*, 61(1):39–59.

[22] Beinlich, I., Suermondt, H., Chavez, R., and Cooper, G. (1989). The alarm monitoring system: A case study with two probabilistic inference techniques for belief networks. In *AIME 89*, pages 247–256. Springer.

[23] Bellazzi, R., Sacchi, L., and Concaro, S. (2009). Methods and tools for mining multivariate temporal data in clinical and biomedical applications. In *Engineering in Medicine and Biology Society, 2009. EMBC 2009. Annual International Conference of the IEEE*, pages 5629–5632. IEEE.

[24] Belle, A., Thiagarajan, R., Soroushmehr, S., Navidi, F., Beard, D., and Najarian, K. (2015). Big data analytics in healthcare. *BioMed research international*, 2015.

[25] Bertsimas, D. and Mourtzinou, G. (1997). Transient laws of non-stationary queueing systems and their applications. *Queueing Systems*, 25(1):115–155.

[26] Billings, J. (2013). Background of NYU ED algorithm. Available at http://wagner.nyu.edu/faculty/billings/nyued-background (accessed September, 2015).

[27] Blair, E. and Lawrence, C. (1981). A queueing network approach to health care planning with an application to burn care in new york state. *Socio-economic planning sciences*, 15(5):207–216.

[28] Blanchard, J., Haywood, Y., and Scott, C. (2003). Racial and ethnic disparities in health: an emergency medicine perspective. *Academic emergency medicine*, 10(11):1289–1293.

[29] Bodenmann, P., Velonaki, V., Ruggeri, O., Hugli, O., Burnand, B., Wasserfallen, J., Moschetti, K., Iglesias, K., Baggio, S., and Daeppen, J. (2014). Case management for frequent users of the emergency department: study protocol of a randomised controlled trial. *BMC health services research*, 14(1):264.

[30] Broyles, J. and Cochran, J. (2011). A queueing-based statistical approximation of hospital emergency department boarding. In *41st International Conference on Computers & Industrial Engineering*, pages 122–127, Los Angeles, CA.

[31] Burt, C., McCaig, L., and Valverde, R. (2006). Analysis of ambulance transports and diversions among us emergency departments. *Annals of emergency medicine*, 47(4):317–326.

[32] Campello, F., Ingolfsson, A., and Shumsky, R. (2016). Queueing models of case managers. *Management Science*.

[33] Catlin, A. and Cowan, C. (2015). History of health spending in the United States, 1960-2013. Centers for Medicare and Medicaid Services. Available at https://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/NationalHealthExpendData/Downloads/HistoricalNHEPaper.pdf (accessed June 15, 2017).

[34] Center for Infectious Disease Research and Policy (2013). Tularemia: Current, comprehensive information on pathogenesis, microbiology, epidemiology, diagnosis, treatment, and prophylaxis. Available at http://www.cidrap.umn.edu/infectious-disease-topics/tularemia-_Overview_1+CIDRAP&overview&1-2 (accessed September 6th, 2013).

[35] Centers for Disease Control and Prevention (2007). Bioterrorism overview. Available at https://emergency.cdc.gov/bioterrorism/pdf/bioterrorism_overview.pdf (accessed March 30, 2015).

[36] Centers for Disease Control and Prevention (CDC) (2015). Health expenditures. Centers for Medicare and Medicaid Services. Available at https://www.cdc.gov/nchs/fastats/health-expenditures.htm (accessed June 15, 2017).

[37] Chan, C. (2008). Intelligent value-based customer segmentation method for campaign management: A case study of automobile retailer. *Expert systems with applications*, 34(4):2754–2762.

[38] Chou, D. (2011). Electronic medical records: A practical guide for primary care. *JAMA*, 305(17):1810–1813.

[39] Cichosz, S., Johansen, M., and Hejlesen, O. (2016). Toward big data analytics: review of predictive models in management of diabetes and its complications. *Journal of diabetes science and technology*, 10(1):27–34.

[40] Clayton, D. and Hills, M. (2013). *Statistical models in epidemiology*. OUP Oxford.

[41] Cochran, J. and Bharti, A. (2006). A multi-stage stochastic methodology for whole hospital bed planning under peak loading. *International Journal of Industrial and Systems Engineering*, 1(1-2):8–36.

[42] Cochran, J. and Broyles, J. (2010). Developing nonlinear queuing regressions to increase emergency department patient safety: Approximating reneging with balking. *Computers & Industrial Engineering*, 59(3):378–386.

[43] Cochran, J. and Roche, K. (2009). A multi-class queuing network analysis methodology for improving hospital emergency department performance. *Computers & Operations Research*, 36(5):1497–1512.

[44] Committee on the Future of Emergency Care in the United States Health Care System/Board on Health Care Services (2006). Hospital-based emergency care: at the breaking point.

[45] Concaro, S., Sacchi, L., Cerra, C., Fratino, P., and Bellazzi, R. (2011). Mining health care administrative data with temporal association rules on hybrid events. *Methods of information in medicine*, 50(2):166–179.

[46] Cooke, M., Wilson, S., and Pearson, S. (2002). The effect of a separate stream for minor injuries on accident and emergency department waiting times. *Emergency Medicine Journal*, 19(1):28–30.

[47] Crane, S., Collins, L., Hall, J., Rochester, D., and Patch, S. (2012). Reducing utilization by uninsured frequent users of the emergency department: combining case management and drop-in group medical appointments. *The Journal of the American Board of Family Medicine*, 25(2):184–191.

[48] De Bruin, A., Koole, G., and Visser, M. (2005). Bottleneck analysis of emergency cardiac in-patient flow in a university setting: an application of queueing theory. *Clinical and investigative medicine*, 28(6):316.

[49] De Bruin, A., Van Rossum, A., Visser, M., and Koole, G. (2007). Modeling the emergency cardiac in-patient flow: an application of queuing theory. *Health Care Management Science*, 10(2):125–137.

[50] de Vericourt, F. and Jennings, O. (2008). Nurse-to-patient ratios in hospital staffing: a queueing perspective. ESMT Working Paper No. 08-005, Duke University, Durham, NC.

[51] Defraeye, M. and Van Nieuwenhuyse, I. (2011). Setting staffing levels in an emergency department: opportunities and limitations of stationary queuing models. *Review of Business and Economics*, 56(1):73–100.

[52] Deo, S. and Gurvich, I. (2011). Centralized vs. decentralized ambulance diversion: A network perspective. *Management Science*, 57(7):1300–1319.

[53] Derlet, R. and Richards, J. (2000). Overcrowding in the nation's emergency departments: complex causes and disturbing effects. *Annals of emergency medicine*, 35(1):63–68.

[54] Derlet, R. and Richards, J. (2002). Emergency department overcrowding in florida, new york, and texas. *Southern medical journal*, 95(8):846–850.

[55] Ding, R., McCarthy, M., Desmond, J., Lee, J., Aronsky, D., and Zeger, S. (2010). Characterizing waiting room time, treatment time, and boarding time in the emergency department using quantile regression. *Academic Emergency Medicine*, 17(8):813–823.

[56] Dong, G. and Taslimitehrani, V. (2016). Pattern aided classification. In *Proceedings of the 2016 SIAM International Conference on Data Mining*, pages 225–233. SIAM.

[57] Dong, M. and He, D. (2007). A segmental hidden semi-markov model (HSMM)-based diagnostics and prognostics framework and methodology. *Mechanical systems and signal processing*, 21(5):2248–2266.

[58] Doran, K., Raven, M., and Rosenheck, R. (2013). What drives frequent emergency department use in an integrated health system? national data from the veterans health administration. *Annals of Emergency Medicine*, 62(2):151–159.

[59] Doupe, M., Palatnick, W., Day, S., Chateau, D., Soodeen, R., Burchill, C., and Derksen, S. (2012). Frequent users of emergency departments: developing standard definitions and defining prominent risk factors. *Annals of emergency medicine*, 60(1):24–32.

[60] Dove, H., Duncan, I., and Robb, A. (2003). A prediction model for targeting low-cost, high-risk members of managed care organizations. *Am J Manag Care*, 9(5):381–9.

[61] Eick, S., Massey, W., and Whitt, W. (1993). Mt/g/∞ queues with sinusoidal arrival rates. *Management Science*, 39(2):241–252.

[62] Enders, P. (2010). *Applications of stochastic and queueing models to operational decision making*. PhD thesis, Carnegie Mellon University.

[63] Federation of American Scientists (2014). Biological threat agents information. Available at http://fas.org/programs/ssp/bio/resource/biothreatagents.html (accessed April 4th, 2014).

[64] Feldman, Z., Mandelbaum, A., Massey, W., and Whitt, W. (2008). Staffing of time-varying queues to achieve time-stable performance. *Management Science*, 54(2):324–338.

[65] Fiems, D., Koole, G., and Nain, P. (2007). Waiting times of scheduled patients in the presence of emergency requests. Working paper, VU University Amsterdam, Amsterdam. Available at http://www.math.vu.nl/~koole/publications/2005report1/art.pdf (accessed January 05, 2017).

[66] Finkelstein, A., Taubman, S., Allen, H., Wright, B., and Baicker, K. (2016). Effect of medicaid coverage on ed use–further evidence from oregon's experiment. *New England Journal of Medicine*, 375(16):1505–1507.

[67] Fomundam, S. and Herrmann, J. (2007). A survey of queuing theory applications in healthcare. Technical Report 24, Institute for Systems Research, University of Maryland, College Park, MD.

[68] Forbes, J. and Cooper, M. (2013). Mechanisms of diabetic complications. *Physiological reviews*, 93(1):137–188.

[69] Forsberg, H., Athlin, Å., and von Thiele Schwarz, U. (2015). Nurses' perceptions of multitasking in the emergency department: Effective, fun and unproblematic (at least for me)–a qualitative study. *International emergency nursing*, 23(2):59–64.

[70] Freshwater, E. and Crouch, R. (2015). Technology for trauma: testing the validity of a smartphone app for pre-hospital clinicians. *International emergency nursing*, 23(1):32–37.

[71] Fuda, K. and Immekus, R. (2006). Frequent users of Massachusetts emergency departments: a statewide analysis. *Annals of emergency medicine*, 48(1):16–e1.

[72] Gardner, B., Lally, P., and Wardle, J. (2012). Making health habitual: the psychology of 'habit-formation' and general practice. *Br J Gen Pract*, 62(605):664–666.

[73] Garnett, O., Mandelbaum, A., and Reiman, M. (2002). Designing a call center with impatient customers. *Manufacturing & Service Operations Management*, 4(3):208–227.

[74] Garrett, C. (2008). The effect of nurse staffing patterns on medical errors and nurse burnout. *AORN journal*, 87(6):1191–1204.

[75] Gautam, N. (2012). *Analysis of queues: methods and applications*. CRC Press, Boca Raton, FL.

[76] Gilboy, N., Tanabe, T., Travers, D., and Rosenau, A. (2011). *Emergency Severity Index (ESI): A Triage Tool for Emergency Department Care, version 4, implementation handbook*. AHRQ Publication, Rockville, MD.

[77] Gordon, J., Billings, J., Asplin, B., and Rhodes, K. (2001). Safety net research in emergency medicine proceedings of the academic emergency medicine consensus conference on "the unraveling safety net". *Academic Emergency Medicine*, 8(11):1024–1029.

[78] Green, L. (2002). How many hospital beds? *INQUIRY: The Journal of Health Care Organization, Provision, and Financing*, 39(4):400–412.

[79] Green, L. (2006). Queueing analysis in healthcare. In Hall, R., editor, *Patient Flow: Reducing Delay in Health Care Delivery, International Series in Operations Research and Management Science*, volume 91, pages 281–307. Springer, New York, 1 edition.

[80] Green, L. and Kolesar, P. (1995). On the accuracy of the simple peak hour approximation for markovian queues. *Management science*, 41(8):1353–1370.

[81] Green, L. and Kolesar, P. (1997). The lagged PSA for estimating peak congestion in multiserver markovian queues with periodic arrival rates. *Management Science*, 43(1):80–87.

[82] Green, L., Kolesar, P., and Soares, J. (2001). Improving the SIPP approach for staffing service systems that have cyclic demands. *Operations Research*, 49(4):549–564.

[83] Green, L., Kolesar, P., and Whitt, W. (2007). Coping with time-varying demand when setting staffing requirements for a service system. *Production and Operations Management*, 16(1):13–39.

[84] Green, L., Soares, J., Giglio, J., and Green, R. (2006). Using queueing theory to increase the effectiveness of emergency department provider staffing. *Academic Emergency Medicine*, 13(1):61–68.

[85] Grover, C., Close, R., Villarreal, K., and Goldman, L. (2010). Emergency department frequent user: pilot study of intensive case management to reduce visits and computed tomography. *Western Journal of Emergency Medicine*, 11(4).

[86] Guillemin, J. (1999). *Anthrax: the investigation of a deadly outbreak*. University of California Press, Oakland, California.

[87] Gupta, D. (2013). Queueing models for healthcare operations. In Denton, B., editor, *Handbook of Healthcare Operations Management: Methods and Applications*, volume 91, pages 19–44. Springer-Verlag, New York.

[88] Gupta, D. and Denton, B. (2008). Appointment scheduling in health care: Challenges and opportunities. *IIE transactions*, 40(9):800–819.

[89] Hagtvedt, R., Ferguson, M., Griffin, P., Jones, G., and Keskinocak, P. (2009). Cooperative strategies to reduce ambulance diversion. In Rossetti, M., Hill, R., Johansson, B.and Dunkin, A., and Ingalls, R., editors, *Proceedings of the 2009 Winter Simulation Conference*, pages 1861–1874, NJ. IEEE Press Piscataway.

[90] Halfin, S. and Whitt, W. (1981). Heavy-traffic limits for queues with many exponential servers. *Operations research*, 29(3):567–588.

[91] Hall, R., Belson, D.and Murali, P., and Dessouky, M. Modeling patient flows through the health care system. In Hall, R., editor, *Patient Flow: Reducing Delay in Healthcare Delivery*, pages 1–44.

[92] Haussmann, R. (1970). Waiting time as an index of quality of nursing care. *Health services research*, 5(2):92.

[93] Healthcare Cost and Utilization Project (2014). Clinical classifications software (CCS) for ICD-9-CM. Agency for Healthcare Research and Quality, Rockville, MD. Available at https://www.hcup-us.ahrq.gov/toolssoftware/ccs/ccs.jsp (accessed September, 2015).

[94] Healthcare Cost and Utilization Project (2015). Chronic Condition Indicator (CCI) for ICD-9-CM. Agency for Healthcare Research and Quality, Rockville, MD. Available at https://www.hcup-us.ahrq.gov/toolssoftware/chronic/chronic.jsp (accessed September, 2015).

[95] Himmelstein, D. and Woolhandler, S. (2016). The current and projected taxpayer shares of US health costs. *American journal of public health*, 106(3):449–452.

[96] Holty, J., Bravata, D., Liu, H., Olshen, R., McDonald, K., and Owens, D. (2006). Systematic review: a century of inhalational anthrax cases from 1900 to 2005. *Annals of internal medicine*, 144(4):270–280.

[97] Hong, R., Baumann, B., and Boudreaux, E. (2007). The emergency department for routine healthcare: race/ethnicity, socioeconomic status, and perceptual factors. *The Journal of emergency medicine*, 32(2):149–158.

[98] Höppner, F. and Klawonn, F. (2002). Finding informative rules in interval sequences. *Intelligent Data Analysis*, 6(3):237–255.

[99] Hsia, R. and Tabas, J. (2009). Emergency care: the increasing weight of increasing waits. *Archives of internal medicine*, 169(20):1836–1838.

[100] Hsieh, J., Li, A., and Yang, C. (2013). Mobile, cloud, and big data computing: contributions, challenges, and new directions in telecardiology. *International journal of environmental research and public health*, 10(11):6131–6153.

[101] Hu, X., Barnes, S., Bjarnadottir, M., and Golden, B. (2017a). Intelligent selection of frequent emergency department patients for case management: A machine learning framework based on claims data. *IIE Transactions on Healthcare Systems Engineering*, to appear.

[102] Hu, X., Barnes, S., and Golden, B. (2014). Early detection of bioterrorism: monitoring disease using an agent-based model. In Tolk, A., Diallo, S., Ryzhov, I., Yilmaz, L.and Buckley, S., and Miller, J., editors, *Proceedings of the 2014 Winter Simulation Conference*, pages 310–321, NJ. IEEE Press Piscataway.

[103] Hu, X., Barnes, S., and Golden, B. (2017b). Applying queueing theory to the study of emergency department operations: A survey and a discussion of comparable simulation studies. *International Transactions in Operational Research*, March 7.

[104] Huang, Q.and Thind, A., Dreyer, J., and Zaric, G. (2010). The impact of delays to admission from the emergency department on inpatient outcomes. *BMC emergency medicine*, 10(1):16.

[105] Huang, J., Carmeli, B., and Mandelbaum, A. (2015). Control of patient flow in emergency departments, or multiclass queues with deadlines and feedback. *Operations Research*, 63(4):892–908.

[106] Huang, X. (1995). A planning model for requirement of emergency beds. *Mathematical Medicine and Biology*, 12(3-4):345–353.

[107] Hunt, K., Weber, E., Showstack, J., Colby, D., and Callaham, M. (2006). Characteristics of frequent users of emergency departments. *Annals of emergency medicine*, 48(1):1–8.

[108] Izady, N. and Worthington, D. (2012). Setting staffing requirements for time dependent queueing networks: The case of accident and emergency departments. *European Journal of Operational Research*, 219(3):531–540.

[109] Jennings, O., Mandelbaum, A., Massey, W., and Whitt, W. (1996). Server staffing to meet time-varying demand. *Management Science*, 42(10):1383–1394.

[110] Johns Hopkins Bloomberg School of Public Health (2015). Primary care visits available to most uninsured but at a high price. Available at http://www.jhsph.edu/news/news-releases/2015/primary-care-visits-available-to-most-uninsured-but-at-a-high-price.html (accessed September, 2016).

[111] Jun, J., Jacobson, S., and Swisher, J. (1999). Application of discrete-event simulation in health care clinics: A survey. *Journal of the operational research society*, pages 109–123.

[112] Kao, E. and Tung, G. (1981). Bed allocation in a public health care delivery system. *Management Science*, 27(5):507–520.

[113] Kazahaya, G. (2005). Harnessing technology to redesign labor cost management reports: labor costs typically represent over 50 percent of a hospital's total operating expenses. can the data management process be harnessed to create meaningful labor cost management tools? *Healthcare Financial Management*, 59(4):94–101.

[114] Kc, D. and Terwiesch, C. (2009). Impact of workload on service time and patient safety: An econometric analysis of hospital operations. *Management Science*, 55(9):1486–1498.

[115] Keehan, S., Sisko, A., Truffer, C., Smith, S., Cowan, C., Poisal, J., and Clemens, M. (2008). Health spending projections through 2017: the baby-boom generation is coming to medicare. *Health Affairs*, 27(2):w145–w155.

[116] Kendall, D. (1953). Stochastic processes occurring in the theory of queues and their analysis by the method of the imbedded markov chain. *The Annals of Mathematical Statistics*, pages 338–354.

[117] Kim, S. and Whitt, W. (2014). Are call center and hospital arrivals well modeled by nonhomogeneous poisson processes? *Manufacturing & Service Operations Management*, 16(3):464–480.

[118] Kolb, E., Peck, J., Schoening, S., and Lee, T. (2008). Reducing emergency department overcrowding-five patient buffer concepts in comparison. In Mason, S., Hill, R., Mönch, L., Rose, O., Jefferson, T., and Fowler, J., editors, *Proceedings of the 2008 Winter Simulation Conference*, pages 1516–1525, NJ. IEEE, Piscataway.

[119] Komaroff, A. (2015). Medicaid expansion in Oregon led to more short-term ED use. Medicaid expansion in Oregon led to more short-term ED use. Available at http://www.jwatch.org/na33444/2014/01/28/medicaid-expansion-oregon-led-more-short-term-ed-use (accessed September, 2015).

[120] Komashie, A., Mousavi, A., Clarkson, P., and Young, T. (2015). An integrated model of patient and staff satisfaction using queuing theory. *IEEE journal of translational engineering in health and medicine*, 3:1–10.

[121] Konrad, R., DeSotto, K., Grocela, A., McAuley, P., Wang, J., Lyons, J., and Bruin, M. (2013). Modeling the impact of changing patient flow processes in an emergency department: Insights from a computer simulation study. *Operations Research for Health Care*, 2(4):66–74.

[122] Koole, G. and Mandelbaum, A. (2002). Queueing models of call centers: An introduction. *Annals of Operations Research*, 113(1):41–59.

[123] Kumar, G. and Klein, R. (2013). Effectiveness of case management strategies in reducing emergency department visits in frequent user patient populations: a systematic review. *The Journal of emergency medicine*, 44(3):717–729.

[124] Kuo, Y., Leung, J., and Graham, C. (2012). Simulation with data scarcity: developing a simulation model of a hospital emergency department. In Laroque, C., Himmelspach, J., Pasupathy, R., Rose, O., and Uhrmacher, A., editors, *Proceedings of the Winter Simulation Conference*, page 87, NJ. Proceedings of the 2012 Winter Simulation Conference, Piscataway.

[125] LaCalle, E. and Rabin, E. (2010). Frequent users of emergency departments: the myths, the data, and the policy implications. *Annals of emergency medicine*, 56(1):42–48.

[126] Lakshmi, C. and Iyer, S. (2013). Application of queueing theory in health care: A literature review. *Operations research for health care*, 2(1):25–39.

[127] Lally, P. and Gardner, B. (2013). Promoting habit formation. *Health Psychology Review*, 7(sup1):S137–S158.

[128] Lambert, D. (1992). Zero-inflated poisson regression, with an application to defects in manufacturing. *Technometrics*, 34(1):1–14.

[129] Law, H., Oraka, E., and Mannino, D. (2011). The role of income in reducing racial and ethnic disparities in emergency room and urgent care center visits for asthma–United States, 2001–2009. *Journal of Asthma*, 48(4):405–413.

[130] Lazer, D., Kennedy, R., King, G., and Vespignani, A. (2014). The parable of Google Flu: traps in big data analysis. *Science*, 343(6176):1203–1205.

[131] Lee, E., Atallah, H., Wright, M., Post, E., C., T., Wu, D., and L.L., H. (2015). Transforming hospital emergency department workflow and patient care. *Interfaces*.

[132] Lee, K. and Davenport, L. (2006). Can case management interventions reduce the number of emergency department visits by frequent users? *The health care manager*, 25(2):155–159.

[133] Leemis, L. (1991). Nonparametric estimation of the cumulative intensity function for a nonhomogeneous poisson process. *Management Science*, 37(7):886–900.

[134] Lemon, S., Mahmoud, A., Mack, A., and Knobler, S. (2005). *The threat of pandemic influenza: are we ready? workshop summary*. National Academies Press, Washington D.C.

[135] Lemus, J., Chacko, M., and Claudius, I. (2013). Need for intervention in families presenting to the emergency department with multiple children as patients. *Western Journal of Emergency Medicine*, 14(5):525.

[136] Leonard, K. (2016). Could universal health care save u.s. taxpayers money? U.S. News & World Report. Available at https://www.usnews.com/news/blogs/data-mine/2016/01/22/could-universal-health-care-save-us-taxpayers-money (accessed June 15, 2017).

[137] Lin, D., Patrick, J., and Labeau, F. (2014). Estimating the waiting time of multi-priority emergency patients with downstream blocking. *Health care management science*, 17(1):88–99.

[138] Liu, C., Wang, F., Hu, J., and Xiong, H. (2015). Temporal phenotyping from longitudinal electronic health records: A graph based framework. In *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, pages 705–714. ACM.

[139] Liu, S. and Chen, J. (2009). Using data mining to segment healthcare markets from patients' preference perspectives. *International journal of health care quality assurance*, 22(2):117–134.

[140] Liu, S., Thomas, S., Gordon, J., and Weissman, J. (2005). Frequency of adverse events and errors among patients boarding in the emergency department. *Academic Emergency Medicine*, 12(5 Supplement 1):49.

[141] Lober, W., Karras, B., Wagner, M., Overhage, J., Davidson, A., Fraser, H., Trigg, L., Mandl, K., Espino, J., and Tsui, F. (2002). Roundtable on bioterrorism detection: information system–based surveillance. *Journal of the American Medical Informatics Association*, 9(2):105–115.

[142] Long, W. (1996). Temporal reasoning for diagnosis in a causal probabilistic knowledge base. *Artificial intelligence in medicine*, 8(3):193–215.

[143] Luo, W., Cao, J., Gallagher, M., and Wiles, J. (2013). Estimating the intensity of ward admission and its effect on emergency department access block. *Statistics in medicine*, 32(15):2681–2694.

[144] Lynn, J., Straube, B., Bell, K., Jencks, S., and Kambic, R. (2007). Provide better health care for all: The ?bridges to health? model. *Milbank Quarterly*, 85(2):185–208.

[145] Maa, J. (2011). The waits that matter. *New England Journal of Medicine*, 364(24):2279–2281.

[146] Macal, C. and Michael, J. N. (2007). Agent-based modeling and simulation: Desktop ABMS. In *the 2007 Winter Simulation Conference. Washington, DC, USA*.

[147] Madsen, T. and Kofoed-Enevoldsen, A. (2011). Five easy equations for patient flow through an emergency department. *Dan Med Bull*, 58(10):A4318.

[148] Mahaffy, J. and Dockery, J. (2013). Influenza modeling with a discrete sir model. Available at http://www.math.montana.edu/~umsfjdoc/m430/Influenza.pdf (accessed July 1st, 2014).

[149] Maman, S. (2009). *Uncertainty in the demand for service: The case of call centers and emergency departments*. Technion-Israel Institute of Technology, Faculty of Industrial and Management Engineering.

[150] Mandelbaum, A., Momčilović, P., and Tseytlin, Y. (2012). On fair routing from emergency departments to hospital wards: QED queues with heterogeneous servers. *Management Science*, 58(7):1273–1291.

[151] Mandelberg, J., Kuhn, R., and Kohn, M. (2000). Epidemiologic analysis of an urban, public emergency department's frequent users. *Academic emergency medicine*, 7(6):637–646.

[152] Marcilio, I., Hajat, S., and Gouveia, N. (2013). Forecasting daily emergency department visits using calendar variables and ambient temperature readings. *Academic emergency medicine*, 20(8):769–777.

[153] Marshall, D., Burgos-Liz, L., Pasupathy, K., Padula, W., IJzerman, M., Wong, P., Higashi, M., Engbers, J., Wiebe, S., Crown, W., and Osgood, N. (2016). Transforming healthcare delivery: Integrating dynamic simulation modelling and big data in health economics and outcomes research. *PharmacoEconomics*, 34(2):115–126.

[154] Massey, W., Parker, G., and Whitt, W. (1996). Estimating the parameters of a nonhomogeneous poisson process with linear rate. *Telecommunication Systems*, 5(2):361–388.

[155] Massey, W. and Whitt, W. (1994). An analysis of the modified offered-load approximation for the nonstationary erlang loss model. *The Annals of applied probability*, pages 1145–1160.

[156] Mayhew, L. and Smith, D. (2008). Using queuing theory to analyse the government's 4-h completion time target in accident and emergency departments. *Health care management science*, 11(1):11–21.

[157] McCarthy, M., Zeger, S., Ding, R., Aronsky, D., Hoot, N., and Kelen, G. (2008). The challenge of predicting demand for emergency department services. *Academic Emergency Medicine*, 15(4):337–346.

[158] McConnell, K., Richards, C., Daya, M., Weathers, C., and Lowe, R. (2006). Ambulance diversion and lost hospital revenues. *Annals of emergency medicine*, 48(6):702–710.

[159] McQuarrie, D. (1983). Hospitalization utilization levels: The application of queuing theory to a controversial medical economic problem. *Minnesota Medicine*, 66(11):679–686.

[160] Medicaid and CHIP Payment and Access Commission (2014). Revisiting emergency department use in medicaid. MAC facts: Key findings on medicaid and CHIP. Available at https://www.macpac.gov/wp-content/uploads/2015/01/MACFacts-EDuse_2014-07.pdf (accessed September, 2016).

[161] Mehrabi, S., Sohn, S., Li, D., Pankratz, J., Therneau, T., Sauver, J., Liu, H., and Palakal, M. (2015). Temporal pattern and association discovery of diagnosis codes using deep learning. In *Healthcare Informatics (ICHI), 2015 International Conference on*, pages 408–416. IEEE.

[162] Mihaylova, B., Briggs, A., O'hagan, A., and Thompson, S. (2011). Review of statistical methods for analysing healthcare resources and costs. *Health economics*, 20(8):897–916.

[163] Mitsa, T. (2010). *Temporal data mining*. CRC Press.

[164] Moll, H. (2010). Challenges in the validation of triage systems at emergency departments. *Journal of clinical epidemiology*, 63(4):384–388.

[165] Moore, G., Gerdtz, M., Manias, E., Hepworth, G., and Dent, A. (2007). Sociodemographic and clinical characteristics of re-presentation to an australian inner-city emergency department: implications for service delivery. *BMC Public Health*, 7(1):320.

[166] Mörchen, F. and Ultsch, A. (2007). Efficient mining of understandable patterns from multivariate interval time series. *Data mining and knowledge discovery*, 15(2):181.

[167] Morrison, M., Murphy, T., and Nalder, C. (2003). Consumer preferences for general practitioner services. *Health marketing quarterly*, 20(3):3–19.

[168] Mullahy, J. (1986). Specification and testing of some modified count data models. *Journal of econometrics*, 33(3):341–365.

[169] National Center for Health Statistics (1980). Estimates of selected comparability ratios on dual coding of 1976 death certificates by eighth and ninth revisions of the international classifications of diseases. *Monthly vital statistics report*, 28(11):1–19.

[170] National Drug Code Directory (2015). List of FDA approved drugs. Available at https://www.accessdata.fda.gov/scripts/cder/ndc/ (accessed September, 2015).

[171] Neufeld, E., Viau, K., Hirdes, J., and Warry, W. (2016). Predictors of frequent emergency department visits among rural older adults in ontario using the resident assessment instrument-home care. *Australian Journal of Rural Health*, 24(2):115–122.

[172] Nosek, R. and Wilson, J. (2001). Queuing theory and customer satisfaction: a review of terminology, trends, and applications to pharmacy practice. *Hospital pharmacy*, 36(3):275–279.

[173] Okin, R., Boccellari, A., Azocar, F., Shumway, M., O'brien, K., Gelb, A., Kohn, M., Harding, P., and Wachsmuth, C. (2000). The effects of clinical case management on hospital service use among ed frequent users. *The American journal of emergency medicine*, 18(5):603–608.

[174] Olshaker, J. and Rathlev, N. (2006). Emergency department overcrowding and ambulance diversion: the impact and potential solutions of extended boarding of admitted patients in the emergency department. *The Journal of emergency medicine*, 30(3):351–356.

[175] Olson, D. and Delen, D. (2008). *Advanced data mining techniques*. Springer Science & Business Media.

[176] Orr, J. (2008). The good, the bad, and the four hour target. *BMJ: British Medical Journal*, 337.

[177] Pajouh, F. and Kamath, M. (2010). Applications of queueing models in hospitals. In *Proceedings of the Fifth Midwest Association for Information Systems Conference (MWAIS 2010)*, Moorhead, MN. Paper 23.

[178] Palaniappan, S. and Awang, R. (2008). Intelligent heart disease prediction system using data mining techniques. In *Computer Systems and Applications, 2008. AICCSA 2008. IEEE/ACS International Conference on*, pages 108–115. IEEE.

[179] Palvannan, R. and Teow, K. (2012). Queueing for healthcare. *Journal of medical systems*, 36(2):541–547.

[180] Panayiotopoulos, J. and Vassilacopoulos, G. (1984). Simulating hospital emergency departments queuing systems:(GI/G/m (t)):(IHFF/N/∞). *European journal of operational research*, 18(2):250–258.

[181] Park, Y., Lee, Y., Lee, G., Lee, J., and Shin, S. (2015). Smartphone applications with sensors used in a tertiary hospital–current status and future challenges. *Sensors*, 15(5):9854–9869.

[182] Patient Protection and Affordable Care Act (2010). 42 U.S.C. §18001.

[183] Patrick, J., Puterman, M., and Queyranne, M. (2008). Dynamic multipriority patient scheduling for a diagnostic resource. *Operations research*, 56(6):1507–1525.

[184] Patterson, K. and Pyle, G. (1991). The geography and mortality of the 1918 influenza pandemic. *Bulletin of the History of Medicine*, 65(1):4–21.

[185] Paul, S., Reddy, M., and DeFlitch, C. (2010). A systematic review of simulation studies investigating emergency department overcrowding. *Simulation*, 86(8-9):559–571.

[186] Peck, J., Benneyan, J., Nightingale, D., and Gaehde, S. (2012). Predicting emergency department inpatient admissions to improve same-day patient flow. *Academic Emergency Medicine*, 19(9):E1045–E1054.

[187] Pereira, M., Singh, V., Hon, C., McKelvey, T., Sushmita, S., and De Cock, M. (2016). Predicting future frequent users of emergency departments in California state. In *Proceedings of the 1st Workshop on Methods and Applications for Healthcare Analytics (MAHA) in conjunction with ACM BCB*, volume 2016.

[188] Pfuntner, A., Wier, L., and Elixhauser, A. (2013). An overview of hospital stays in the United States, 2011: HCUP Statistical Brief# 166. Agency for Healthcare Research and Quality, Rockville, MD. Available at https://www.hcup-us.ahrq.gov/reports/statbriefs/sb166.pdf (accessed June 15, 2017).

[189] Pham, J., Patel, R., Millin, M., Kirsch, T., and Chanmugam, A. (2006). The effects of ambulance diversion: a comprehensive review. *Academic Emergency Medicine*, 13(11):1220–1227.

[190] Pike, M., Proctor, D., and Wyllie, J. (1963). Analysis of admissions to a casualty ward. *British journal of preventive & social medicine*, 17(4):172–176.

[191] Pitts, S., Niska, R., Xu, J., and Burt, C. (2008). National hospital ambulatory medical care survey: 2006 emergency department summary. *Natl Health Stat Report*, 7(7):1–38.

[192] Preater, J. (2002). Queues in health. *Health Care Management Science*, 5(4):283–283.

[193] Presidents Council of Advisors on Science and Technology (2014). Report to the president. better health care and lower costs: Accelerating improvement through systems engineering. The White House, Washington DC. Available at https://www.whitehouse.gov/sites/default/files/microsites/ostp/PCAST/pcast_systems_engineering_in_healthcare_-_may_2014.pdf (accessed September 29, 2016).

[194] Rais, A. and Viana, A. (2011). Operations research in healthcare: a survey. *International transactions in operational research*, 18(1):1–31.

[195] Ram, S., Zhang, W., Williams, M., and Pengetnze, Y. (2015). Predicting asthma-related emergency department visits using big data. *IEEE journal of biomedical and health informatics*, 19(4):1216–1223.

[196] Ramsey, S., Newton, K., Blough, D., McCulloch, D., Sandhu, N., and Wagner, E. (1999). Patient-level estimates of the cost of complications in diabetes in a managed-care population. *Pharmacoeconomics*, 16(3):285–295.

[197] Ratanamahatana, C. and Keogh, E. (2005). Three myths about dynamic time warping data mining. In *Proceedings of the 2005 SIAM International Conference on Data Mining*, pages 506–510. SIAM.

[198] Raven, M. and Gould, D. (2012). *Time and again: frequent users of emergency department services in New York City*. United Hospital Fund.

[199] Ridge, J., Jones, S., Nielsen, M., and Shahani, A. (1998). Capacity planning for intensive care units. *European journal of operational research*, 105(2):346–355.

[200] Roberts, F. (2003). Challenges for discrete mathematics and theoretical computer science in the defense against bioterrorism. *Bioterrorism: Mathematical modeling applications in homeland security*, pages 1–34.

[201] Roche, K. and Cochran, J. (2007). Improving patient safety by maximizing fast-track benefits in the emergency department: a queuing network approach. In *Proceedings of IIE Annual Conference*, pages 619–624, Memphis, TN.

[202] Rossana, R. and Seater, J. (1995). Temporal aggregation and economic time series. *Journal of Business & Economic Statistics*, 13(4):441–451.

[203] Sacchi, L., Dagliati, A., and Bellazzi, R. (2015). Analyzing complex patients' temporal histories: new frontiers in temporal data mining. *Data Mining in Clinical Medicine*, pages 89–105.

[204] Saghafian, S., Austin, G., and Traub, S. (2015). Operations research/management contributions to emergency department patient flow optimization: Review and research prospects. *IIE Transactions on Healthcare Systems Engineering*, 5(2):101–123.

[205] Saghafian, S., Hopp, W., Van Oyen, M., Desmond, J., and Kronick, S. (2012). Patient streaming as a mechanism for improving responsiveness in emergency departments. *Operations Research*, 60(5):1080–1097.

[206] Saghafian, S., Hopp, W., Van Oyen, M., Desmond, J., and Kronick, S. (2014). Complexity-augmented triage: A tool for improving patient safety and operational efficiency. *Manufacturing & Service Operations Management*, 16(3):329–345.

[207] Schneider, S., Gallery, M., Schafermeyer, R., and Zwemer, F. (2003). Emergency department crowding: a point in time. *Annals of emergency medicine*, 42(2):167–172.

[208] Schull, M., Mamdani, M., and Fang, J. (2004). Community influenza outbreaks and emergency department ambulance diversion. *Annals of emergency medicine*, 44(1):61–67.

[209] Schweigler, L., Desmond, J., McCarthy, M., Bukowski, K., Ionides, E., and Younger, J. (2009). Forecasting models of emergency department crowding. *Academic Emergency Medicine*, 16(4):301–308.

[210] Shahar, Y. (1997). A framework for knowledge-based temporal abstraction. *Artificial intelligence*, 90(1-2):79–133.

[211] Shanthikumar, J. and Sargent, R. (1983). A unifying view of hybrid simulation/analytic models and modeling. *Operations research*, 31(6):1030–1052.

[212] Sharif, A., Stanford, D., Taylor, P., and Ziedins, I. (2014). A multi-class multi-server accumulating priority queue with application to health care. *Operations Research for Health Care*, 3(2):73–79.

[213] Shmueli, A., Sprung, C., and Kaplan, E. (2003). Optimizing admissions to an intensive care unit. *Health Care Management Science*, 6(3):131–136.

[214] Shumway, M., Boccellari, A., O'brien, K., and Okin, R. (2008). Cost-effectiveness of clinical case management for ED frequent users: results of a randomized trial? *The American journal of emergency medicine*, 26(2):155–164.

[215] Siddharthan, K., Jones, W., and Johnson, J. (1996). A priority queuing model to reduce waiting times in emergency care. *International Journal of Health Care Quality Assurance*, 9(5):10–16.

[216] Silberholz, J., Anderson, D., Golden, B., Harrington, M., and Hirshon, J. (2013). The impact of the residency teaching model on the efficiency of the emergency department at an academic center. *Socio-Economic Planning Sciences*, 47(3):183–190.

[217] Silvestrini, A. and Veredas, D. (2008). Temporal aggregation of univariate and multivariate time series models: a survey. *Journal of Economic Surveys*, 22(3):458–497.

[218] Simmons, F. M. (2005). CEU: Hospital overcrowding: An opportunity for case managers. *The Case Manager*, 16(4):52–54.

[219] Solberg, L., Asplin, B., Weinick, R., and Magid, D. (2003). Emergency department crowding: consensus development of potential measures. *Annals of emergency medicine*, 42(6):824–834.

[220] Soril, L., Leggett, L., Lorenzetti, D., Noseworthy, T., and Clement, F. (2015). Reducing frequent visits to the emergency department: a systematic review of interventions. *PloS one*, 10(4):e0123660.

[221] Spillane, L., Lumb, E., Cobaugh, D., Wilcox, S., Clark, J., and Schneider, S. (1997). Frequent users of the emergency department: can we intervene? *Academic Emergency Medicine*, 4(6):574–580.

[222] Sprivulis, P., Da Silva, J., Jacobs, I., Frazer, A., and Jelinek, G. (2006). The association between hospital overcrowding and mortality among patients admitted via Western Australian emergency departments. *Medical Journal of Australia*, 184(5):208.

[223] Stanford, D., Taylor, P., and Ziedins, I. (2014). Waiting time distributions in the accumulating priority queue. *Queueing Systems*, 77(3):297–330.

[224] Taubenberger, J. and Morens, D. (2006). 1918 influenza: the mother of all pandemics. *Emerging Infectious Diseases*, 12(1):15–23.

[225] Taubman, S., Allen, H., Wright, B., Baicker, K., and Finkelstein, A. (2014). Medicaid increases emergency-department use: evidence from Oregon's health insurance experiment. *Science*, 343(6168):263–268.

[226] Taylor, I. and Templeton, J. (1980). Waiting time in a multi-server cutoff-priority queue, and its application to an urban ambulance service. *Operations Research*, 28(5):1168–1188.

[227] Taylor, R., Pare, J., Venkatesh, A., Mowafi, H., Melnick, E., Fleischman, W., and Hall, M. (2016). Prediction of in-hospital mortality in emergency department patients with sepsis: A local big data–driven, machine learning approach. *Academic Emergency Medicine*.

[228] Taylor, T. (2000). Position statement on the critical state of emergency care in Arizona. In *Arizona emergency care crisis meeting summary. Phoenix, AZ: Governor's administrative offices conference*.

[229] The U.S. Department of Health and Human Services (2017). Who is eligible for medicaid? Available at https://www.hhs.gov/answers/medicare-and-medicaid/who-is-eligible-for-medicaid/index.html (accessed May 2017).

[230] Thompson, S., Nunez, M., Garfinkel, R., and Dean, M. (2009). OR practice–efficient short-term allocation and reallocation of patients to floors of a hospital during demand surges. *Operations research*, 57(2):261–273.

[231] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288.

[232] Triandis, H. (1977). *Interpersonal behavior*. Brooks/Cole Publishing Company Monterey, CA.

[233] Tucker, J., Barone, J., Cecere, J., Blabey, R., and Rha, C. (1999). Using queueing theory to determine operating room staffing needs. *Journal of Trauma and Acute Care Surgery*, 46(1):71–79.

[234] Van der Heijden, M., Velikova, M., and Lucas, P. (2014). Learning bayesian networks for clinical time series analysis. *Journal of biomedical informatics*, 48:94–105.

[235] Vass, H. and Szabo, Z. (2015). Application of queuing model to patient flow in emergency department. case study. *Procedia Economics and Finance*, 32:479–487.

[236] Vollmann, T., Berry, W., and Whybark, D. (1993). *Integrated production and inventory management: revitalizing the manufacturing enterprise*. McGraw-Hill Professional Publishing.

[237] Volpatti, C., Leathley, M., Walley, K., and Dodek, P. (2000). Time-weighted nursing demand is a better predictor than midnight census of nursing supply in an intensive care unit. *Journal of critical care*, 15(4):147–150.

[238] Vuik, S., Mayer, E., and Darzi, A. (2016). A quantitative evidence base for population health: applying utilization-based cluster analysis to segment a patient population. *Population Health Metrics*, 14(1):44.

[239] Wagner, M., Tsui, F., Espino, J., Dato, V., Sittig, D., Caruana, R., McGinnis, L., Deerfield, D., Druzdzel, M., and Fridsma, D. (2001). The emerging science of very early detection of disease outbreaks. *Journal of public health management and practice*, 7(6):51–59.

[240] Wallis, L., Fleming, J., Hasselberg, M., Laflamme, L., and Lundin, J. (2016). A smartphone app and cloud-based consultation system for burn injury emergency care. *PloS one*, 11(2):e0147253.

[241] Wang, F., Lee, N., Hu, J., Sun, J., Ebadollahi, S., and Laine, A. (2013). A framework for mining signatures from event sequences and its applications in healthcare data. *IEEE transactions on pattern analysis and machine intelligence*, 35(2):272–285.

[242] Weiss, A., Schechter, M., and Chang, G. (2013). Case management for frequent emergency department users. *Psychiatric Services*, 64(7):715–716.

[243] Welch, S. (2012). Using data to drive emergency department design: a meta-synthesis. *HERD: Health Environments Research & Design Journal*, 5(3):26–45.

[244] Welch, S., Asplin, B., Stone-Griffith, S., Davidson, S., Augustine, J., and Schuur, J. (2011). Emergency department operational metrics, measures and definitions: results of the second performance measures and benchmarking summit. *Annals of emergency medicine*, 58(1):33–40.

[245] Whitt, W. (1993). Approximations for the GI/G/m queue. *Production and Operations Management*, 2(2):114–161.

[246] Whitt, W. (2007). What you should know about queueing models to set staffing requirements in service systems. *Naval Research Logistics (NRL)*, 54(5):476–484.

[247] Wilensky, U. (1999). Netlogo: Center for connected learning and computer-based modeling. Available at http://ccl.northwestern.edu/netlogo/ (accessed July 1st, 2014).

[248] Wiler, J., Bolandifar, E., Griffey, R., Poirier, R., and Olsen, T. (2013). An emergency department patient flow model based on queueing theory principles. *Academic Emergency Medicine*, 20(9):939–946.

[249] Wiler, J., Griffey, R., and Olsen, T. (2011). Review of modeling approaches for emergency department patient flow and crowding research. *Academic Emergency Medicine*, 18(12):1371–1379.

[250] Williams, S. and Heller, A. (2007). Patient activation among medicare beneficiaries: Segmentation to promote informed health care decision making. *International Journal of Pharmaceutical and Healthcare Marketing*, 1(3):199–213.

[251] Wu, J., Grannis, S., Xu, H., and Finnell, J. (2016). A practical method for predicting frequent use of emergency department care using routinely available electronic registration data. *BMC emergency medicine*, 16(1):12.

[252] Xu, K. and Chan, C. (2016). Using future information to reduce waiting times in the emergency department via diversion. *Manufacturing & Service Operations Management*, 18(3):314–331.

[253] Yahav, I., Barnes, S., Golden, B., and Wasil, E. (2013). Early detection of bioterrorism: Monitoring disease diffusion through a multilayered network. In *IIE Annual Conference. Proceedings*, page 2561. Institute of Industrial and Systems Engineers (IISE).

[254] Yankovic, N. and Green, L. (2011). Identifying good nursing levels: A queuing approach. *Operations research*, 59(4):942–955.

[255] Yom-Tov, G. and Mandelbaum, A. (2014). Erlang-R: A time-varying queue with reentrant customers, in support of healthcare staffing. *Manufacturing & Service Operations Management*, 16(2):283–299.

[256] Young, B., Lin, E., Von Korff, M., Simon, G., Ciechanowski, P., Ludman, E., Everson-Stewart, S., Kinder, L., Oliver, M., and Boyko, E. (2008). Diabetes complications severity index and risk of mortality, hospitalization, and healthcare utilization. *The American journal of managed care*, 14(1):15.

[257] Zayas-Caban, G., Xie, J., Green, L., and Lewis, M. Optimal control of an emergency room triage and treatment process. Research Paper No. 14-51, Columbia Business School, Columbia University, New York, NY.

[258] Zeltyn, S., Marmor, Y., Mandelbaum, A., Carmeli, B., Greenshpan, O., Mesika, Y., Wasserkrug, S., Vortman, P., Shtub, A., Lauterman, T., et al. (2011). Simulation-based models of emergency departments:: Operational, tactical, and strategic staffing. *ACM Transactions on Modeling and Computer Simulation (TOMACS)*, 21(4):24.

[259] Zonderland, M., Boucherie, R., Carter, M., and Stanford, D. (2015). Modeling the effect of short stay units on patient admissions. *Operations research for health care*, 5:21–27.