DATA-DRIVEN QUALITY OF SERVICE IMPROVEMENTS IN HOSPITALS

Dongqing Yu

A dissertation submitted to the faculty of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Statistics and Operations Research.

Chapel Hill
2016

Approved by:

Haipeng Shen

Serhan Ziya

Nilay Argon

Chuanshu Ji

Jason Katz

ABSTRACT

DONGQING YU: Data-Driven Quality of Service Improvements in Hospitals
(Under the direction of Haipeng Shen and Serhan Ziya)

In recent years, there has been an increasing interest in developing novel methods for effective and efficient healthcare service delivery and data analytics has widely been recognized as being essential for decision-making at various healthcare service settings. This dissertation consists of three research projects each aiming to improve decisions at three different healthcare settings with one being related to critical care delivery and the other two being related to inpatient patient flow management. Each project combines statistical analysis of hospital data with techniques and methodologies from operations research.

The first project concerns care delivery in the cardiac intensive care unit (CICU). We analyze a prospective study to describe admissions and care practices within the CICU of the UNC Hospital, and also evaluate the influence of an open versus closed model of care on patient outcomes and resource consumption. The second and third projects study management of patient flow from emergency department (ED) to hospital internal wards (IWs). Both projects develop effective inpatient flow management policies with the objective of reducing ED boarding time, which is defined as the time between the decision for admission for an ED patient and the time the patient is physically admitted to an IW. Delayed admission to IWs has been identified as a key factor for ED overcrowding and is a big challenge for many hospitals. We approach the problem from two different angles: early discharging patients to free up beds in IWs, and early requesting beds for ED patients based on early prediction of need for IW beds. In both projects, we develop relevant statistical models for analyzing the detailed hospital patient flow data and build mathematical decision models to develop new methods. We also build simulation models and use these models to investigate the benefits of the proposed methods. Simulation studies suggest significant potential improvements in various performance measures of interest.

ACKNOWLEDGEMENTS

Pursuing the Ph.D. degree is a long, and most of the time, stressful journey. I am deeply grateful to all the people who have helped me during this journey.

First of all, I would like to thank my two advisors, Professor Haipeng Shen and Professor Serhan Ziya. They have dedicated an enormous amount of time, numerous efforts, and great patience to guide me on my thesis research. Without their support and continuous encouragement, I would not be able to finish my thesis. They are great mentors not only in research but also in life; every meeting with them is enjoyable and inspiring. I am truly lucky to have the opportunity to work with and learn from them.

I would like to thank Professor Chuanshu Ji, Professor Nilay Argon, and Professor Jason Katz for serving on my dissertation committee, providing insightful comments to my thesis. More specifically, I would like to thank Professor Chuanshu Ji for offering generous help during my first two years study. I would like to thank Professor Nilay Argon for teaching me and answering my questions regarding both research and life. I would like to thank Professor Jason Katz for the opportunity to work on the CICU project and the practical insights from a physician perspective.

I would like to thank my friends: Qing Feng, Siliang Gong, Gen Li, Siying Li, Jenny Shi, Dong Wang, Jie Xiong, Tan Xu, Dan Yang, Qunqun Yu, and Yu Zhang. Your company and support have brought me a lot of joy during this long journey.

I would also like to thank Minghui Liu. He always believes in me and gives me vigorous support during this journey. His encouragement is particularly important when I experienced setbacks during my research. He deserves as much credit as I do for completing this thesis.

Last but not the least, I want to pass my deepest thanks to my parents for all their endless love and unconditional support. Without them, I cannot finish this long journey. It is my privilege to be your daughter. I love you!

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1
Introduction

In recent years, there has been an increasing amount of attention on effectively and efficiently delivering healthcare services. Besides hospital administrators and medical practitioners (physicians, nurses, ...), academic researchers have joined forces in this exciting and important endeavor. Among them are queueing and operations management researchers who have taken the queueing-network perspective of patient flows in the healthcare delivery systems in general, and hospitals in particular, aiming at ways that appropriately balance quality and efficiency to streamline operations, reduce congestion and delay, increase system throughput, while providing better quality of service to patients.

This dissertation continues a recent popular research thread in this area and makes contributions through three projects. The research combines detailed hospital patient-flow data with queueing-theoretical perspectives, and uses data-driven approaches to better match hospital resources with demand for healthcare services (Armony et al., 2011; Shi et al., 2014). The three projects share a unique and common novelty: they effectively combine tools from statistics, stochastic modeling, and optimization. Modern hospitals consist of many medical units, among which our research considers efficient operations in the following three such units: cardiac intensive care unit (CICU), emergency departments (ED), and hospital internal wards (IWs). All studies take advantage of real patient flow data collected at two hospitals: the UNC Hospital and the Rambam Hospital in Israel.

1.1 Cardiac Intensive Care Unit (CICU)

Intensive care units (ICUs) are hospital wards specialized in the care of patients with critical illnesses that require continuous monitoring and treatment. ICUs are exceedingly costly, consuming more than 20% of total hospital costs despite constituting less than 10%

of hospital beds in the US (Chalfin et al., 1995). CICU is the one dedicating to critically ill cardiovascular patients. A growing evidence supports that patients now occupying CICU have become increasingly susceptible to multisystem organ injury and more frequent consumers of costly critical care resources (Katz et al., 2010; Morrow et al., 2012).

Care models in the ICU have classically been described as either closed or open, depending upon the presence or absence of a dedicated critical care team. While a closed model has been shown to improve patient outcomes in medical and surgical ICUs, the merits of various care models have not been previously explored in the CICU setting. Given the complexity and expenditure in the CICU, there is a compelling need to better understand and develop optimal practice models for the effective and efficient care delivery in the CICU.

Chapter 2 investigates patient conditions and compare the two care models in the current CICU in many ways. The data that support this study is prospectively collected from the UNC Hospital. We identify and describe variations among demographics, comorbidities and admission diagnoses, along with variability in the use of both cardiac and non-cardiac critical care resources. To thoroughly compare the two care models in the CICU from both clinical and fiscal perspectives, we evaluate patient outcomes in terms of CICU and hospital mortality, resource utilization, and length of stay. We find no significant impact of the CICU structure on either CICU or hospital mortality. Although some significant differences in the resource utilizations are found, it is not the case that one model is consistently more resource-conservative than the other. We do find that patients have shorter length of stay in the closed unit. This study sheds lights for the first time on understanding the pros and cons of such operational models in the CICU.

1.2 Emergency Departments (ED) and Internal Wards (IW)

Long waiting times and length of stays in ED are ubiquitous in many parts of the world. In ED, a long waiting time is more than an inconvenience as it can result in many adverse outcomes including death (Bernstein et al., 2009; Sun et al., 2013). Besides a variety of reasons behind prolonged waiting times and extended ED stays, the inability to promptly move admitted patients from ED to inpatient units has been identified as

a key contributor (Asplin et al., 2003; Trzeciak and Rivers, 2003; Institute of Medicine, 2007; Abraham and Reddy, 2010; National Center for Health Statistics, 2013). The time of waiting for admission to an inpatient bed, known as the ED boarding time, refers to the duration between the time when doctors decide to admit a patient and the time when the patient is physically transferred to an inpatient bed. Prolonged ED boarding times not only prevent the admitted patients from getting proper level of care in inpatient units, but also result in ED congestion and decrease the hospital throughput, as such patients still occupy ED resources that otherwise could have been used for the current and incoming ED patients. Therefore, effective inpatient flow management is essential to proved better quality of care to patients.

Chapters 3 and 4 focus on the patient flow from the ED to one of the inpatient units – internal ward (IW), and aim to develop insights into effective inpatient flow management from the perspective of the availability of inpatient beds and the timeliness of admission decision, respectively. The data that support these two studies are from the Rambam Hospital in Israel, which have been analyzed in detail in Armony et al. (2011). The data are inter-departmental, in that they record patient flows from ED to IW, while ignoring detailed time stamps within each unit (ED or IW). Our work and contributions in each chapter are briefly summarized below.

Active Bed Management through Patient Early Discharges

It is important to understand the inpatient flow problem from a strict bed demand and supply perspective. Demand for beds on any given day can simply exceed the available bed capacity even after daily patient discharges are completed. Even if the number of available beds post-discharge is enough to meet the daily bed demand, the bulk of bed requests can occur before most of the discharges are complete, which results in a temporary bed shortage during the day. This "misalignment" of bed requests (demand) and patient discharges (supply), which is very common in practice, is one of the main causes of the prolonged ED boarding times.

To address the "misalignment" issue, researchers have proposed to discharge patients in inpatient units earlier in the day (Howell et al., 2008; Vicellio et al., 2008; Powell et al., 2012; Shi et al., 2014). The effectiveness of the early discharge in reducing ED boarding

time has been demonstrated. However, the realization of this potential benefit in practice requires some efforts, which are not cost-free. In addition, because of daily fluctuations in patient arrival and discharge, the severity of the bed shortage problem also changes from day to day. Therefore, it would be essential that hospitals could be provided with a clear guidance on how to manage patient discharges, in order to achieve the pre-specified service target at the beginning of each day.

The main objective of Chapter 3 is to provide a data-driven patient early discharge plan. We investigate the effects of active bed management within IW, through anticipating needs for IW beds and freeing up bed capacity by early discharging IW patients who are medically feasible. We demonstrate that it can reduce ED boarding times by early discharging feasible patients in IW, which matches the finding in the literature. Our main contribution is to provide hospitals a data-driven solution to the bed shortage problem through the following framework.

To begin with, we identify that the initial IW occupancy, the number of arrivals to IW, and the number of discharges from IW are three leading factors to the ED boarding time; we then build a statistical model to characterize the relationship. Secondly, we consider two approaches for discharging patients early: the first one is to discharge patients earlier in the same day, while the other one is to discharge patients one day ahead; we then formulate two discharge policies through math programming to determine the smallest number of early discharges necessary in order to meet a particular target level for boarding times. Finally, we build a discrete event simulation model and carry out a wide range of simulation studies to illustrate potential benefits of our proposed approaches and the corresponding policies. To the best of our knowledge, our work is the first to provide a practical guideline for hospitals to implement early discharge based on their particular needs.

Active Bed Management via Early Bed Requests

Bed request and preparation for admitting a patient from ED to inpatient units won't start until the admission decision is certain, which usually happens at the end of ED service when all test results and diagnoses are available. A typical ED patient spends several hours in the ED before the admission decision is made. In the literature, a lot of studies predict the probability of admission at the time of ED triage, which makes early admission decisions

possible. If the decision of needing a bed can be made for an ED patient while he is still in the course of ED service, then it is expected that by the time when this patient is ready for the transfer, one bed will have been made available or in preparation, so that the transfer can happen smoothly without much delay.

There are studies considering the early bed request option for each individual ED patient one by one, which demonstrate that a significant amount of time waiting for admission can be saved. However, making early bed request decision for each individual patient is too sensitive to the prediction accuracy. Therefore, it is meaningful to investigate the aggregated bed demand in the near future and request inpatient beds in a batch in advance to take care of that aggregated demand.

The main objective of Chapter 4 is to estimate the aggregated bed demand and then use the estimate to reduce ED boarding times. We investigate the effects of active timing of admission decision making within ED, through anticipating needs for IW beds. First, we demonstrate that it can reduce ED boarding times by early requesting an inpatient bed for an ED patient based on the prediction of her/his own admission probability, which is consistent with the findings in the literature. In addition, we illustrate the compensation and cost for the realization of early bed request. Secondly, we formulate a bed request policy through math programming to determine the aggregated bed demand in the future and the appropriate number of bed requests needed in advance, in order to achieve the right balance between the benefits and costs of early bed requests. Thirdly, we conduct a variety of empirical studies based on the real hospital data. Fourthly, we develop and calibrate the simulation model using empirical results to capture inpatient flow from ED to IW. Finally, we conduct a series of simulation studies to demonstrate the good performance of our proposed bed request policy. We systematically discuss the pros and cons of early bed requests on the ED boarding time and other measures of service quality in hospitals. In addition, to the best of our knowledge, our study is the first to use the aggregated bed demand estimation based on the prediction of admission probability for ED patients to guide the bed request decision.

CHAPTER 2

Investigating the Benefits of Closed Unit Structure in the Cardiac Intensive Care Unit

2.1 Introduction

The coronary care unit (CCU), with its inception dating back to the 1960s, has been credited with significantly improving the care and survival of patients hospitalized with acute myocardial infarction (e.g., Braunwald, 1998). Until recently, however, little is known about the modern CCU and its influence on contemporary patient outcomes. In spite of a limited evidence-base, these specialized units have nonetheless become pervasively embedded within today's healthcare systems – particularly those within academic medical centers. What has been more recently acknowledged is that the CCU has evolved considerably over the last several decades (e.g., Katz et al., 2007), underscored by a growing patient population admitted with increasing illness severity, for a myriad of cardiovascular maladies, and with advancing critical illness (e.g., Katz et al., 2010). At the same time, patients now occupying cardiac intensive care units (CICUs) have become increasingly susceptible to multisystem organ injury and more frequent consumers of costly critical care resources (e.g., Katz et al., 2010; Morrow et al., 2012).

Given the multiplicity and complexity of disease within today's CICU, there is a compelling need to better understand and develop optimal practice models for the delivery of effective and efficient care. This is particularly true in an era in which past dysfunctional processes and ineffective team dynamics have placed critically ill patients at enhanced and undue risk for medical error and poor outcomes (e.g., Soumerai and Avorn, 2001). In the intensive care unit (ICU), two models of healthcare delivery are most commonly employed – an open and closed model. In the open staffing model, every admitted patient has his own physician who determines the need for ICU admission and discharge, and who makes

all primary management decisions. In the closed model, on the other hand, all patients are managed by a single practitioner or team who is responsible for directing clinical care while the patient is in the ICU.

In the critical care literature, a lot of evidence suggests that the closed model of ICU care can both improve patient outcomes and reduce critical care expenditures (e.g., Carson et al., 1996; Multz et al., 1998; Ghorra et al., 1999). Though a recent survey of medical directors has provided some clarity on the contemporary CICU landscape, and described a predominance of closed care units within the United States (OMalley et al., 2013), organizational models have never been directly studied in the cardiovascular critical care population.

In order to prospectively describe admissions and care practices within the CICU of a large, tertiary care, academic hospital, and also to evaluate the influence of an open versus closed model of care on patient outcomes and resource consumption, the following study is conducted.

2.2 Description of Models of Care and Data

2.2.1 Patient Population and Models of Care

The University of North Carolina (UNC) is an 805-bed tertiary care, not-for-profit teaching hospital embedded within a large healthcare system owned by the state of North Carolina. It operates a 13-bed CICU dedicated to the management of all patients admitted with a primary cardiovascular diagnosis who require critical care monitoring and support. Prior to July 2013, the UNC CICU operated in an open model of care delivery, and had done so for nearly two decades. As part of this model, multiple physician-led teams cared for patients admitted to the CICU and then continued to manage these individuals once they were moved out of the unit to either an intermediate-care (stepdown) or general cardiology bed. In addition to the other patients who never required CICU-level care during their admission, these teams were also ultimately responsible for discharging patients following their acute hospitalization. Each team was composed of an attending-level faculty cardiologist, a cardiology fellow, several junior and senior Internal Medicine residents, and

medical students, and each team followed only patients that had been admitted directly by them. Each night, one of the teams was "on-call" and responsible for hospital admissions in addition to providing medical support to all other hospitalized cardiac patients as needed.

In July 2013, as part of a planned structural transition, care delivery in the CICU was changed from an open to a closed model. In the closed model, one physician-led team was tasked with caring primarily for all cardiology patients requiring intensive care. This team was composed of an attending-level faculty cardiologist, two cardiology fellows and two residents paired to work alternating 12-hour shifts, along with 1-2 interns and medical students present only during the daytime hours. When a patient was deemed stable enough to transfer out of the CICU they were then received by a separate cardiology floor/intermediate-care team who managed these individuals until hospital discharge.

There was no difference in training or expertise between attending physicians who participated in the open or closed CICU; however, only a small cohort of these cardiologists (10 out of 25) was ultimately required to staff the closed unit. The remainder of the inpatient staff cardiologists directed the care of cardiovascular patients needing floor or stepdown beds. Only 1 cardiologist had advanced critical care training, and he attended during both the open and closed CICU periods for an equal amount of time.

2.2.2 Data

In light of the anticipated structural transition, data were collected prospectively with the goal of addressing changes in outcomes that could potentially be attributable to the model of care delivery. From November 2012 to June 2013, while the CICU was operating in an open format, data were collected on all consecutive patients $\geq$18 years of age admitted to the unit with a primary cardiovascular diagnosis. During a planned 2-month transition period during which the unit was changed from an open to closed model of care, no data were recorded. Beginning in September 2013 through March 2014, data were once again collected prospectively on consecutive CICU patients admitted to the closed unit. Patients were excluded if they were admitted by a non-cardiac service or managed primarily by a surgical team.

Baseline demographic and clinical variables, admission and discharge diagnoses, medical comorbidities, resource use, and outcomes were recorded for all patients. Trained abstractors were used for data collection, and serial assessments of data quality and completeness were performed throughout the process. Disease severity at presentation was assessed using the modified Acute Physiology and Chronic Health Evaluation II (APACHE II) Score and the Simplified Acute Physiology Score II (SAPS II), both of which were initially derived and have been subsequently validated for use in the intensive care setting (e.g., Knaus et al., 1985; Le Gall et al., 1993). APACHE II and SAPS II scores for all eligible patients were determined from clinical information obtained during the first 24 hours of admission to the CICU. In each case, a higher score is indicative of greater illness. In addition, admission source was captured for all patients, and discharge disposition was recorded for those individuals who ultimately left the hospital alive.

2.3 Data Analysis

Patient baseline demographic and clinical variables are presented as means with standard deviations for continuous variables and as frequencies with percentages for categorical data. To evaluate differences between patients treated in the open versus closed models, categorical variables are compared using the chi-squared or Fishers exact test where appropriate; continuous variables are compared using the non-parametric Wilcoxon rank sum test. Logistic regression, adjusting for patient disease severity at presentation, is used to examine the impact of unit structure on the binary outcome of CICU mortality. Because APACHE II and SAPS II scores are highly correlated, only the APACHE II score is used for modeling purposes. CICU mortality is also analyzed by time-to-event survival analysis. A Kaplan-Meier curve is plotted and compared using the log-rank test. In addition, Cox proportional hazard models are fitted to estimate hazard ratios and 95% confidence intervals for comparisons of unit structure on mortality for a number of pre-specified subgroups. These subgroups include: age ($<$75 or $\geq$75 years), sex, race, BMI ($<$30, 30-35, $\geq$35 kg/m^2), primary admission diagnosis (heart failure, acute MI, cardiogenic shock, cardiac arrest and sepsis), APACHE II score ($<$22 or $\geq$22), SAPS II score ($<$42 or $\geq$42), and the use of inva-

sive mechanical ventilation. All reported p-values are two-sided and one-sided if necessary and considered statistically significant for $p < 0.05$. This study was reviewed and approved by our institutions Office of Human Research Ethics.

2.3.1 Demographics and Comorbidities

The entire study population consisted of 670 patients, 332 (49.6%) of whom were admitted to the open model CICU and 338 (50.4%) of whom were admitted during the closed model of care. Baseline characteristics are shown in Table 2.1. Demographic variables were largely similar between the two groups, although there were more Black patients admitted during the closed CICU study period. Medical comorbidities were also quite similar between the two cohorts, with the exception of prior MI which was more common among closed model patients and history of pulmonary hypertension and previous CABG which were both more commonly found during the open model of care.

2.3.2 Admission Source

Patients placed in the CICU during the open model study period were more often admitted from the Emergency Department (ED), while transfers to the CICU from a floor/general ward bed were more common in the closed model of care. Additional admission sources can be seen in Table 2.2.

2.3.3 Admission Diagnoses and Severity-of-Illness

Table 2.2 also shows the primary diagnosis for CICU admission, along with other secondary diagnoses determined during hospitalization. In general, these were similar between the two study populations. The most common reasons for admission to the CICU in general were acute ischemia/infarction (STEMI and NSTE-ACS), acute heart failure, cardiogenic shock, and arrhythmia, although Table 2.2 underscores the broad myriad of conditions that were ultimately felt to require treatment in the intensive care setting. Only cardiac arrest was found to be a statistically more common reason for admission in the closed CICU.

Table 2.1: Baseline characteristics of the study population

	Open Unit (n=332)	Closed Unit (n=338)	p-value two-sided	p-value one-sided
Demographics				
Age (years), mean (SD)	63 (15)	63 (15)	0.70	
Sex, No. (%)			0.69	
Male	207 (62)	205 (61)		
Female	125 (38)	133 (39)		
Race, No. (%)			0.03	
White	192 (58)	188 (56)	0.59	
Black	86 (26)	113 (33)	0.03	0.02
Other	54 (16)	37 (11)	0.05	
Weight (kg), mean (SD)	87.1 (23.2)	87.3 (25.1)	0.55	
BMI (kg/m^2), mean (SD)	29.5 (7.5)	29.8 (8.2)	0.94	
Comorbidities, No. (%)				
CAD	167 (50)	180 (53)	0.49	
MI	72 (22)	128 (38)	<0.001	<0.001
PCI	70 (21)	63 (19)	0.44	
CABG	54 (16)	36 (11)	0.04	0.02
CHF	186 (56)	152 (45)	0.004	0.003
CVD	46 (14)	33 (10)	0.12	
PVD	40 (12)	25 (7)	0.05	0.03
CKD	80 (24)	80 (24)	0.93	
ESRD	25 (8)	26 (8)	1.00	
HTN	235 (71)	221 (65)	0.14	
Hyperlipidemia	163 (49)	142 (42)	0.07	
Diabetes Mellitus	118 (36)	122 (36)	0.94	
Chronic Lung Disease	84 (25)	67 (20)	0.10	
Chronic Liver Disease	18 (5)	17 (5)	0.86	
Cancer (within 5 years)	29 (9)	39 (12)	0.25	
Severe Valvular Disease	19 (6)	30 (9)	0.14	
Pulmonary Hypertension	34 (10)	18 (5)	0.02	0.01
History of ICD/PPM	74 (22)	63 (19)	0.25	

Abbreviations: BMI: Body Mass Index, CAD: Coronary Artery Disease, MI: Myocardial Infarction, PCI: Percutaneous Coronary Intervention, CABG: Coronary Artery Bypass Graft, CHF: Congestive Heart Failure, CVD: Cerebrovascular Disease, PVD: Peripheral Vascular Disease, CKD: Chronic Kidney Disease, ESRD: End-Stage Renal Disease, HTN: Hypertension, ICD/PPM: Implantable Cardioverter-Defibrillator/Permanent Pacemaker

Disease severity was determined by both the APACHE II and SAPS II scoring systems. There was considerable correlation between the two scores (correlation of 0.89 among open CICU patients and 0.85 among closed unit patients). While statistically significant differences did exist between the two patient cohorts, numerically these differences were small. The mean APACHE II score was 18 +/− 10 in the open study group and 16 +/− 11 in the closed one. These scores would predict an average anticipated ICU mortality of 32% and 28%, respectively. Similarly, the mean SAPS II score was 36 +/− 18 in the open

model CICU and 33 +/− 18 during the closed model of care. These scores would predict an average anticipated ICU mortality of 25% and 20%, respectively. In aggregate, these results suggest that less ill patients may have been admitted during the closed CICU study period as determined by previously validated ICU severity-of-illness measures. Finally, consistent with these findings, but not reaching statistical significance, more delirious patients (CAM-ICU positive) were admitted to the open compared to the closed CICU (Table 2.2).

2.3.4 Resource Consumption and Procedures

There was considerable variability in the use of both cardiovascular and non-cardiovascular critical care resources throughout the study. In aggregate, over 1/3 of patients treated in the CICU underwent coronary angiography, 25% had a central venous catheter and intra-arterial line, and more than 20% required invasive mechanical ventilation (Table 2.3). In the open model of CICU care, patients utilized significantly more inotropic agents ($p < 0.001$), non-invasive positive-pressure ventilation ($p < 0.001$), anti-arrhythmic medications ($p < 0.001$), and pericardiocentesis ($p = 0.03$). On the other hand, patients in the closed model CICU utilized significantly more intra-arterial lines ($p < 0.001$), vasopressor agents ($p = 0.01$), transthoracic echocardiography ($p = 0.02$), and veno-arterial extracorporeal membrane oxygenation (VA-ECMO) ($p = 0.02$).

2.3.5 Patient Outcomes

Despite the aforementioned predicted mortality rates ranging from 20% to 33% by APACHE II and SAPS II scoring systems, overall CICU and hospital mortality were only 11.5% and 14.5%, respectively. There is no statistical difference in either CICU or hospital mortality when comparing open versus closed unit models (Table 2.4). In addition, we investigate the difference of survival in CICU between two unit structures using survival functions: we plot the Kaplan-Meier curve for each unit structure in Figure 2.1, and conduct the logrank test. Under the null hypothesis that they two are not different, the 1-df chi-square statistic has the value of 0.7, with the corresponding p-value of 0.388. It suggests that survival functions are not statistically significantly different between the open and

Table 2.2: Primary admission source, disease severity, and diagnoses

	Open Unit (n=332)	Closed Unit (n=338)	p-value two-sided	p-value one-sided
Source of Admission, No. (%)			<0.001	
ED	104 (31)	78 (23)	0.02	0.01
Other Hospital	97 (29)	98 (29)	1.00	
Clinic	28 (8)	1 (0)	<0.001	<0.001
Floor Bed	100 (30)	160 (47)	<0.001	<0.001
Operating Room or Procedural Area	3 (1)	0 (0)	0.12	
Severity of Illness, mean (SD)				
APACHE II score	18 (10)	16 (11)	<0.001	<0.001
APACHE II-Predicted Mortality (%)	32 (25)	28 (26)	<0.001	<0.001
SAPS II score	36 (18)	33 (18)	<0.001	<0.001
SAPS II-Predicted Mortality (%)	25 (27)	20 (26)	<0.001	<0.001
Delirium (CAM-ICU Positive)	65 (20)	55 (16)	0.31	
Primary Admission Diagnosis, No. (%)				
STEMI	36 (11)	43 (13)	0.47	
NSTE-ACS	45 (14)	56 (17)	0.28	
Acute Heart Failure	30 (9)	47 (14)	0.05	
Cardiogenic Shock	52 (16)	35 (10)	0.05	
Cardiac Arrest				
Primary Rhythm - VT/VF	6 (2)	6 (2)	1.00	
Primary Rhythm - PEA/Asystole	12 (4)	29 (9)	0.01	0.01
Arrhythmia	37 (11)	24 (7)	0.08	
Conduction Disease	12 (4)	13 (4)	1.00	
Acute Valvular Disease	8 (2)	5 (1)	0.42	
Cardias Tamponade	10 (3)	6 (2)	0.32	
Acute Respiratory Failure	9 (3)	8 (2)	0.81	
Sepsis/Infection	13 (4)	16 (5)	0.71	
DVT/PE	2 (1)	2 (1)	1.00	
Other Secondary Diagnoses, No. (%)				
Acute Respiratory Failure	28 (8)	29 (9)	1.00	
Acute Renal Failure	55 (17)	48 (14)	0.45	
Sepsis/Infection	6 (2)	15 (4)	0.07	
Cardiac Arrest	10 (3)	6 (2)	0.32	
PE/DVT	11 (3)	14 (4)	0.68	
Acute (Non-Intracerebral) Hemorrhage	4 (1)	10 (3)	0.18	
Acute Heart Failure Exacerbation	30 (9)	11 (3)	0.002	0.001
Cardiogenic Shock	9 (3)	11 (3)	0.82	
CVA/TIA	4 (1)	11 (3)	0.11	

Abbreviations: APACHE: Acute Physiology and Chronic Health Evaluation, SAPS: Simplified Acute Physiology Score, CAM-ICU: Confusion Assessment Method for the ICU, STEMI: ST-segment Elevation Myocardial Infarction, NSTE-ACS: Non-ST-segment Elevation Acute Coronary Syndrome, VT/VF: Ventricular Tachycardia/Ventricular Fibrillation, PEA: Pulseless Electrical Activity, DVT/PE: Deep Vein Thrombosis/Pulmonary Embolus, CVA/TIA: Cerebrovascular Accident/Transient Ischemic Attack

closed units, and confirms our finding in Table 2.4. Visually Kaplan-Meier curves appear to separate from each other after 10 days. It suggests that patients who stay longer in the CICU might benefit from the closed model structure. Further, we still do not find significant

Table 2.3: Resource utilization, No. (%)

	Open Unit (n=332)	Closed Unit (n=338)	p-value two-sided	p-value one-sided
Coronary Angiography	119 (36)	130 (38)	0.52	
PCI	72 (22)	76 (22)	0.85	
Central Venous catheter	77 (23)	68 (20)	0.35	
Intra-arterial Line	48 (14)	121 (36)	<0.001	<0.001
Vasopressor Use	58 (17)	84 (25)	0.02	0.01
Inotrope Use	70 (21)	36 (11)	<0.001	<0.001
Mechanical Ventilation				
Invasive	64 (19)	85 (25)	0.08	0.04
Non-Invasive	41 (12)	9 (3)	<0.001	<0.001
Transthoracic Echocardiogram (TTE)	255 (77)	283 (84)	0.03	0.02
Transesophageal Echocardiogram (TEE)	8 (2)	10 (3)	0.81	
IABP	14 (4)	20 (6)	0.38	
VA-ECMO	0 (0)	6 (2)	0.03	0.02
Anti-arrhythmic Therapy	86 (26)	47 (14)	0.0001	<0.001
TH/TTM	14 (4)	7 (2)	0.12	
Temporary Pacer	3 (1)	1 (0)	0.37	
PPM Implantation	15 (5)	15 (4)	1.00	
ICD	10 (3)	17 (5)	0.24	
Ablation	16 (5)	15 (4)	0.86	
DCCV/Defibrillation	21 (6)	19 (6)	0.75	
Bronchoscopy	3 (1)	2 (1)	0.68	
Thoracentesis	6 (2)	5 (1)	0.77	
Paracentesis	0 (0)	2 (1)	0.50	
Pericardiocentesis	14 (4)	5 (1)	0.04	0.03
Lumbar Puncture	1 (0)	1 (0)	1.00	
Endoscopy	10 (3)	6 (2)	0.32	

Abbreviations: PCI: Percutaneous Coronary Intervention, IABP: Intraaortic Balloon Pump, VA-ECMO: Veno-Arterial Extracorporeal Membrane Oxygenation, TH/TTM: Therapeutic Hypothermia/Targeted Temperature Management, PPM: Permanent Pacemaker, ICD: Implantable Cardioverter-Defibrillator

difference in the CICU mortality between two unit structures when comparing a number of pre-specified patient subgroups (Figure 2.2), though more point estimates consistently favor the closed model structure. The analysis detail is left in Section 2.5.3.

From a length-of-stay perspective, patients spent an average of 1 day less in the CICU during the closed model of care ($p = 0.02$). Among patients who did survive their acute hospitalization, many (25%) required additional post-discharge resources including referral to skilled nursing facilities, rehabilitation centers, transfer to other acute care hospitals, or the use of Hospice (Table 2.4). These requirements were similar regardless of whether patients were admitted to the open or closed CICU. There were, however, slightly greater rates of hospital-to-hospital transfers following discharge from the closed CICU ($p = 0.01$).

Table 2.4: Primary outcomes and disposition

	Open Unit (n=332)	Closed Unit (n=338)	*p*-value two-sided	*p*-value one-sided
Clinical Outcomes				
Mortality, No. (%)				
In CICU	43 (13)	34 (10)	0.28	
In Hospital	47 (14)	51 (15)	0.74	
LOS (days), median (IQR)				
In CICU	3 (1, 5)	2 (1, 5)	0.04	0.02
In Hospital	5 (3, 11)	5 (3, 10)	0.77	
Patient Disposition*, No. (%)			0.06	
Home	239 (84)	227 (79)	0.19	
Rehab/SNF	35 (12)	38 (13)	0.80	
Hospice	10 (4)	13 (5)	0.67	
Transferred to Another Acute Care Hospital	0 (0)	7 (2)	0.02	0.01
Other Transfer	1 (0)	1 (0)	0.75	

* For patients who survived index hospitalization (open, n=285; closed, n=286)
Abbreviations: LOS: Length of Stay, SNF: Skilled Nursing Facility

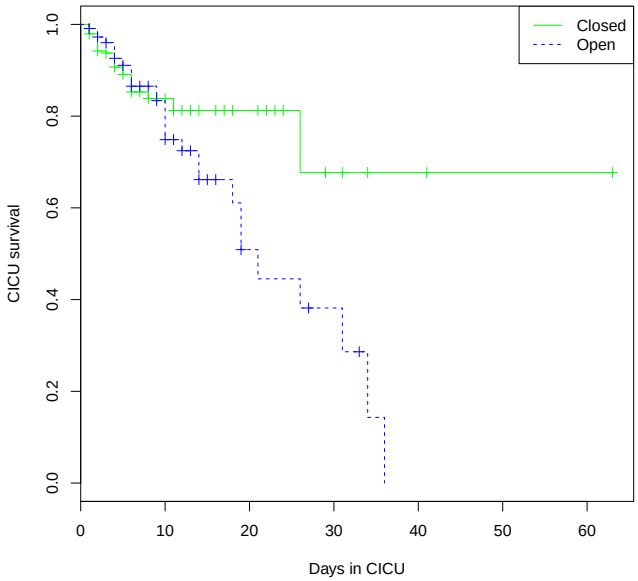


Figure 2.1: Kaplan-Meier survival curve for the CICU mortality

2.4 Discussion

Emerging data has helped to characterize a striking evolution in the CICU. Once developed solely for the management of patients suffering from acute myocardial infarction, a burgeoning retrospective evidence-base now purports that the contemporary CICU is currently home to an increasingly complex, resource intensive, and diverse population of

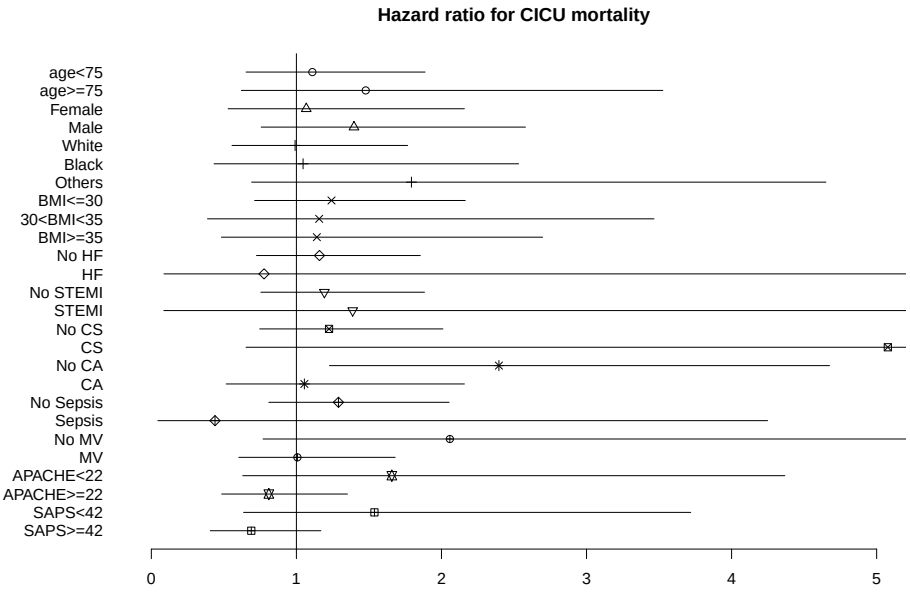


Figure 2.2: Hazard ratio for CICU mortality among pre-specified patient subgroups

patients presenting with a wide variety of cardiac and non-cardiac critical illnesses (e.g., Katz et al., 2007; Katz et al., 2010). We report similar findings in what we believe to be the first prospective evaluation focusing on unselected patients hospitalized in an academic, tertiary care CICU. Confirming other observational reports, we describe a heterogeneous group of patients, admitted with a variety of primary diagnoses, who require frequent critical care resources and long-term care support.

2.4.1 Is the CICU Beneficial?

Although extensively employed within modern hospitals and healthcare systems, the true benefits of the CICU have never actually been validated. In fact, all prior studies addressing the merits of CICU care have been largely experiential or historical reports, published in an era predating contemporary treatment protocols and analyzed with limited scientific rigor (e.g., Killip and Kimball, 1967). Despite this, these specialized units have flourished, particularly within academic settings. While the current study does not directly assess the effectiveness of CICU care for the management of patients with cardiovascular critical illness over other settings, our findings certainly do not refute the commonly held

notion that the CICU is a beneficial tool (e.g., Braunwald, 1998). We found that overall CICU mortality at our institution was slightly greater than 11%. Well-studied ICU calculators tell us that we should have expected a mortality rate in the range of 25-30% based upon presenting illness severity. While the APACHE II and SAPS II models may be poor predictors of outcome in cardiovascular critical care cohorts - and admittedly have not been validated among such populations - this finding should be interpreted with cautious optimism. Whether this is the direct result of CICU care, the result of other non-ICU related processes, or a combination of the two remains unclear. However, given the pervasive nature of the CICU within contemporary healthcare systems, there is little presented here which should undermine the enthusiasm for this care delivery platform.

2.4.2 What is the Best Way to Deliver CICU Care?

Accepting then that the CICU is an important component of cardiovascular care, the next question to ask is how best to use the CICU. While we have a wealth of data helping to inform care practices for specific disease states - including STEMI (e.g., O'Gara et al., 2013; Jernberg et al., 2011) and cardiogenic shock (e.g., Hochman et al., 1999) - we know very little about optimal models and structures of care for aggregate CICU patient populations. We believe our study is, in fact, the first to address this topic. This is not to suggest that care models have not been previously examined in general ICU settings; on the contrary, there is indeed a fairly substantial body of evidence supporting closed models of ICU care in a variety of non-cardiac critical care units (e.g., Carson et al., 1996; Multz et al., 1998; Ghorra et al., 1999). No data, however, exists among today's CICU patients.

We found no significant impact of CICU structure on patient mortality. Though nominally improved CICU death rates were found during the closed model of care, the lower-than-expected mortality rendered our study insufficiently powered to truly assess this robust endpoint. Additional study, in considerably larger and multicenter cohorts, will be needed to address survival as a product of CICU structure. Careful review of the Kaplan-Meier survival curves, however, may prove somewhat illustrative and add insight into future study design. While virtually superimposed within the first several days of CICU admission, these

curves begin to substantially diverge after a little over a week. This might suggest that the true benefit of a closed unit in the CICU might exist predominantly for patients with protracted critical illness. Perhaps patients that are discharged early from the CICU may be too well to benefit from the comprehensive critical care that may result from a closed ICU model, while those requiring longer stays in the CICU may have a more complex disease phenotype that warrants more structured critical care delivery. While all of this is speculative, it should stimulate additional investigation.

Critical care resource utilization, while variable among CICU models, was certainly substantial in this contemporary cohort of critically ill cardiovascular patients. This mirrors other retrospective and observational series (e.g., Katz et al., 2010; OMalley et al., 2013). While it is unclear if one unit model is more resource-conservative than the other, it is prudent to point out that patients were more quickly discharged with shorter lengths-of-stay in the closed CICU. Although length-of-stay is undoubtedly a complex and often confounded metric of critical care, this observation is nonetheless important. ICUs are exceedingly costly, consuming greater than 20% of total hospital costs despite constituting less than 10% of hospital beds in the US (e.g., Chalfin et al., 1995). Future analyses of models within the contemporary CICU must be able to address these costs-of-care, and resource consumption certainly plays a major role. Other fiscally relevant topics to examine will include professional fees, critical care documentation and billing, and ICU recidivism, among others.

2.4.3 Limitations

Our study has several limitations which merit discussion. First, this is a single institution study from a US, university-based, academic medical center. As a result, our findings may not be generalizable to other healthcare settings or regions. Nonetheless, we believe that this is the most comprehensive and first prospective evaluation of contemporary CICU care delivery, and should therefore represent an important foundation for future study. Second, while attempting to assess the impact of care delivery models on cardiac critical care outcomes, we cannot completely exclude that there may have been temporal changes in

practice patterns that could have influenced our findings. This confounding should have been minimized, at least in part, by the rather short timeline for investigation and the consistency in care team members promoted throughout the study period. Additionally, we must also acknowledge that this study does not specifically address the benefit or liability of having critical care-trained physicians as part of the CICU. This has often been linked with structural studies advocating for closed units in other ICU settings, but was not assessed here. Undoubtedly, given that recommendations for additional training in cardiac intensive care have found their way into recent guidelines and scientific statements (e.g., Morrow et al., 2012; OGara et al., 2015), future study would be prudent in order to better understand the optimal role of cardiac intensivists. Finally, cost of care was not described in this study. Given the enormous expenditures associated with ICU admissions (e.g., Hochman et al., 1999), this will be an important subject for further investigation.

2.4.4 Conclusions

In summary, the contemporary CICU continues to admit complex patients with multiple comorbidities, for diverse cardiovascular critical illnesses, and with a high-expected risk for adverse outcomes. With that being said, there is considerable variability among CICU patients, and this must be better understood in order to develop optimal admission and discharge criteria moving forward. While not associated with an improvement in mortality among those treated in a CICU, the closed model of care resulted in decreased lengths-of-stay. It will be important to utilize these findings in order to continue to develop collaborative research efforts evaluating key components of the CICU. Not only does this have the potential to impact patient care directly, but it may also influence other research efforts. The CICU is often home to many of our contemporary cardiovascular clinical trial participants. It is possible that differences in CICU care delivery may have already affected past trail results, and failure to both understand existing variability and to standardize future practice may lead to additional confounding. Finally, given the differences in care and resources seen in academic and community hospitals across the US, attention should also fo-

cus on how to develop individualized CICUs that can appropriately cater to an institutions specific needs.

2.5 Additional Statistical Analysis

2.5.1 CICU Mortality

We use the logistic regression to model the CICU mortality, which is denoted by p. The model structure indicator (open= 0, closed= 1) and important clinical variables are used as the covariates, including gender, race, age, BMI, APACHE II score, and CICU LOS. The fitted model is

$$\log \frac{p}{1-p} = -5.60 + 0.15 \times \text{APACHE II score} + 0.04 \times \text{LOS in CICU}, \tag{2.1}$$

where covariates are significant at the level of 0.05. Patient's APACHE II score and LOS in CICU have positive effects on the mortality rate. In other words, severer patients and patients who have longer stays in CICU are more likely to die in CICU than other patients. The scatter plot of the mortality rate predicted by Equation (2.1) and the score based mortality rate for each individual patient is depicted in Figure 2.3, where different colors are used to distinguish different unit structures. The mortality rates predicted by Model (2.1) are consistently lower than what predicted by the the APACHE II scoring system. Previously, in Section 2.3.5, we find that actual mortality rates are also lower than what predicted by the the APACHE II scoring system. This finding suggests that it is worthwhile to develop new models for the severity-of-illness score and the corresponding predicted mortality rate for CICU patients in the future study.

2.5.2 Model of Length of stays

We are interested in length of stays (LOSs) within the following three stages: ICU, hospital post-CICU, and hospital since CICU admission. Each LOS is in the unit of one day and is defined as follows:

- LOS in CICU = Date of CICU discharge − Date of CICU admission + 1,

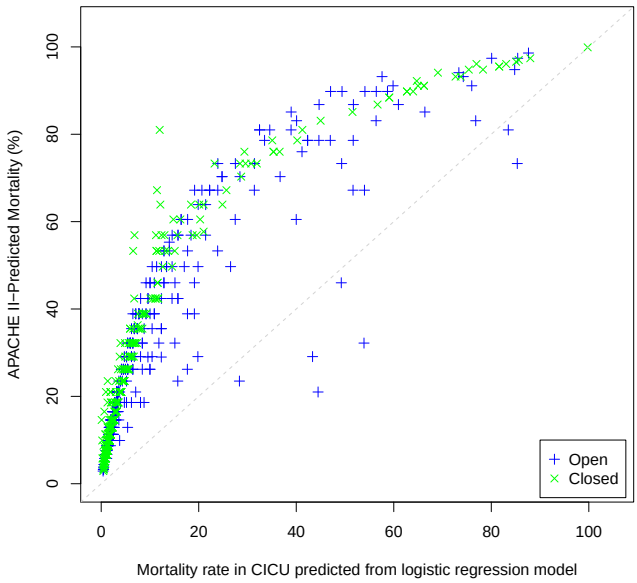


Figure 2.3: Mortality rates predicted by Model (2.1) versus APACHE II score predicted mortality rates

- LOS in hospital post-CICU = Date of hospital discharge − Date of CICU discharge + 1,
- LOS in hospital since CICU admission = Date of hospital discharge − Date of CICU admission + 1.

Note that, a patient, who is admitted to and discharged from the CICU on the same day, has the LOS in CICU as 1 day. We use multiple linear regression to model each of the three stages of LOS. We take the log transformation for each LOS variable. The pool of explanatory variables consists of categorical variables: unit indicator (open= 0, closed= 1), acute heart failure (HF: No= 0, Yes= 1), ST-segment Elevation Myocardial Infarction (STEMI: No= 0, Yes= 1), cardiogenic shock (CS: No= 0, Yes= 1), and cardiac arrest (CA: No= 0, Yes= 1); and continuous variables: age and APACHE II score. We use stepwise selection to choose the best models, which are shown as Models (2.2) - (2.4). Each variable included in the three models is significant at the level of 0.05. From the sign of each significant variable, we can see that: patients who had higher APACHE II scores or suffered the HF and CS, consistently have longer LOSs in each of the three stages; patients who

were implemented with STEMI consistently have shorter LOSs in all three stages; patients who had CA have shorter LOSs in hospital post-CICU and hospital since CICU admission; patients who were admitted in the closed unit have longer LOSs in hospital post-CICU than those who were admitted in the open unit, but not significantly different LOSs in CICU or the LOSs in hospital since CICU admission.

$$\begin{aligned}\log\{\text{LOS in CICU} + 1\} =& 1.054 + 0.011 \times \text{APACHE II} + 0.227 \times \text{HF} + 0.902 \times \text{CS} \\ & - 0.228 \times \text{STEMI}\end{aligned} \tag{2.2}$$

$$\begin{aligned}\log\{\text{LOS in post-CICU} + 1\} =& 1.068 + 0.206 \times \text{Unit} + 0.261 \times \text{HF} + 0.449 \times \text{CS} \\ & - 0.280 \times \text{STEMI} - 0.614 \times \text{CA}\end{aligned} \tag{2.3}$$

$$\begin{aligned}\log\{\text{LOS in hospital} + 1\} =& 1.678 + 0.010 \times \text{APACHE II} + 0.272 \times \text{HF} \\ & + 0.869 \times \text{CS} - 0.334 \times \text{STEMI} - 0.452 \times \text{CA}\end{aligned} \tag{2.4}$$

2.5.3 Subgroup Analysis

In this section, we explain how the subgroup studies visualized in Figure 2.2 are conducted. By comparing the CICU mortality for patients who were admitted into the open and closed units, we do not find any significant difference, as shown in Table 2.4. We doubt that impacts of unit structure on the CICU mortality might be different within finely stratified subgroups (e.g., female versus male). We divide each suspected characteristic into two or three subgroups, the detailed partitions are shown in Table 2.5.

We model the hazard function of the CICU mortality using the Cox regression. For each characteristic, we include the unit indicator (Unit= 0 stands for the open unit, and Unit= 1 stands for the closed unit), the subgroup indicator (Z), and the interaction term of the unit indicator and the subgroup indicator as covariates in the Cox regression. The regression model can be expressed as follows:

$$\lambda_i(t) = \lambda_0(t) \exp\{\beta_1 \text{Unit}_i + \beta_2 Z_i + \beta_3 \text{Unit}_i \times Z_i\}, \tag{2.5}$$

where $\lambda_i(t)$ is the hazard of the ith patient at time t, and $\lambda_0(t)$ is the baseline hazard at time t. Unit_i is the indicator of whether patient i was admitted in the closed ($\text{Unit}_i = 0$) or the open ($\text{Unit}_i = 1$) unit; Z_i is the subgroup value of the study characteristic for patient i. If the interaction term is significant, then we can conclude that impacts of unit structure on the CICU mortality are different for among patient subgroups.

As

$$\frac{\exp\{\beta_1 \times 1 + \beta_2 \times j + \beta_3 \times 1 \times j\}}{\exp\{\beta_1 \times 0 + \beta_2 \times j + \beta_3 \times 0 \times j\}} = \exp\{\beta_1 + \beta_3 j\}, \tag{2.6}$$

$\exp\{\beta_1 + \beta_3 j\}$ is the hazard ratio of the open unit to the closed unit within subgroup $Z_i = j$, listed in Table 2.5. The value of the hazard ratio means that, within the corresponding subgroup, the hazard of the open unit is that number of times of the closed unit. For example, 1.068 in the second row of Table 2.5 means that, for females, the hazard in the open unit is 1.0683 times of that in the closed unit, which implies that the closed unit is better for females than the open unit. Therefore, a hazard ratio that is larger than 1 supports that the closed unit is better for that subgroup of patients. The hazard ratio for each subgroup together with the 95% confidence interval are plotted in Figure 2.2.

Although there are more hazard ratios greater than 1, which support that the closed unit is better than the open unit for more subgroups of patients, neither of them is significant at the level of 0.05, as indicated by p-values listed in the last column of Table 2.5. This finding might be caused by the limited sample size, which is reflected by the long width of each confidence interval shown in Figure 2.2.

Table 2.5: Hazard ratio of the CICU mortality in open versus closed unit within each subgroup

Subgroup	Open	Closed	Hazard ratio	Lower 95%	Upper 95%	p-value
Age						0.575
< 75	29/255	26/266	1.110	0.653	1.887	
≥ 75	14/77	8/72	1.478	0.620	3.526	
Sex						0.582
Female	14/125	18/133	1.068	0.529	2.158	
Male	29/207	16/205	1.397	0.757	2.579	
Race						0.251
White	25/192	21/188	0.991	0.556	1.767	
Black	8/86	11/113	1.046	0.432	2.532	
Others	10/54	2/37	1.793	0.691	4.650	
BMI						0.876
≤ 30	28/198	21/206	1.242	0.712	2.164	
30-35	7/65	5/59	1.157	0.387	3.465	
≥ 35	8/69	8/73	1.141	0.483	2.697	
HF						0.728
No	42/302	30/291	1.159	0.725	1.854	
Yes	1/30	4/47	0.777	0.087	7.018	
STEMI						0.916
No	42/296	33/295	1.193	0.755	1.883	
Yes	1/36	1/43	1.388	0.086	22.310	
CS						0.188
No	32/280	33/303	1.225	0.747	2.010	
Yes	11/52	1/35	5.077	0.653	39.492	
CA						0.101
No	31/314	12/303	2.396	1.228	4.675	
Yes	12/18	22/35	1.055	0.516	2.158	
Sepsis						0.362
No	42/319	31/322	1.290	0.810	2.053	
Yes	1/13	3/16	0.439	0.045	4.248	
MV						0.207
No	12/229	6/245	2.058	0.770	5.503	
Yes	31/103	28/93	1.007	0.603	1.681	
APACHE II						0.200
<22	10/238	7/260	1.658	0.630	4.367	
≥ 22	33/94	27/78	0.810	0.485	1.352	
SAPS II						0.126
<42	11/230	9/268	1.538	0.636	3.718	
≥ 42	32/102	25/70	0.689	0.406	1.169	

CHAPTER 3

Controlling Emergency Department Boarding Times via Active Bed Management

3.1 Introduction

Long waiting times and length of stays in ED are ubiquitous in many parts of the world. In ED, a long waiting time is more than an inconvenience as it can lead to many adverse outcomes including death (Bernstein et al., 2009; Sun et al., 2013). Numerous studies have investigated the reasons behind extended ED stays and waiting times. While there are a number of contributing factors, most of these studies identified long ED boarding times, i.e., the times that ED patients spend waiting to be transferred to an inpatient unit, as one of the primary reasons behind ED crowding (Asplin et al., 2003; Trzeciak and Rivers, 2003; National Center for Health Statistics, 2013). In this chapter, we consider the sub-network consisting of the ED and IW at the daily level, and identify primary factors that affect ED boarding times. To reduce ED boarding times, we investigate the effects of active bed management within IW, through anticipating needs for IW beds and freeing up bed capacity by early discharging IW patients that are medically feasible.

A long boarding time for a patient is an indicator of poor quality of service provided to the patient. It signals the fact that the patient unnecessarily occupies scarce ED space and resources, when in fact these resources could have been used for other patients who are being treated in ED, or waiting for admission to ED. Most often, a long boarding time is a direct consequence of the delay in identifying a bed in the hospital (often IW) to which the patient can be transferred. Thus, as many studies have already concluded, in order to alleviate ED overcrowding problem, focusing on the operations within ED alone is unlikely to be productive; one needs to take a higher system-level view and address the main source

of the patient flow problem: the bed bottleneck in the hospital (e.g., Asplin et al., 2003; Japsen, 2003; Olshaker and Rathlev, 2006; Howell et al., 2008; Hoot and Aronsky, 2008).

To better understand the patient flow problem from a strict bed demand and supply perspective, it is useful to highlight the various ways that the hospital admission can become a bottleneck. First of all, demand for beds on any given day can simply exceed the available bed capacity even after daily patient discharges are completed. Second, even if the number of available beds post-discharges is enough to meet the daily bed demand, the bulk of the bed requests can occur before most of the discharges complete, which results in a temporary bed unavailability during the day. This "misalignment" of bed requests (demand) and patient discharges (supply), which is very common in practice, is one of the main causes of the prolonged ED boarding times.

Third, even when a bed is physically available, admissions to the hospital can still be a bottleneck. This is because when a decision is made to admit a patient, having a bed available in the hospital does not mean that the patient will be transferred right away. For reasons mostly related to the general crowdedness of the hospital, the patient can still have a long boarding time. Specifically, as we demonstrate in this chapter, boarding times are longer when the occupancy level of the hospital (number of patients) is higher. One might think that as long as at least one bed is available when demand for a hospital bed arises, it should not be difficult to identify the empty beds and transfer the patient right away. If so, as long as the occupancy level is below the maximum capacity, there will not be delays in boarding the patients. Unfortunately, that is not quite the case in practice for various reasons. First, the bed assigned to a patient is not arbitrary. One needs to spend effort to assign "the right patient to the right bed", taking into account patient characteristics and different specialty areas within the hospital. Circumstances frequently force the hospitals to be flexible and utilize the aggregate bed capacity. However, if there is no direct match between the patient type and the available bed, which is more likely to happen when the occupancy level is high, it takes longer to determine where exactly the patient should be transferred. Second, while the availability of beds may seem like the hard constraint, what is in fact as critical is the staffing constraint. Unfortunately, higher levels of bed occupancy coincide often with higher levels of workload for the staff. That in turn

means that it would take the staff in any particular unit longer to prepare newly vacated beds for new admissions, longer to respond to a transfer request from ED, and generally be more reluctant to admit new patients, which would further increase their workload. Finally, there are multiple decision makers when it comes to transferring a patient from ED to one of the internal units within the hospital. As a result, hospitals, concerned with both efficiency and fairness, end up adopting rather elaborate patient transfer processes, which tend to take longer when the occupancy level is higher.

Partially motivated by the close relationship between the hospital occupancy level and ED crowding, hospitals are paying more and more attention to utilizing their bed capacities efficiently by adopting different forms of active bed management. As Powell et al. (2012) and Shi et al. (2014) discuss in details, some hospitals focus their efforts on patient discharges, specifically, changing their operations/staffing so that patients are discharged as early in the day as possible and thereby the bed crunch problem typically felt in the middle of the day (i.e., busier periods) is partially averted. Others have been initiating broader efficiency improvement efforts that aim for better bed management. According to reports by Institute of Medicine (2007) and Vicellio et al. (2008) on the state of hospital-based emergency medicine delivery, some hospitals have created "bed czars" or "bed teams" whose job is essentially to achieve the most efficient use of hospital beds in coordination with various units in the hospital. Among the responsibilities of a bed czar is to account for all the inpatient beds and "ensure rapid bed turnaround". Hospital beds are frequently occupied by patients who in fact no longer need the level of services provided for patients in those beds. A report by Audit Commission (2003) finds that the median percentage of such beds to be 5% with some trusts reporting more than 20%. This suggests that there is significant potential for improving bed management without sacrificing the quality of care provided to the patients.

The realization of this potential in practice, however, requires some efforts, which are not cost-free. It might require hiring new staff members whose responsibilities would mainly be efficient management of hospital beds and/or allocating some of the existing staff members to the task of freeing up beds by carrying out tasks that will make it possible to push the discharge times of some of the patients earlier (Howell et al., 2008; Vicellio et al., 2008).

For a severely overcrowded hospital which has a bed bottleneck problem around the clock every day of the week, the cost of putting in a constant effort to identify available beds can be perfectly justified. For many hospitals, however, because of daily fluctuations in patient arrivals and discharges, the severity of the bed bottleneck problem also changes from day to day. For some days, there may be no need to aggressively seek space for new patients, while for some other days, there can be a significant bed shortage. Such hospitals may adopt more of a dynamic policy for allocating resources to the task of creating new bed space. Every day, by taking into account the number of available beds, the number of patients who are expected to be discharged over the next 24 to 48 hours, and the prediction for the number of new bed requests to the internal wards, the hospital can determine whether or not a bed bottleneck problem is likely to occur on that day and if yes how severe it will be. This can then guide decisions regarding how aggressive the hospital needs to be (if at all) when speeding up the discharge process of patients who are already medically safe to be discharged.

The main objective of this study is to develop a framework, which can be used to make this determination and investigate the potential benefits of adopting it. Specifically, the study makes the following three contributions.

1. Using data from an Israeli hospital, we investigate how ED boarding times depend on the hospital occupancy level, the number of new arrivals, and the number of discharges for every day of the week, and quantify this relationship for this particular hospital.

2. We develop two optimization problems whose objective is to determine the smallest number of early discharges necessary in order to meet a particular target level for boarding times. An early discharge may either refer to moving an afternoon discharge to the morning or moving a tomorrow's discharge to today afternoon. Such early discharges can only be done for a selective group of patients who are already medically safe to be discharged. Note that our framework only offers a target for the number of early discharges. The final decision of discharges has to be made by a qualified medical staff. The first problem is meant to be solved by the hospital once every 24 hours and it essentially determines the least the hospital should do to achieve the

target level of boarding time it aims to provide for its patients. The second problem is meant to be solved by the hospital one day in advance to realize an ideal initial occupancy target based on the historical experience.

3. We carry out an extensive simulation study to demonstrate the potential benefits of using either early-discharge approach, in comparison with the baseline scenario under which no early discharge is allowed. We also carry out a sensitivity analysis with respect to the bed request rates to IW.

3.2 Literature Review

There is an extensive literature on methods improving the patient flow from ED to IW. Here we focus on the discharge process for patients in IW. Rubino et al. (2007) and Vicellio et al. (2008) point out the importance of adopting the "discharge by noon" target in freeing up the occupied inpatient beds and improving patient throughput.

Armony et al. (2011) provide broad exploratory data analyses on the patient flow inside ED, IW, and transfer process from ED to IW in the Rambam Hospital, which is also the source of the data supporting our study. In the Rambam Hospital, a patient who is decided to be hospitalized in IW by an ED physician is assigned and transferred to one of the five IWs based on a certain routing policy. Although IW tries to admit patients within four hours from the decision of hospitalization, significantly longer delays exist. Therefore, they propose that reducing the waiting times in ED is essential to prevent the clinical consequences of long delays. They identify a series of causes for the delays, e.g., the ward occupancy, bed capacity, delayed IW discharges and so on. They observe that the longest delay happens in the early morning, and follows with a consistent decline up until noon, due to the fact that the physicians' morning round is performed in the early morning but completed in the afternoon. These findings motivate that it is meaningful to investigate the ways to speed up the discharge process in IW, and convince us about potential subsequent benefits.

Both Powell et al. (2012) and Shi et al. (2014) study the effect of the timing of discharges on the ED boarding times by comparing the performance of different discharge distributions

over a day. Powell et al. (2012) test three alternative discharge polices having the same total volume as the original policy, but different distributions from the original one with peak around 3pm in the model capturing the patient flow into and out of the inpatient beds at the hourly level. The first tested discharge policy is shifting the original discharge distribution 1, 2, 3 and 4 hours earlier; the second one is uniformly discharging 75% patients from 7am to noon and the remaining 25% from noon to 8pm; and the third one is uniformly discharging all patients from 7am to 4pm. For the performance measure, they sum all the number of patients who board during each hour over a day and get the total daily admitted patient boarding hours. All the three alternatives are proved to be able to decrease the total daily admitted patient boarding hours by various degrees.

The original discharge policy in the general wards of the hospital studied by Shi et al. (2014) has a slim and high peak between 2-3pm and only 12.7% of patients are discharged by noon. At one point, this hospital pushes forward the discharge process moderately, which results in an additional peak earlier than the original 2-3pm peak, and around 26% of the patients are discharged before noon. They observe slight reductions in two performance measures: the average boarding times and the fraction of patients whose boarding times are longer than 6 hours. To study the effect of the timing of discharge on relieving the bed problem thoroughly, they propose to push forward the discharge process in a more aggressive way to make that additional peak appear as early as 8-9am. To evaluate the performance of this progressive policy, they simulate the inpatient operations in the real hospital at the hourly level by building a stochastic network. The simulation results show that this more aggressive discharge policy almost stabilizes the two performance measures. We extend the idea in the above two papers, and consider moving tomorrow's discharges to today afternoon, in addition, moving discharges in the afternoon to the morning of the same day.

Crawford et al. (2013) study the effect of the timing of discharges on the ED boarding times by comparing different trigger strategies for letting patients leave earlier than the regular discharge schedule. One static strategy is discharging a patient when her estimated risk of readmission is acceptable; and two proactive strategies are discharging patients when either the percent of patients waiting for an ED bed post triage or the percent of patients

waiting for an inpatient unit bed post treatment in ED is above a specified threshold. They evaluate the performance of various strategies by building a discrete-event simulation model of patient pathway through a hospital that comprises of an ED and several inpatient units at the hourly resolution. Based on the simulation results, they conclude that, compared with the static early discharge strategy, the two proactive ones can significantly reduce ED waiting and boarding times, ambulance diversion duration, and percentage of leave without treatment. In addition, the performance improvements of the proactive early discharge strategies are sensitive to the patient arrival rates.

Besides trying different ways for early discharge and demonstrate the effects as the above literature did, this study goes one step beyond - given the effects of reducing ED boarding times of the various approaches, we start with a pre-specified ED boarding time target, and suggest specific numbers of early discharges in order to achieve that target.

3.3 A Short Description of the Patient Flow

In this chapter, in order to study the delays in the transfer from the ED to IW, we focus on patient flow within the subnetwork consisting of the ED, IW and the transfers of patients from the ED to IW, which is referred to as ED+IW. In the hospital that we are studying, a very high percentage of the patients visiting hospital stay within this subnetwork. Among patients who enter hospital from ED, 13% of those are hospitalized in IW. Switching attention to the IW, 96% of the internal patients come from ED, who are referred to as ED-to-IW patients. The other 4% of the patients in IW are from other units inside hospital, who are referred to as In-to-IW patients. These two incoming streams to IW are shown as red solid and dashed line in Figure 3.1. We will concentrate on ED-to-IW patients and their delays within the transfer process in the rest of the chapter.

When an ED patient is decided to be transferred to hospital inside at the end of the treatment in ED, the request for a bed in IW is sent out and the corresponding routing process is put in. The routing process is implemented by a software called "The Justice Table" in the hospital that we are studying. Once an ED patient is assigned to be hospitalized in IWs, one of the five IWs of the same function is chosen by a Head Nurse. If

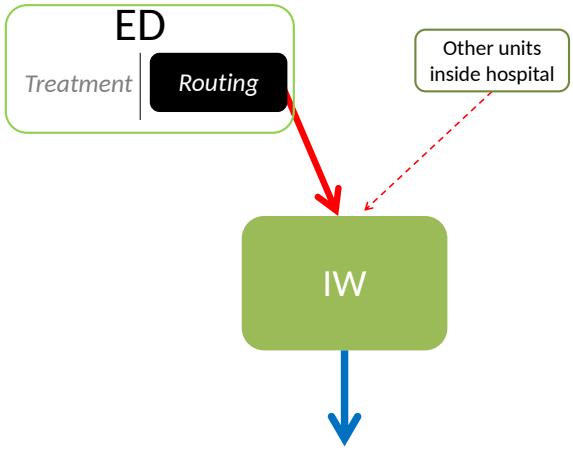


Figure 3.1: Patient Flows within ED+IW Subnetwork

this patient is refused by the first assigned IW, he will be waiting for the reassignment to another IW. When this patient is eventually accepted by an IW, preparation for his arrival begins. In order to complete the transfer, a bed, medical equipment and staff must be ready for the transfer. Even if there is an available (empty) bed in this particular IW, this patient might still suffer delay caused by the lack of either equipment or staff or both. For detailed procedures that happen in the routing process, readers could refer to Armony et al. (2011). Up to the point that all the requirements are ready, this patient will remain in the ED and receive care from ED staff.

As we don't have detailed data regarding the transfer process, we put the routing mechanism inside a "black box", as shown in Figure 3.1, to investigate the transfer delays at the daily level. Of interest is the total amount of time that a transfer patient is held within the "black box". Once the patient spends necessary time inside the "black box", then the transfer is completed and he can enter IW. The total amount of time that an ED-to-IW patient spends in the "black box", from the time of assignment to IW to the time of admission into IW, is often referred to as the ED boarding time in the literature.

Delays in the transfer from the ED to IW should be tightly coupled with IW occupancy level and are affected by operational decisions such as IW admission behaviors and discharge policies. Detailed analyses of the relationships between the ED boarding time and these

impacting factors will be provided in the following sections. Moreover, we propose policies to reduce the prolonged ED boarding time and investigate their benefits.

3.4 ED Boarding Times and Internal Ward Occupancy

In this section, we analyze the department-level patient flow data from the Rambam Hospital, and identify system-level factors that can affect ED boarding times. We first describe in Section 3.4.1 various variables in the hospital data, and how we process them to obtain factors that are potentially related to ED boarding times. Section 3.4.2 then demonstrates the lognormality of the ED boarding times. Finally, Section 3.4.3 builds regression models to characterize how the mean and standard deviation of the lognormal ED boarding time depend on various hospital factors.

3.4.1 Rambam Hospital Data and Processing

The Rambam Hospital is a large hospital in Israel that operates one ED treating on average 247 patients every day, and five IWs hospitalizing on average 1,000 patients every month. The data are provided to us through the courtesy of the Technion SEE Lab. The data include the following time stamps for the ith patient's stay within the ED+IW subnetwork:

- EDR_i — time that an IW bed request is sent out for an ED-to-IW patient;
- INR_i — time that an IW bed request is sent out for an In-to-IW patient;
- ADM_i — time of admission to IW;
- $FIRST_i$ — time that the first procedure is performed in IW;
- DIS_i — time of discharge from IW.

For each individual patient that enters the ED+IW subnetwork on the same day t, we can use the above time stamps to derive the following three daily system-level factors:

- N_t — initial IW occupancy level of Day t, the number of patients whose times of admission to IW (ADM_i) are earlier than 00:00 of Day t and times of discharge from IW (DIS_i) are later than 0:00 of Day t;
- A_t — daily arrivals to IW on Day t, the number of patients whose times of admission to IW (ADM_i) are later than 00:00 of Day t but no later than 23:59 of Day t;
- D_t — daily discharges from IW on Day t, the number of patients whose times of discharge from IW (DIS_i) are later than 00:00 of Day t but no later than 23:59 of Day t.

Summary statistics of these three factors are provided below in Table 3.1.

Table 3.1: Hospital Summary Statistics

Variable	Mean	Standard deviation	First quantile	Median	Third quantile
Initial IW occupancy level (#)	179.6	14.9	170.0	181.0	188.8
Initial IW occupancy level (%)	85.5	7.1	81.0	86.2	89.9
Daily arrivals to IW	33.6	8.0	28.0	34.0	39.8
Daily discharges from IW	33.6	14.1	27.3	37.0	43.0

Notes. $N = 306$ days between October 2006 to October 2007 (excluding the months January to March in 2007, when one of the IWs was in charge of an additional sub-ward).

In addition, for the ith ED-to-IW patient, we use the difference between the time that the first procedure is performed in IW ($FIRST_i$) and the time that an IW bed request is sent out (EDR_i) as the proxy of the ED boarding time of that patient. This can be an overestimation for the actual delay time in the transfer process. However, as shown by Elkin and Rozenberg (2007), a significant portion of this time period is indeed spent within ED, so this is a reasonable estimate for the ED boarding time.

3.4.2 Lognormal ED Boarding Times

Figure 3.2 plots the histogram of all the ED boarding times within the study period (consisting of 306 days), which is clearly right-skewed. An ED-to-IW patient needs to board for 2.9 hours on average until the transfer is completed, while the median boarding time is 2.1 hours. Out of all the ED-to-IW patients, a quarter of them can be boarded within one

hour. The operational goal of the hospital is to board patients within 4 hours (the vertical dash line); however, 21% of the ED-to-IW patients have to wait for more than 4 hours!

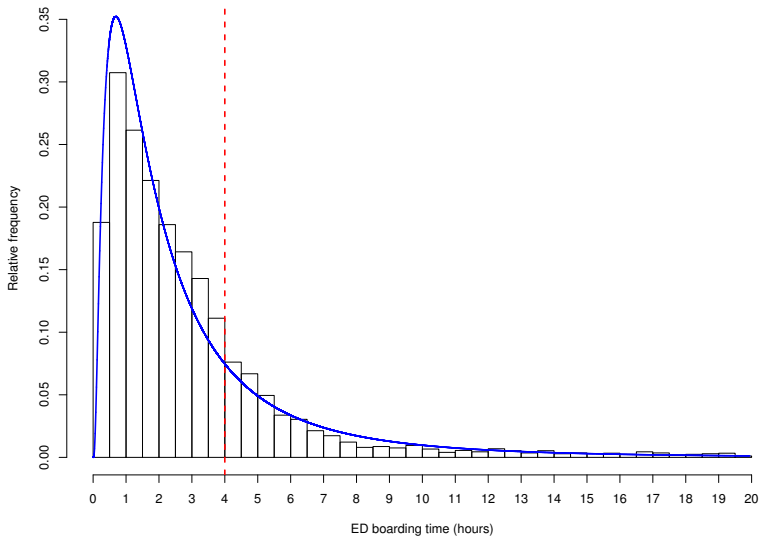


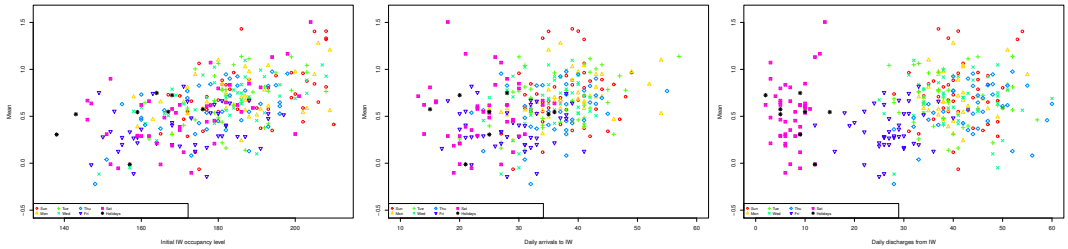
Figure 3.2: Distribution of ED Boarding Time

To aid the hospital decision at the beginning of each day, it helps to know the distribution of ED boarding times within that day. For each day, we perform the Kolmogorov-Smirnov test to check whether the ED boarding times follow a lognormal distribution. Only one day (out of the 306 days) failed the test at the significance level of 0.05. Therefore, the ED boarding times within a day can be well approximated as lognormally distributed. Thompson et al. (2009) have also demonstrated the lognormallity of ED boarding times. In other service systems such as telephone contact centers, Brown et al. (2005) find that lognormal distributions are reasonable to model service times there (i.e. durations of conversations between callers and service representatives.)

3.4.3 System-level Factors and ED Boarding Times

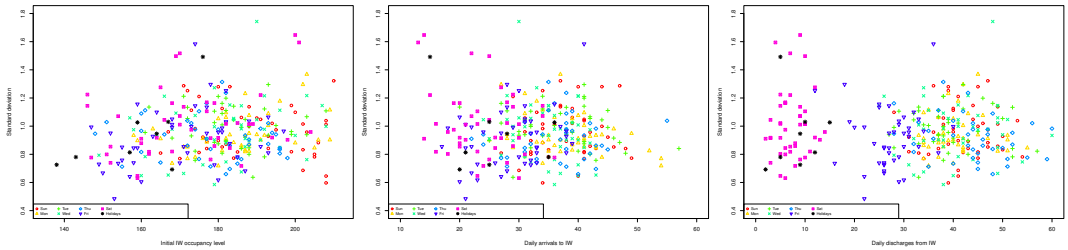
Lognormal distributions are characterized by two parameters: the mean and the standard deviation on the log scale. Let B_t denote a random ED boarding time on Day t; therefore $\log B_t$ is normally distributed. Below we investigate how the system-level factors can affect the mean (m_t) and the standard deviation (s_t) of $\log B_t$, respectively.

Figures 3.3 and 3.4 plot m_t and s_t against the three system-level factors, respectively. To highlight the day-of-the-week effect, the points are coded using different colors and symbols according to their corresponding day of the week as shown in the legend. In addition, Israeli holidays are shown as black asterisks. In both figures, there exists no clear day-of-the-week effect in Panels (a) and (b). Interestingly in Panel (c) of both figures, there are approximately three clusters related to day-of-the-week and holiday. From left to right, the first cluster, with the lowest number of discharges, consists of points from Saturdays and holidays; the second cluster is in the middle and corresponds to Fridays; while the third cluster has the highest number of discharges and includes all weekdays. Note that in Israel, Sunday is the first weekday; Friday is a half-day weekend; and Saturday is a whole-day weekend. Therefore, we conjecture that the variability in the number of discharges is due to lower staffing levels in weekends than those in weekdays.



(a) N_t (b) A_t (c) D_t

Figure 3.3: Mean of Log-transformed ED Boarding Time against: (a) Initial IW Occupancy, (b) Daily Arrivals to IW, and (c) Daily Discharges from IW



(a) N_t (b) A_t (c) D_t

Figure 3.4: Standard Deviation of Log-transformed ED Boarding Time against: (a) Initial IW Occupancy, (b) Daily Arrivals to IW, and (c) Daily Discharges from IW

In Panel (c), the black asterisks (for the 8 holidays listed in Table 3.2) are clustered together with the Saturdays. Staffing level is usually low during major holidays, so this

clustering supports our previous conjecture that the number of discharges is low when the staffing level is low. As shown in Panels (a) and (b), these holidays do not systematically differ from the other days in terms of initial IW occupancy level and number of daily arrivals to IW. Therefore, we label and treat these days as Saturdays when we build regression models for m_t and s_t later on.

Table 3.2: 8 Israeli Holidays within The Study Period

Date	Reason	Date	Reason
Oct. 2, 2006	Yom Kippur	May 23, 2007	Pentecost
Apr. 3, 2007	First day of Passover	Sep. 13, 2007	New Year
Apr. 9, 2007	Last day of Passover	Sep. 27, 2007	Sukkot I
Apr. 24, 2007	Independence Day	Oct. 4, 2007	Shmini Atzeret

Figures 3.3 and 3.4 suggest that we can predict m_t and s_t using IW occupancy level (N_t), number of arrivals to IW (A_t), and number of discharges from IW (D_t). Based on Panel (c) in both figures, we include day-of-the-week indicators for weekdays, Fridays and Saturdays, when modeling the effect of D_t on m_t and s_t. We perform model selection to identify the final model, the parameter estimates of which are provided in Table 3.3 for the mean and the standard deviation of the log-transformed ED boarding time, respectively.

Table 3.3: Effect of System-level Factors on ED Boarding Times

	m_t	s_t
Intercept	-1.4974 (0.1902)***	0.5165 (0.1316)***
Initial IW occupancy (N_t)	0.0101 (0.0010)***	0.0028 (0.0008)***
Arrivals (A_t)	0.0124 (0.0021)***	
Discharge (Weekdays & Friday) (D_t)	-0.0034 (0.0011)**	-0.0019 (0.0007)**
Friday Only	-0.1227 (0.0414)**	
Model fit F-test (Pr> F)	<0.001	<0.001
Adjusted R-squared	0.3721	0.0413

Notes. Standard errors are in parentheses. ** 0.01 significance; *** 0.001 significance.

Based on the above coefficients, the fitted models for m_t and s_t are shown below in Models (3.1) and (3.2), respectively:

$$m_t = \begin{cases} -1.4974 + 0.0101N_t + 0.0124A_t - 0.0034D_t + \epsilon_t^m, & \text{Weekdays,} \\ -1.6202 + 0.0101N_t + 0.0124A_t - 0.0034D_t + \epsilon_t^m, & \text{Friday,} \\ -1.4974 + 0.0101N_t + 0.0124A_t + \epsilon_t^m, & \text{Saturday,} \end{cases} \tag{3.1}$$

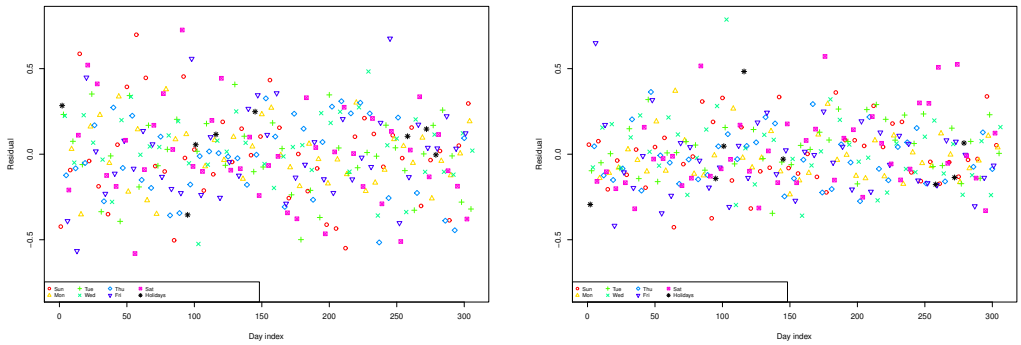
$$s_t = \begin{cases} 0.5165 + 0.0028N_t - 0.0019D_t + \epsilon_t^s, & \text{Weekdays \& Friday,} \\ 0.5165 + 0.0028N_t + \epsilon_t^s, & \text{Saturday.} \end{cases} \tag{3.2}$$

We can make the following observations based on the fitted models. The coefficient estimates for the initial IW occupancy level are significantly positive (0.0101, $p < 0.001$) and (0.0028, $p < 0.001$), respectively, indicating that IW occupancy increases the mean and the standard deviation of the log-transformed ED boarding time. The number of arrivals to IW has a significant positive effect for m_t (0.0124, $p < 0.001$), although it is not significant for s_t. Daily discharge is generally associated with reduction in both the mean and standard deviation of the (log) ED boarding time on Weekdays and Fridays, as indicated by the negative coefficients of the discharge terms in m_t (-0.0034, $p < 0.01$) and in s_t (-0.0019, $p < 0.01$), while discharge on Saturday does not significantly affect neither the mean nor the standard deviation of the log-transformed ED boarding time. The coefficient estimate for the indicator of Fridays in m_t is negative (-0.1227, $p < 0.01$), indicating lower number of discharges on average than other days of a week.

We now perform residual diagnostics on the two fitted models. Figure 3.5 shows the residual plots (ϵ_t^m and ϵ_t^s) with the day-of-the-week highlighted by symbols and colors, revealing randomness among the residuals. The normal quantile plots in Figure 3.6 suggest that the residuals are normally distributed. In addition, we test the independence between ϵ_t^m and ϵ_t^s and find that they are significantly dependent ($p < 0.001$), with a correlation of -0.3. Hence, we can model the residuals using the following bivariate normal distribution:

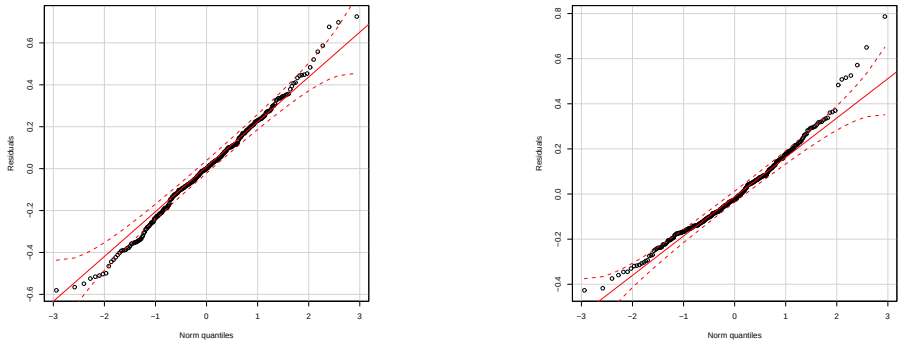
$$\begin{pmatrix} \epsilon_t^m \\ \epsilon_t^s \end{pmatrix} \sim N \left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 0.0560 & -0.0140 \\ -0.0140 & 0.0339 \end{pmatrix} \right].$$

For any given day t, we introduce a day-of-week indicator w_t, where $w_t \in \{1 : \text{Sunday}, \cdots, 7 : \text{Saturday}\}$. At the beginning of any day t, the above fitted Model (3.1) and (3.2) suggest that, given the initial IW occupancy level (N_t), the number of arrivals to IW (A_t), the number of discharges from IW (D_t), and its day-of-week indicator (w_t), we



(a) Model (3.1) (b) Model (3.2)

Figure 3.5: Residual Plots of Model (3.1) and (3.2)



(a) Model (3.1) (b) Model (3.2)

Figure 3.6: Normal Quantile Plots of Residuals of Model (3.1) and (3.2)

can predict the distribution of the log-transformed ED boarding times on Day t as:

$$\log\{B_t(w_t, N_t, A_t, D_t)\} \sim N\Big(m_t(w_t, N_t, A_t, D_t),\, s_t^2(N_t, D_t)\Big), \tag{3.3}$$

where m_t and s_t are specified in Models (3.1) and (3.2). This distribution will be used to guide our proposal for early-discharge policies in Section 3.5.

Models (3.1) and (3.2) assume linear effects of the initial IW occupancy level on the mean and the standard deviation of the log ED boarding time. Below we check whether the effect can be quadratic, which makes sense in that the higher the occupancy level, the bigger the effect on the boarding time. Models (3.4) and (3.5) use the square of the occupancy level (instead of the linear term), with the estimated coefficients listed in Table 3.4. The coefficient estimates for the quadratic term are positive (0.00003, $p < 0.001$) and (0.00001, $p < 0.001$) in both models, respectively. It reflects the fact that when the system is more

congested, the effect of the initial IW occupancy level on the ED boarding time is larger. For the other system-level factors, the coefficient estimates are similar to those in Models (3.1) and (3.2).

Table 3.4: Effect of System-level Factors on ED Boarding Times

	m_t	s_t
Intercept	-0.6173 (0.1088)***	0.7713 (0.0657)***
Initial IW occupancy square (N_t^2)	0.00003 (0.000003)***	0.00001 (0.000002)***
Arrivals (A_t)	0.0125 (0.0021)***	
Discharge (Weekdays & Friday) (D_t)	-0.0033 (0.0011)**	-0.0019 (0.0007)**
Friday Only	-0.1227 (0.0414)**	
Model fit F-test (Pr> F)	<0.001	<0.001
Adjusted R-squared	0.3756	0.0387

Table 3.5: *Notes.* Standard errors are in parentheses. ** 0.01 statistical significance; *** 0.001 statistical significance.

$$m_t = \begin{cases} -0.6173 + 0.00003N_t^2 + 0.0125A_t - 0.0033D_t + \epsilon_t^{mq}, & \text{Weekdays} \\ -1.6201 + 0.00003N_t^2 + 0.0125A_t - 0.0033D_t + \epsilon_t^{mq}, & \text{Friday} \\ -1.4974 + 0.00003N_t^2 + 0.0125A_t + \epsilon_t^{mq}, & \text{Saturday} \end{cases} \tag{3.4}$$

$$s_t = \begin{cases} 0.7713 + 0.00001N_t^2 - 0.0019D_t + \epsilon_t^{sq}, & \text{Weekdays \& Friday} \\ 0.7713 + 0.00001N_t^2 + \epsilon_t^{sq}, & \text{Saturday} \end{cases} \tag{3.5}$$

3.5 Determining the target number for early discharges or internal ward occupancy

In this section, we propose two policies which can be used to guide the hospitals in making decisions regarding how many patients to early-discharge on any given day so as to keep emergency department boarding times at a level that is acceptable to the hospital. More specifically, both policies determine "suggested" target levels (for the number of early discharges or equivalently the number of occupied beds in the internal ward) which the hospital should strive for so as to ensure that the percentage of patients whose boarding

times exceed a particular time limit (e.g. 4 hours) is not more than a particular value (e.g. 20 percent). Policy 1 is dynamic in the sense that the target number of early discharges for each day is determined at the beginning of the day by taking into account the number of patients currently in the internal wards, number of discharges expected on that day and the following day, and the predicted number of new ED arrivals on that day based on the historical data. On the other hand, Policy 2 determines target occupancy levels for the internal ward for each day of the week in advance based on the historical data alone.

We consider two types of early discharges. In most hospitals - including the Rambam hospital, where our data come from - a very high percentage of the patients are discharged from the hospital in the afternoon. Once a bed is vacated, it is not immediately available for admitting new patients since it needs to be cleaned. Therefore, a bed vacated by a patient on a given day can be available for a new patient only late in the afternoon by which time the number of bed requests from the emergency department would have already peaked. In order to prevent this misalignment between the peak of the new bed requests and bed availability in the internal wards, patients could be discharged earlier in the day. This is the first type of early-discharge we consider and we call this *same day early-discharge.* Patients who go through same day early-discharge are discharged on the same day they were regularly scheduled to be discharged but they leave early in the morning. Specifically, we assume that these patients leave early enough that it would be reasonable to assume that the beds they vacate are available the whole day when predicting the boarding time distribution for that day. Admittedly, this is an optimistic assumption. Clearly, in reality, some of the new bed requests will arrive before the early-discharge process of some of the patients is over. However, because the boarding time distribution is estimated for a random patient on a given day, as long as a vast majority of the patients arrive after early discharges are complete (which would happen if patients are discharged by 10 am), this assumption would serve as a good approximation.

On any given day, in addition to the patients whose discharge times can be moved to earlier in the day, patients who are normally going to be discharged tomorrow can also be discharged early. Specifically, we assume that these patients can be early-discharged today instead of tomorrow but they cannot be discharged as early in the day as same day early-

discharges. Their beds become available today but only in the afternoon just like the beds of those patients who are scheduled to be discharged today and are not early discharged. We call this type of early discharge a *one-day ahead early-discharge.* However, we do not consider one-day ahead early-discharge on Saturday based on the findings in Section 3.4.

3.5.1 Policy 1: Daily dynamic determination of the target number for early discharges

Let $\tilde{D}_t^1$ denote the number of patients who will be discharged on Day t at regular discharge times, which we call *regular discharges*, if none of the patients are early discharged on that day (Note that $\tilde{D}_t^1 = D_t$.) Similarly, let $\tilde{D}_t^2$ denote the number of regular discharges for day $t+1$ as of the beginning of Day t. (The actual number of regular discharges on day $t+1$ can be higher if at least one of the patients admitted today stays only one day at the hospital. Therefore, it is possible that $\tilde{D}_t^2 \neq D_{t+1}$). Also let y_t^1 denote the number of same day early discharges and y_t^2 denote the number of one-day ahead early discharges for day t. Then, we must have $y_t^1 \leq \tilde{D}_t^1$ and $y_t^2 \leq \tilde{D}_t^2$.

Recall that the hospital sees N_t patients in the general ward at the beginning of Day t, and in the absence of any early discharges, D_t patients will be discharged from the hospital over the course of the same day at regular discharge times. However, with the early discharges, before the arrival of the bulk of the patients of the day, the number of occupied beds in the hospital will drop to $N_t - y_t^1$ and the number of non-early discharges will be $D_t - y_t^1 + y_t^2$. Note that the effect of one-day ahead early-discharge patients on the boarding times will be like the regular discharge patients. Then, using the results of Section 3.4, the hospital can compute

$$\mathrm{P}\{B_t\left(w_t, N_t - y_t^1, A_t, D_t - y_t^1 + y_t^2\right) > \Theta\}, \tag{3.6}$$

the probability that a randomly chosen patient on Day t will have a boarding time that exceeds Θ. The hospital would like to keep this probability low but also would like to keep the "early discharge cost" low. Thus, the hospital might consider minimizing the total daily discharge cost subject to a constraint on the probability. Specifically, the daily target levels

for early discharges can be determined by solving the following optimization problem:

$$\begin{aligned} \min \quad & f_1 y_t^1 + f_2 y_t^2, \\ \text{s.t.} \quad & \mathrm{P}\left\{ B_t\left(w_t, N_t - y_t^1, A_t, D_t - y_t^1 + y_t^2\right) > \Theta \right\} \leq \alpha, \\ & 0 \leq y_t^1 \leq \tilde{D}_t^1, \\ & 0 \leq y_t^2 \leq \tilde{D}_t^2 \times \mathbf{1}\{w_t \neq 7\}, \end{aligned} \tag{3.7}$$

where α is the predetermined tolerance-level, and $f_1 > 0$ and $f_2 > 0$ are respectively the per patient costs of same day early-discharge and one-day ahead early-discharge. As observed in Section 3.4, number of discharges on Saturdays is extremely lower than all the other days of a week, we believe it is caused by low staffing level on Saturdays. Therefore, it would be appropriate to not consider 1-day-ahead early discharges on Saturdays.

Note that it is difficult to estimate f_1 and f_2 in reality as early discharging a patient might require the involvement of a number of individuals and units within the hospital and changes in the prioritization of certain tasks in the hospital. However, estimation of these cost parameters is not necessary since the solution to the above optimization problem is independent of the precise values of f_1 and f_2. Since early discharging tomorrow's patient today would be more challenging than discharging today's patient earlier in the day, we can assume that $f_1 < f_2$. Then, it is straightforward to show the following result (the proof is immediate and therefore is left in Section 3.8).

Proposition 3.1. *Suppose that $f_1 < f_2$ and for any fixed day t, $\log(B_t)$ has probability cdf $\Phi(\cdot)$ with mean given by $m_t = a_0 + a_1 y_t^1 + a_2 y_t^2$ and standard deviation given by $s_t = b_0 + b_1 y_t^1 + b_2 y_t^2$. Then, if*

1. *$a_2 + \Phi^{-1}(1-\alpha) b_2 < 0$ and $\frac{a_1 + \Phi^{-1}(1-\alpha) b_1}{a_2 + \Phi^{-1}(1-\alpha) b_2} \leq -1$, or*
2. *$a_2 + \Phi^{-1}(1-\alpha) b_2 = 0$,*

we have: if $\exists$ a feasible solution to Problem (3.7) and $\left(y_t^{1*}, y_t^{2*}\right)$ is an optimal solution. Then if $y_t^{2*} > 0$, we have $y_t^{1*} = \tilde{D}_t^1$.

Corollary 1. *For models (3.1) and (3.2) developed for Rambam in Section 3.4, the conditions of Proposition 3.1 hold for days if $\alpha < 0.9632$.*

An implication of Proposition 3.1 is that if there is a feasible solution to Problem (3.7), then an optimal solution can be found by using a simple greedy policy that first increases y_t^1 one by one (and then y_t^2 if y_t^1 hits $\tilde{D}_t^1$ for weekdays and Friday) until the probability constraint is satisfied.

If Problem (3.7) does not have a feasible solution, this would mean that even if the hospital can early discharge all the patients who can be early discharged safely, the service level constraint will not be met. Clearly, however, in such a case the hospital would prefer to early discharge all the patients it can in order to keep the boarding times as low as possible even if the target cannot be met. Thus, Policy 1 can be described as follows.

Description of Policy 1:

At the beginning of every day t (at midnight that marks the beginning of day t), do the following:

Step 1: Set $y_t^1 = \tilde{D}_t^1$ and $y_t^2 = \tilde{D}_t^2$ and compute $\bar{p} = \mathrm{P}\Big\{B_t\left(w_t, N_t - y_t^1, A_t, D_t - y_t^1 + y_t^2\right) > \Theta\Big\}$. If $\bar{p} > \alpha$, set $\delta^1 = y_t^1$, $\delta^2 = y_t^2$, and skip Step 2; otherwise go to Step 2.

Step 2: Set $\delta^1 = y_t^{1*}$ and $\delta^2 = y_t^{2*}$, where (y_t^{1*}, y_t^{2*}) is an optimal solution to Problem (3.7).

Step 3: Set the target level for same-day early discharge to δ^1 and the target level for one-day ahead early discharge to δ^2.

Note that it is possible that $\delta^1 = \delta^2 = 0$ in which case there is no need to early discharge any of the patients.

3.5.2 Policy 1-Var: Variation of Policy 1

Policy 1-Var is a heuristic version of Policy 1. It replaces the probability distributions of IW arrival A_t, ϵ_t^m and ϵ_t^s in the objective function in problem (3.7) by their corresponding mean values. Its solution (y_t^1 and y_t^2) can be expressed in a closed-form, as shown in Expression (3.11) and (3.12) in Section 3.8, so it would be less time consuming. Therefore, Policy 1-Var would be more desirable if it could get comparable solutions as Policy 1.

3.5.3 Policy 2: A look-up table for the internal ward target occupancy level

For implementation purposes, a simpler alternative to Policy 1 would be computing target occupancy level for the hospital in advance based on the historical data rather than determining target early discharge levels every day. One important advantage of having such a static predetermined occupancy level would be that since the staff would be aware of the target levels in advance there would not be last minute "surprises" for the bed teams. Knowing in advance what the target level would be, they can better prepare for early discharges. In this chapter, recognizing the fact that patient demand and staffing levels can differ significantly depending on the day of the week, we will consider setting a different target level depending on the day of the week. It would certainly be even more convenient to have a single target occupancy level that would be valid for every day of the week. However, as we also observed in our simulation study, the benefits of having the target level vary with the day of the week would likely outweigh the additional complexity in most cases.

As we discussed earlier, we can determine the probability distribution for A_w, the number of patient arrivals for weekday w where $w \in \{1, 2, \ldots, 7\}$ indicates a specific day of the week (e.g., Monday through Sunday). (Note that we use this probability distribution in Policy 1 as well.) By analyzing the data, we can also determine the probability distribution for D_w, the number of discharges for day w (in the absence of any early discharges) and for any given midnight occupancy level N, we can compute $\mathrm{P}\Big\{B_w\left(N,\, A_w,\, D_w\right) > \Theta\Big\}$, the probability that a randomly chosen patient on a randomly chosen but specific day of the week w (e.g. a randomly chosen Monday) will have a boarding time that exceeds Θ. The question then is for what values of the occupancy level, this probability will be below α as desired by the hospital. If the probability is a non-decreasing function of the occupancy level N, which we know is the case for the Rambam hospital and likely to be the case for many other hospitals, the problem reduces to finding the largest value for the occupancy level under which the probability is smaller than or equal to α. Thus, the target midnight occupancy level N_w for day of the week $w \in \{1, 2, \cdots, 7\}$ can be determined by solving the

following simple optimization problem:

$$\begin{aligned} \max \quad & N_w, \\ \text{s.t.} \quad & \mathrm{P}\left\{B_w\left(N_w,\, \hat{A}_w,\, \hat{D}_w\right) > \Theta\right\} \leq \alpha. \end{aligned} \tag{3.8}$$

The optimal solution to Problem (3.8), which we call N_w^* will provide the hospital with a target occupancy level the hospital should aim for a given day of the week w based on which the hospital can determine how many patients to early discharge on each day.

Unlike the optimization problem we solve for Policy 1, the solution to Problem 3.8 does not determine how many same-day early discharges and how many one-day ahead early discharges are needed. However, the difference in the way the two types of early discharges affect the occupancy level essentially determines what exactly needs to be done. As we discussed above, a same-day early discharge can be assumed to decrease the occupancy level by the beginning of that day. On the other hand, a one-day ahead early discharge would help reduce the boarding time but the discharge, even though one day earlier, would not happen early enough in the day to have an effect on the occupancy level at the beginning of the day. Therefore, Policy 2 requires that the hospital try to meet the target occupancy level by early discharging the patients who are already going to be discharged that day. However, if that will not be enough to bring the hospital to the target level, Policy 2 still calls for one-day ahead early discharges as many as practically feasible and needed to reach the target level. Specifically, Policy 2 can be described as follows.

Description of Policy 2:

At the beginning of every day t (at midnight that marks the beginning of Day t), do the following:

Step 1: If N_t, the number of patients in the internal wards is less than or equal to $N_{w_t}^*$, the target occupancy level for that day of the week, then set $\delta_1 = \delta_2 = 0$, and skip Step 2; otherwise go to Step 2.

Step 2: If $N_t - D_t \leq N_{w_t}^*$, set $\delta^1 = N_t - N_{w_t}^*$ and $\delta^2 = 0$; otherwise set $\delta^1 = D_t$ and $\delta^2 = min\{N_t - N_{w_t}^* - D_t, \tilde{D}_{t+1}\}$.

Step 3: Set the target level for same-day early discharge to δ^1 and the target level for one-day ahead early discharge to δ^2.

3.5.4 Policy 2-Var: Variation of Policy 2

Policy 2-Var is a variation of Policy 2. Instead of applying different target midnight occupancy levels to different days of a week, it uses a same target level for every day, and the number is just the arithmetic average of N_w^*'s in Policy 2.

3.6 Simulation Study

In this section, we report the results of a simulation study we conducted to investigate the potential benefits of early discharging patients using the two policies we described in Section 3.5. We first describe the simulation model.

3.6.1 Description of the simulation model

The simulation model is not meant to capture the ED or the hospital operations at a very detailed level. The model captures what happens every day in an aggregate manner and it can be seen as a discrete-time model where time proceeds in units of 1 day. The model keeps track of the number of patients in the IW and the remaining length-of-stay (in terms of days) for each patient in the IW. When a new patient is admitted to the IW, a new length-of-stay is generated using the distribution obtained from the historical data. The length-of-stay for a patient can take any value between 1 and L days. We define class i patients as those who will stay i more days in the IW (unless they are early discharged one-day ahead when they are class 1 patients). Note that each patient's class changes everyday. For example, for $i \geq 2$, today's class i patient is tomorrow's class $i-1$ patient. Let X_t^i denote the number of class i patients at the beginning of Day t, and define $\mathrm{X}_t = \{X_t^i : 0 \leq i \leq L-1\}$ as the vector of the number of patients of each class in IW at the beginning of Day t. At the end of every day, the simulation model updates the vector X_t by incorporating new patients, accounting for discharges, and carrying out the necessary class updates as the length-of-stay for each patient needs to be reduced by one. Note that X_t^0

is the number of patients who are scheduled to be discharged on Day t, which is equal to D_t (or $\tilde{D}_t^1$) when implementing Policy 1. Similarly, X_t^1 is the number of patients currently scheduled to be discharged tomorrow given that we are currently on Day t, and is equal to $\tilde{D}_t^2$. The hospital has C beds and therefore $\sum_{i=0}^{L-1} X_t^i \leq C$. (Since the maximum value for the length-of-stay is L, at the beginning of any day, there are no patients with a remaining length-of-stay of L.)

The sequence of events on any given day t is as follows: First, the hospital determines δ^1 and δ^2, the number of same-day and one-day ahead early discharge patients. When making this decision, the hospital knows the occupancy level $N_t = \sum_{i=0}^{L-1} X_t^i$, the probability distribution for $A_{w_t}^1$ and $A_{w_t}^2$, the number of arrivals on that day, and $D_t = X_t^0$ and $\tilde{D}_t^2 = X_t^1$, the maximum number of same-day and one-day ahead early discharges. Then, same-day early discharges leave, the occupancy level is updated to $N_t - \delta^1$, and the scheduled number of afternoon discharges are updated to $D_t - \delta^1 + \delta^2$. The model then generates a_t^1 and a_t^2 the realized values for the number of patients who will arrive on that day and then the boarding time for each patient given the total number of arrivals of the day, updated occupancy level as well as the updated number of afternoon discharges. Necessary updates are done to compute the relevant performance measures (e.g., number of patients whose boarding time exceeds Θ) at the end of the simulation run.

At the end of the day, the vector X_t is updated to determine X_{t+1}. This is done as follows. First, some of the patients may not be accepted to the IW because of the limited bed capacity. Specifically, the total number of admissions to the IW is given by $\min\{a_t^1 + a_t^2, C - N_t + D_t + \delta^2\}$ where the first term is the total number of patient arrivals and the second term is the available number of beds by the end of the day. Thus, the implicit assumption here is that patients wait for an IW bed as long as they will be provided with a bed by the end of the day but they are transferred to a different hospital (or another alternative arrangement is made) if a bed is not going to be available for them some time during the day. For each admitted patient, length-of-stay is generated. For $j \in \{1, 2, \ldots, L\}$, let Z_t^j denote the number of admitted patients with a length-of-stay of j days. (Note that

we must have $\sum_{j=1}^{L} Z_t^j = \min\{a_t^1 + a_t^2, C - N_t + D_t + \delta^2\}$.) Then, X_{t+1} is determined by

$$X_{t+1}^0 = X_t^1 - \delta^2 + Z_t^1,$$

$$X_{t+1}^i = X_t^{i+1} + Z_t^{i+1},\ 1 \le i \le L-2,$$

$$X_{t+1}^{L-1} = Z_t^L.$$

Finally, time is updated to $t+1$ and the same sequence of events are repeated for day $t+1$.

3.6.2 Specification of simulation model parameters

As discussed in previous sections, the IW model is created using data from the Rambam hospital. We try to capture what happens at Rambam in the IW model, in terms of arrivals, occupancy, discharges, and operational goal. We use Poisson distribution to model the probability distribution of each type of arrivals to IW, ED-to-IW and In-to-IW, for each day of a week, respectively. We estimate the mean of each type of arrivals ($p = 1$: ED-to-IW, $p = 2$: In-to-IW) over the same day of a week w ($w \in \{1, \cdots, 7\}$), and denote it as λ_w^p. Table 3.6 lists the estimated values of λ_w^p from the Rambam data. As mentioned before, we use $A_{w_t}^p$ to denote the number of the pth type of arrivals on any given day t, then $A_{w_t}^p \sim \text{Poisson}\,(\lambda_{w_t}^p)$, where w_t is the day-of-the-week indicator of Day t. Further, we use A_{w_t} to denote the total number of arrivals to IW on any given day t. Therefore, $A_{w_t} \sim \text{Poisson}\left(\lambda_{w_t}^1 + \lambda_{w_t}^2\right)$.

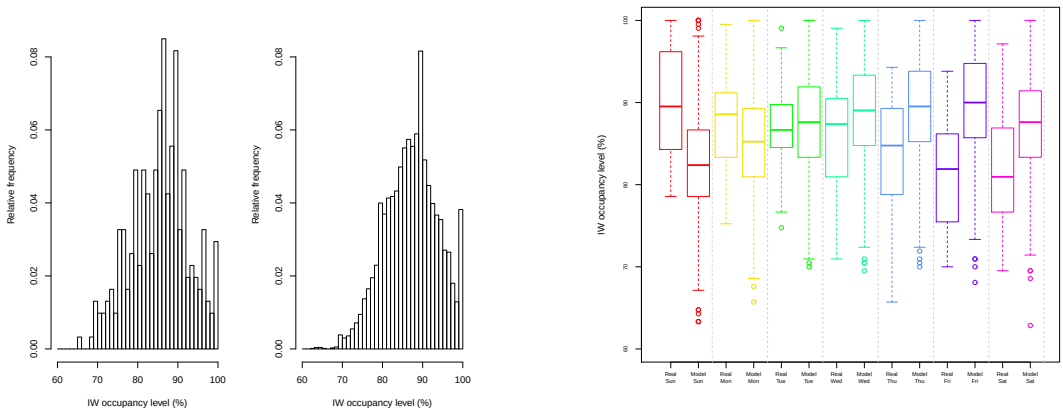
Table 3.6: Estimated Values of λ_w^p, $p \in \{1, 2\}$, $w \in \{1, \cdots, 7\}$

	Sunday ($w = 1$)	Monday ($w = 2$)	Tuesday ($w = 3$)	Wednesday ($w = 4$)	Thursday ($w = 5$)	Friday ($w = 6$)	Saturday ($w = 7$)
ED-to-IW ($p = 1$)	35.41	35.19	33.56	31.90	34.40	26.50	23.08
In-to-IW ($p = 2$)	2.76	2.77	2.73	2.81	2.13	1.69	0.98

To capture the actual IW occupancy in the Rambam hospital, we sample the length-of-stay for each patient who enters the IW model, from the empirical distribution of patient's length-of-stay in the real IW. Besides, the capacity of the IW model is set to be the same as the real IW, 210 beds, which is the maximum number of patients in the IW observed from

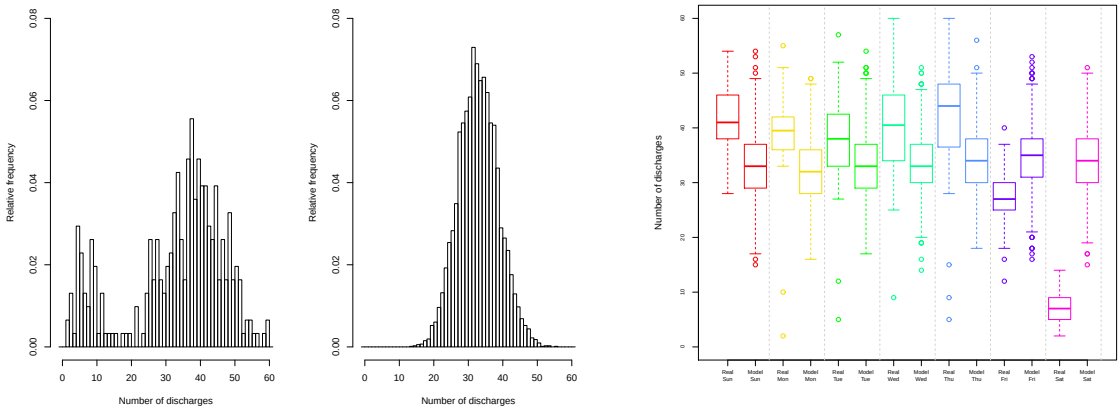
the real hospital data. The service goal in the Rambam hospital is to keep the ED boarding times below 4 hours, therefore the service-levels (Θ) in both the optimization problem (3.7) and (3.8) are set to be 4 hours for all the simulation runs. Intuitively, the smaller the tolerance-level (α) is specified, the stricter the policy is and the more early discharges will be resulted in. In order to verify this trend, we set a series of tolerance-levels.

Simulation results under the baseline scenario, when no early-discharge policy is implemented, verify that the IW model could capture the IW arrivals and occupancy level (shown in Figure 3.7) in the real hospital very well. However, from Figure 3.8, we can see that the IW discharge pattern differs from what happens in the real hospital, especially on Saturdays. Panel (a) compares the histogram of daily number of discharges from IW between the real hospital (left) and IW model (right). It is obvious that there are two modes in the real data, but there is only one in the IW model. The left mode in the real data is Saturdays, the whole-day weekend in Israel, and corresponds to the second pink boxplot from the right of Panel (b). Panel (b) shows that the IW model underestimates the number of discharge during weekdays, while overestimates it during weekends (Friday (purple) and Saturday (pink)). If one only look at the IW discharge in the real hospital, day-of-week effect is easily to be identified: weekends are lower than weekdays. We conjecture that the low level of discharges on weekends are due to low staffing level; however, the lack of data on staffing makes it difficult to demonstrate this conjecture.



(a) Histogram (b) Boxplot

Figure 3.7: Comparison of IW Occupancy Level (%) between Real Hospital and IW Model in Terms of: (a) Histogram, and (b) Boxplot for Each Day-of-week



(a) Histogram (b) Boxplot

Figure 3.8: Comparison of IW Discharges between Real Hospital and IW Model in Terms of: (a) Histogram, and (b) Boxplot for Each Day-of-week

The IW model is simulated as an non-terminating system. Without loss of generality, the first day in each simulation run is set as Sunday. Each run is initialized with the same non-empty stage X_0. We simulate 8008 days consecutively, of which the first 728 days are deleted for the warm-up reason, and the remaining 7280 days are divided into 20 non-overlapping batches of equal length of 364 days (52 weeks). The warm-up period is sufficiently long to populate the system with enough occupancy levels.

3.6.3 Results of the simulation Study

Given the relationship between patient boarding times and the occupancy level of the internal wards, which we have established in Section 3.4, it would be reasonable to expect that any of the early-discharge policies we proposed in Section 3.5 can help in reducing the overall patient boarding times. Our goal in this section is to investigate how much improvement the hospital would get by adopting one of these four policies and which one works best. Making this assessment, however, is not straightforward. A simple comparison of what fraction of the patients exceeds the predetermined level would be misleading and unfair because one also needs to consider the "cost", i.e., the average number of daily early discharges needed, under each policy. After all, it is clear that the policy that achieves the best possible boarding time performance will be one that early discharges all the patients who can be feasibly early discharged. However, such a policy might be undesirable or

even not implementable since that would require early discharging more patients than the hospital can handle.

Another issue to consider is the effect of early-discharges on the daily number of new admissions. By early discharging patients one-day ahead, in addition to helping reduce the boarding times, the hospital also creates additional space for new patients. When the hospital operates at full or close to full capacity, this additional space would lead the hospital to accept more patients than it would in the absence of early discharges, which might in turn have the unintended consequence of increasing boarding times. Of course, the hospital might welcome more patients considering its societal and financial benefits but again such benefit should be assessed in comparison with the "cost" of early discharges and possible increase in boarding times. In short, proper assessment of the different policies we propose requires simultaneous consideration of a number of related performance measures. Here, we will be mainly focusing on three: fraction of patients whose boarding times exceed the predetermined level of 4 hours, average number of early discharges per day, average number of admissions per day.

We first study how the four early-discharge policies perform in comparison with the baseline scenario of having no early discharges. Figure 3.9 demonstrates the fraction of patients whose boarding times exceed the predetermined level of 4 hours for prespecified tolerance levels under different policies. We construct 95% confidence interval (CI) by using the 20 batch means from each simulation run. The gray area on the top of Figure 3.9 is the 95% CI of the fractions under the baseline scenario of having no early discharges in the simulation model, and it is used to tell whether early-discharge policies can help in reducing the overall patient boarding times.

For each discharge policy, we start to set tolerance-level (α) from 0.23, the mean fraction under the baseline scenario, and lower by a constant difference of 0.01 to 0.14. These 11 tolerance-levels are indicated by the small red bars, and labeled on the x-axis as well. We run each discharge policy with these 11 tolerance-levels and construct 95% CI of the fractions, distinguished by different colors as labeled in the legend. The decreasing trend, which exhibits in all discharge policies, supports the previous conjecture that the smaller the tolerance-level (α) is specified, the stricter the policy is and the more early discharges will be

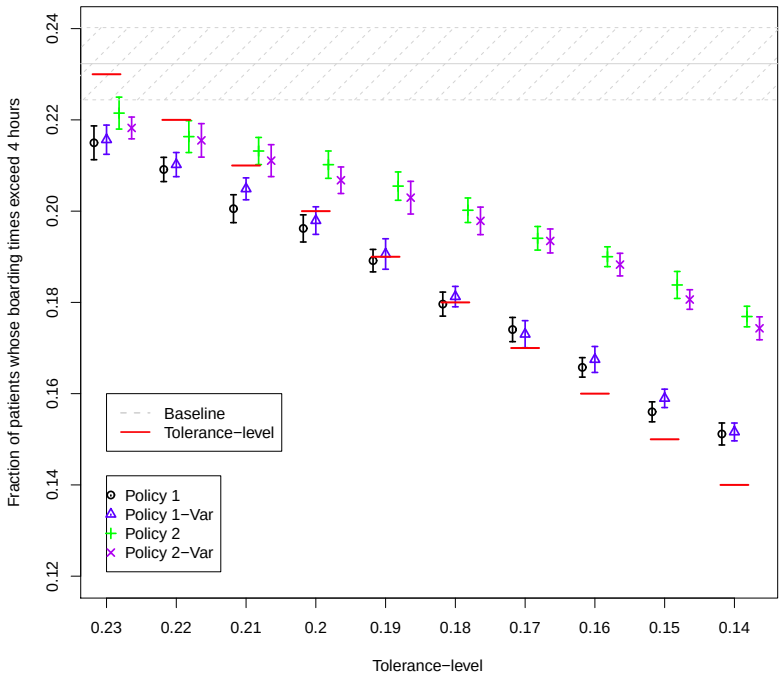


Figure 3.9: 95% Confidence Intervals of Fractions of Patients Whose Boarding Times Exceed the Predetermined Level of 4 Hours Under Different Policies

resulted in. If one looks into each group of four policies with the same tolerance-level, it is easy to notice that Policy 1 always achieves the lowest fractions; Policy 1-Var is comparable to Policy 1; Policy 2 and Policy 2-Var are comparable but with higher fractions. If one looks from group to group, it is easy to notice that the performance of discharge policies firstly beat, then hit, lastly miss the prespecified target when the tolerance-levels decrease.

The comparable performance of Policy 1 and Policy 1-Var makes us comfortable to use Policy 1-Var instead of Policy 1 when computing time is more constraint than computing accuracy. From Figure 3.9 solely, we can also see that Policy 1 and Policy 1-Var are superior to both Policy 2 and Policy 2-Var in regards to boarding time performance. This is not a surprising outcome (although not guaranteed) since unlike Policy 2 and its variation, Policy 1 and its variation take into account of changing occupancy levels day-to-day and thus are attuned to meeting the target boarding time performance. Next, comparisons will be extended by balancing the benefit and the "cost" of early discharges regarding the following

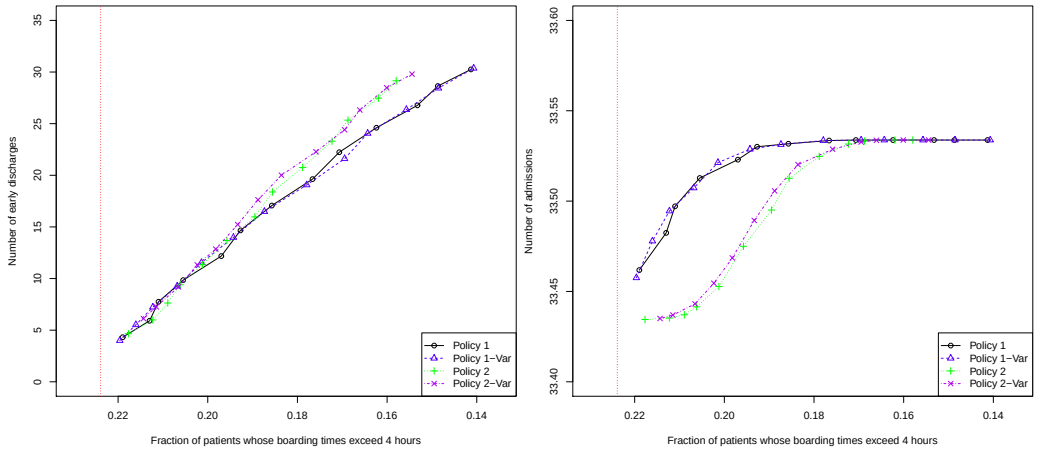
two aspects: average number of early discharges per day and average number of admissions per day.

Panel (a) of Figure 3.10 plots the fraction of patients whose boarding times exceed 4 hours versus the average number of early discharges over each simulation run for different tolerance-levels under different early-discharge polices. Tolerance-levels are chosen from $\{0.27, 0.26, \cdots, 0.14\}$ for Policy 1 and Policy 1-Var, and $\{0.24, 0.23, \cdots, 0.09\}$ for Policy 2 and Policy 2-Var, to make the boarding time performance comparable. The uptrend convinces that better boarding time performance is achieved at the cost of more early discharges. It is clear that the policy that achieves the same boarding time performance by early discharging the smallest number of patients will be the best. Therefore, we eliminate scenarios with worse boarding time performance (larger fractions) and larger number of early discharges. Panel (c) of Figure 3.10 shows the scenarios that cannot be removed based on previous eliminating rule. It is clear that no policy can dominate others consistently; while Policy 1 and Policy 1-Var perform better when targets are stricter.

Panel (b) of Figure 3.10 visualizes the effect of early discharges on the daily number of new admission. The uptrend illustrates that additional space for admitting new patients are created by early discharging patients one-day ahead. It is not surprising that the policy (black dot at the upper right corner) with the largest early discharges realizes the largest number of new admissions. The clear differences between Policy 1 and its variation and Policy 2 and its variation at the moderate levels of target tell us that Policy 1 and its variation are more sensitive to one-day ahead early discharges than the other two. As the display of CIs in Figure 3.9, we find similar behavior patterns between Policy 1 and its variation, and Policy 2 and its variation, respectively.

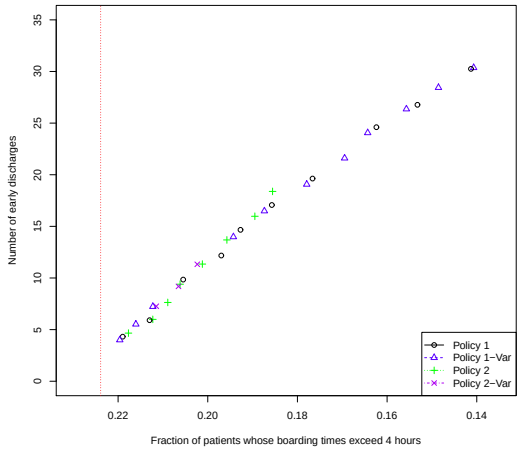
3.6.4 Changing the Arrival Rates

In this section, we analyze the sensitivity of four early-discharge policies to the changes in IW arrival rates. Here, we scale the Poisson arrival rates down and up from the original rates by 10% and 20%, respectively. Intuitively, when more new patients need service in



(a) Early Discharge

(b) Admission



(c) Early Discharge (Dominated)

Figure 3.10: Fraction of Patients Whose Boarding Times Exceed 4 Hours Against: (a) Average Number of Early Discharges per Day, (b) Average Number of Admissions per Day, and (c) Average Number of Early Discharges per Day for Only Dominating Scenarios

IW, the overall ED boarding times should be increased. Figure 3.11 provides evidence to support the effect of arrival rates on the ED boarding times under the baseline scenario.

In the case of each arrival rate, we compare four discharge policies under a series of tolerance-levels, in terms of three aspects same as before: fraction of patients whose boarding times exceed the predetermined level of 4 hours, average number of early discharges per day, average number of admissions per day. From top to bottom, Figure 3.12 shows the trade-off under the scenario with 20% lower, 10% lower, the same as, 10% higher, and 20% higher than the original arrival rate. Two figures in the same row correspond to the same

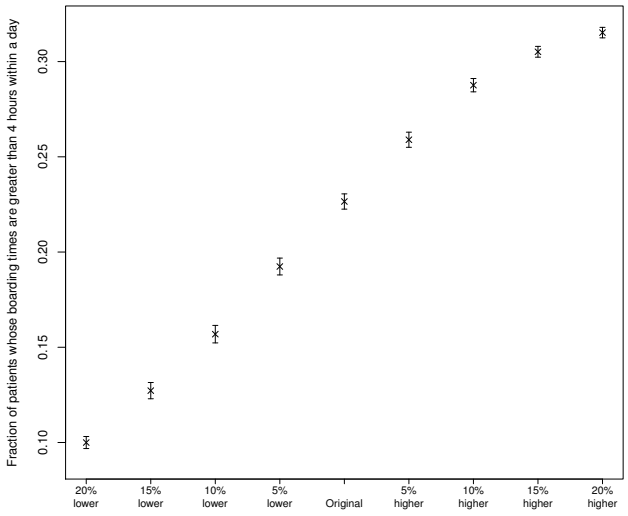
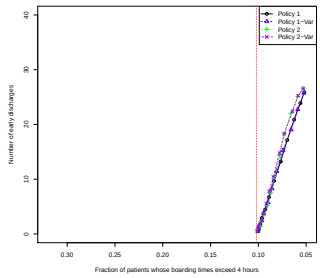


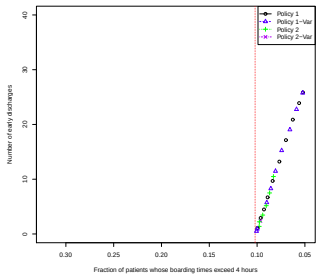
Figure 3.11: 95% Confidence Interval of the Fractions of Patients Whose Boarding Times Exceed 4 Hours Within a Day in the IW Model Under the Baseline Scenario for Different Arrival Rates

arrival rate scenario: the left one is the plot of fraction of patients whose boarding times exceed 4 hours versus average number of early discharges per day; and the right one is the plot of fraction of patients whose boarding times exceed 4 hours versus average number of admissions per day.

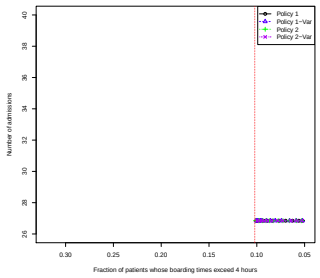
In most cases, we can still observe that the more the early discharges, the better the ED boarding time performance. While the early discharge benefit seems to be violated when the arrival rate is 20% more than that of the real hospital (on the lower left of Figure 3.12), and under the situation with stricter targets (smaller tolerance-levels). After examining the new admissions on the right hand side, we realize that benefit might be diluted by more new admissions. It is still the case that the policy that achieves the same boarding time performance with less early discharges and more new admissions will be the best. Similarly to the scenarios with the original arrival rates, we cannot distinguish one policy from the pool with the best performance under this criterion.



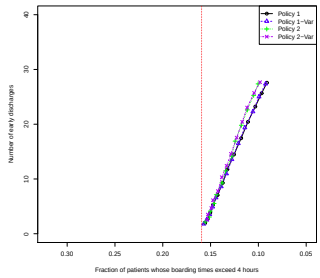
(a.1) 20% Lower, Early Discharge



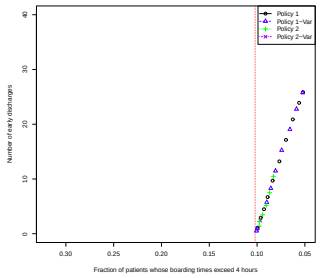
(a.2) 20% Lower, Early Discharge (Dominated)



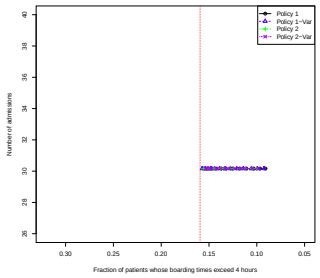
(a.3) 20% Lower, Admission



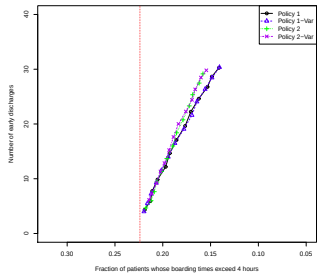
(b.1) 10% Lower, Early Discharge



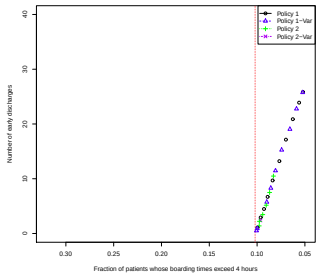
(b.2) 10% Lower, Early Discharge (Dominated)



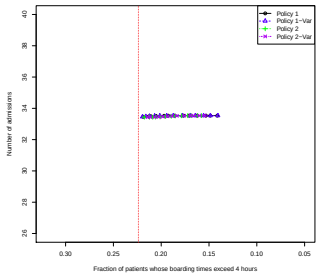
(b.3) 10% Lower, Admission



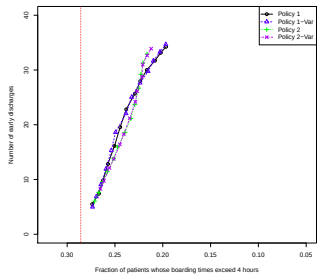
(c.1) Original, Early Discharge



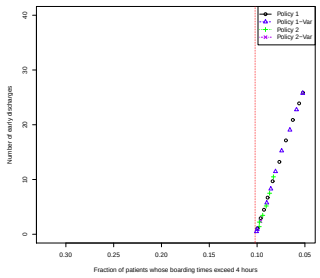
(c.2) Original, Early Discharge (Dominated)



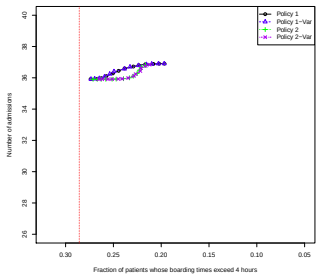
(c.3) Original, Admission



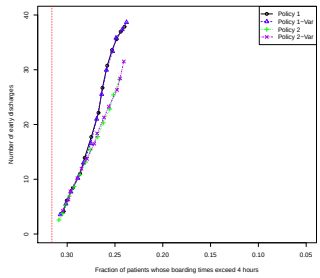
(d.1) 10% Higher, Early Discharge



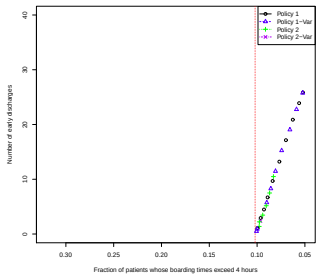
(d.2) 10% Higher, Early Discharge (Dominated)



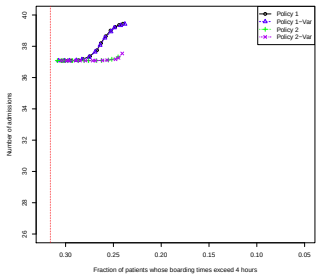
(d.3) 10% Higher, Admission



(e.1) 20% Higher, Early Discharge



(e.2) 20% Higher, Early Discharge (Dominated)



(e.3) 20% Higher, Admission

Figure 3.12: Trade-off Plots for Different Arrival Rates

3.7 Conclusion and Discussion

In this chapter, we identify three leading factors to affect the ED boarding time on the daily operating level: the initial IW occupancy, the number of arrivals to IW, and the number of discharges from IW. We verify that the ED boarding times within a day in our studying hospital follow a lognormal distribution, and in addition, use linear regression to quantify the relationship between the mean and standard deviation of the log-transformed ED boarding time and the three leading factors, respectively. The quantified mathematical relationships are used to guide the later decision making.

We propose two options of early discharge: one is to discharge patients who are expected to be discharged from IW by the end of today leave before the new bulk of arrivals (before noon), and the other is to discharge patients who are expected to be discharged from IW tomorrow by the end of today. To implement these two options, we propose two early discharge policies: one is to dynamically determine the target number of each type of early discharges every day by solving a optimization problem, and the other is to create a look-up table listing the ideal occupancy to start the IW service for each day of the week in advance. The two policies have the same quality of service target: to control the probability of a randomly chosen patient on any day with a boarding time exceeding the target service-level to be not above the tolerance-level in the mind of the hospital administrators. Both policies aim to give the exact discharge plan to the hospital administrators, given that they have a clear target/goal in their mind.

We build a discrete-event simulation model to capture the daily inpatient flow dynamics from ED to IW. Its rationality has been validated. Then, we conduct a series of studies using the simulation model. First, we check how the ED boarding performance of each policy changes with respect to the choice of the tolerance-level. We verify that the early discharge is able to reduce the ED boarding time, in addition, the more early discharges are conducted, the lower the ED boarding time can be achieved. The early discharge is not cost-free, for it needs some extra efforts. Then the policy that achieves the same level of service performance while early discharging less patients is better. We demonstrate that Policy 1 is better than Policy 2. To check the robustness of the better performance of

Policy 1, we conduct a series of sensitivity analyses. The better performance of Policy 1 is consistently observed when the IW arrival rate is either increased or decreased up to 20% of the origin.

Though the conclusions come from the simulation model built upon an extensive empirical study of a single hospital, we believe that similar results can be found in other hospitals based on the similarity in many empirical observations between them. These proposed policies can be generalized to other hospitals as long as certain requirements are satisfied (Proposition 3.1); besides, these insights can help hospital managers choose among different policies to implement. In addition, based on the benefits and costs of the implementation, hospital administrators can choose the desired service-level and tolerance limit to implement these policies.

Our study has limitations in several aspects. First, the effectiveness of the proposed policies is evaluated based on predictions from a single hospital. Thus, our findings might not always be generalizable to other hospital settings, e.g., those do not satisfy the conditions of Proposition 3.1. Second, we recognize that it is very challenging to implement either the "same-day" or the "one-day ahead" early discharge in practice. Because discharging patients by noon is difficult as physicians and nurses are busy with the morning rounds, in addition, early discharging patients one-day ahead might cause deteriorations further readmissions to IW. Last but not the least, discharging more patients, in turn admitting more new patients as more vacant beds are available, might overload the IW. These would require coordination throughout the entire hospital, and additional staffing and resources to be added according to the specific discharge plan of that day. Though the number of early discharges given by the proposed policies may not be completely practical, we believe it can serve as a goal for hospital managers to aim at if they intend to eliminate excessively long boarding times within that day. In addition, our model could help hospital managers estimate the benefits on reducing boarding times gained from certain numbers of two types of early discharges.

3.8 Proof of Proposition (3.1)

Proof. Based on the second and third constraint of problem (3.7), we know that the possible feasible region is inside or on the boundary of the following rectangle:

$$\Big\{ \big(y_t^1, y_t^2\big) : 0 \le y_t^1 \le \tilde{D}_t^1, 0 \le y_t^2 \le \tilde{D}_t^2 \Big\}. \tag{3.9}$$

Therefore, Proposition 3.1 could be rephrased as the optimal solution, if there exits one, is on the lower-right boundary of the above rectangle:

$$\Big\{ \big(y_t^1, y_t^2\big) : 0 \le y_t^1 \le \tilde{D}_t^1, y_t^2 = 0 \Big\} \bigcup \Big\{ \big(y_t^1, y_t^2\big) : y_t^1 = \tilde{D}_t^1, 0 \le y_t^2 \le \tilde{D}_t^2 \Big\}. \tag{3.10}$$

To make the proof clearly, we rewrite the models for mean and standard deviation of log-transformed ED boarding time as:

$$m_t \;=\; a_0 + a_1 y_t^1 + a_2 y_t^2,$$

and

$$s_t \;=\; b_0 + b_1 y_t^1 + b_2 y_t^2,$$

where $a_0 = \beta_0^m + \beta_1^m N_t + \beta_2^m A_t + \beta_3^m D_t + \beta_4^m w_t + \epsilon_t^m$, $a_1 = -\big(\beta_1^m + \beta_3^m\big)$, $a_2 = \beta_3^m$, $b_0 = \beta_0^s + \beta_1^s N_t + \beta_2^s A_t + \beta_3^s D_t + \beta_4^s w_t + \epsilon_t^s$, $b_1 = -\left(\beta_1^s + \beta_3^s\right)$, $b_2 = \beta_3^s$. A_t, ϵ_t^m, and ϵ_t^s are random variables. If we can prove that for any given $A_t = a$, $\epsilon_t^m = e$ and $\epsilon_t^s = f$, $\mathrm{P}\Big\{B_t\left(w_t, N_t - y_t^1, a, D_t - y_t^1 + y_t^2\right) > \Theta\Big\} \le \alpha$, then $\mathrm{P}\Big\{B_t\left(w_t, N_t - y_t^1, A_t, D_t - y_t^1 + y_t^2\right) > \Theta\Big\} \le \alpha$.

Step 1: Find the feasible region.

$$
\begin{aligned}
&\mathrm{P}\left\{B_t\left(w_t, N_t - y_t^1, a, D_t - y_t^1 + y_t^2\right) > \Theta\right\} \leq \alpha \\
\Longleftrightarrow\ &\mathrm{P}\left\{\log B_t\left(w_t, N_t - y_t^1, a, D_t - y_t^1 + y_t^2\right) > \log\Theta\right\} \leq \alpha \\
\Longleftrightarrow\ &\mathrm{P}\left\{\mathrm{Z} > \frac{\log\Theta - m_t}{s_t}\right\} \leq \alpha \\
\Longleftrightarrow\ &\mathrm{P}\left\{\mathrm{Z} \leq \frac{\log\Theta - m_t}{s_t}\right\} \geq 1-\alpha \\
\Longleftrightarrow\ &\frac{\log\Theta - m_t}{s_t} \geq \Phi^{-1}(1-\alpha) \\
\Longleftrightarrow\ &\log\Theta - m_t \geq \Phi^{-1}(1-\alpha)\, s_t \\
\Longleftrightarrow\ &\log\Theta - \left(a_0 + a_1 y_t^1 + a_2 y_t^2\right) \geq \Phi^{-1}(1-\alpha)\left(b_0 + b_1 y_t^1 + b_2 y_t^2\right) \\
\Longleftrightarrow\ &\left(a_2 + \Phi^{-1}(1-\alpha)\, b_2\right) y_t^2 \leq -\left(a_1 + \Phi^{-1}(1-\alpha)\, b_1\right) y_t^1 + \left(\log\Theta - a_0 - \Phi^{-1}(1-\alpha)\, b_0\right)
\end{aligned}
$$

1. When $a_2 + \Phi^{-1}(1-\alpha)\, b_2 = 0$, then the feasible region is

$$0 \leq -\left(a_1 + \Phi^{-1}(1-\alpha)\, b_1\right) y_t^1 + \left(\log\Theta - a_0 - \Phi^{-1}(1-\alpha)\, b_0\right).$$

2. When $a_2 + \Phi^{-1}(1-\alpha)\, b_2 < 0$, then the feasible region is

$$y_t^2 \geq \underbrace{-\frac{a_1 + \Phi^{-1}(1-\alpha)\, b_1}{a_2 + \Phi^{-1}(1-\alpha)\, b_2}}_{\triangleq g} y_t^1 + \underbrace{\frac{\log\Theta - a_0 - \Phi^{-1}(1-\alpha)\, b_0}{a_2 + \Phi^{-1}(1-\alpha)\, b_2}}_{\triangleq h}.$$

Step 2: We are to prove that the optimal solution is on the lower-right boundary (3.10) under certain circumstance for each of the above two cases respectively.

1. When $a_2 + \Phi^{-1}(1-\alpha)\, b_2 = 0$,

$$y_t^1 \geq \frac{\log\Theta - a_0 - \Phi^{-1}(1-\alpha)\, b_0}{a_1 + \Phi^{-1}(1-\alpha)\, b_1}$$

2. When $a_2 + \Phi^{-1}(1-\alpha)\, b_2 < 0$ and $g \leq -1$.

   (a) Assume $\left(y_t^{1*}, 0\right)$ is the first point on the lower-right boundary that satisfies the constraints ($0 \leq y_t^{1*} \leq \tilde{D}_t^1$), then we are to prove that no point under $y_t^1 + y_t^2 = y_t^{1*}$

satisfies the constraints. Otherwise, assume we have $\left(y_t^{1\prime}, y_t^{2\prime}\right)$ such that $y_t^{1\prime} + y_t^{2\prime} \leq y_t^{1*} - 1$ and satisfying the constraints, namely $y_t^{2\prime} \geq g y_t^{1\prime} + h$. Then,

$$0 \geq g y_t^{1\prime} - y_t^{2\prime} + h \geq g y_t^{1\prime} + g y_t^{2\prime} + h \geq g\left(y_t^{1\prime} + y_t^{2\prime}\right) + h.$$

That is to say that $\left(y_t^{1\prime} + y_t^{2\prime}, 0\right)$ also satisfies the constraints, while $y_t^{1\prime} + y_t^{2\prime} \leq y_t^{1*} - 1 < y_t^{1*}$, so it is a contradiction.

(b) Assume $\left(\tilde{D}_t^1, y_t^{2*}\right)$ is the first point on the lower-right boundary that satisfies the constraints ($0 < y_t^{2*} \leq \tilde{D}_t^2$), then we are to prove that no point under $y_t^1 + y_t^2 = \tilde{D}_t^1 + y_t^{2*}$ satisfies the constraints. Otherwise, assume we have $\left(y_t^{1\prime}, y_t^{2\prime}\right)$ such that $y_t^{1\prime} + y_t^{2\prime} \leq \tilde{D}_t^1 + y_t^{2*} - 1$ and satisfying the constraints, namely $y_t^{2\prime} \geq g y_t^{1\prime} + h$. Then,

i. if $y_t^{1\prime} + y_t^{2\prime} \geq \tilde{D}_t^1$,

$$y_t^{1\prime} + y_t^{2\prime} - \tilde{D}_t^1 \geq y_t^{1\prime} + g y_t^{1\prime} + h - \tilde{D}_t^1 - g\tilde{D}_t^1 + g\tilde{D}_t^1 = (1+g)\left(y_t^{1\prime} - \tilde{D}_t^1\right) + g\tilde{D}_t^1 + h \geq g\tilde{D}_t^1 + h,$$

that is to say $\left(\tilde{D}_t^1, y_t^{1\prime} + y_t^{2\prime} - \tilde{D}_t^1\right)$ also satisfies the constraints, while $y_t^{1\prime} + y_t^{2\prime} - \tilde{D}_t^1 \leq y_t^{2*} - 1 \leq y_t^{2*}$, so it is a contradiction;

ii. if $y_t^{1\prime} + y_t^{2\prime} < \tilde{D}_t^1$,

$$0 \geq g y_t^{1\prime} - y_t^{2\prime} + h \geq g y_t^{1\prime} + g y_t^{2\prime} + h \geq g\left(y_t^{1\prime} + y_t^{2\prime}\right) + h,$$

that is to say $\left(y_t^{1\prime} + y_t^{2\prime}, 0\right)$ also satisfies the constraints, so it is a contradiction.

Solution: We can come up with a closed-form solution to the optimal problem to (3.7), when we replace A_t, ϵ_t^m and ϵ_t^s by a fixed value (e.g., mean value), respectively.

1. If $a_2 + \Phi^{-1}(1-\alpha) b_2 < 0 \,\&\, g \leq -1 \,\&\, \frac{\log\Theta - a_0 - \Phi^{-1}(1-\alpha) b_0}{a_1 + \Phi^{-1}(1-\alpha) b_1} \leq \tilde{D}_t^1$, or $a_2 + \Phi^{-1}(1-\alpha) b_2 = 0$,

$$\left(y_t^1,\ y_t^2\right) = \left(\max\left\{0, \lceil \frac{\log\Theta - a_0 - \Phi^{-1}(1-\alpha) b_0}{a_1 + \Phi^{-1}(1-\alpha) b_1} \rceil\right\},\ 0\right); \tag{3.11}$$

2. if $a_2 + \Phi^{-1}(1-\alpha) b_2 < 0 \,\&\, g \leq -1 \,\&\, \frac{\log\Theta - a_0 - \Phi^{-1}(1-\alpha) b_0}{a_1 + \Phi^{-1}(1-\alpha) b_1} > \tilde{D}_t^1$,

$$\left(y_t^1,\ y_t^2\right) = \left(\tilde{D}_t^1,\ \lceil \frac{\log\Theta - a_0 - \Phi^{-1}(1-\alpha) b_0 - \left(a_1 + \Phi^{-1}(1-\alpha) b_1\right)\tilde{D}_t^1}{a_2 + \Phi^{-1}(1-\alpha) b_2} \rceil\right); \quad (3.12)$$

where $\lceil . \rceil$ is the ceiling function, namely the smallest integer greater than the inside value.

□

CHAPTER 4

Investigating the Benefits of Early Bed Request on Emergency Department Performance

4.1 Introduction

Emergency department (ED) patients admitted to a hospital's inpatient unit often endure excessive wait times for the transfer. This prolonged wait is not only a result of the ED operations, but is a consequence of hospital-wide operations. Improvements in the transfer delay may result from making changes in patient work flow. Nearly half of ED patients who are eventually hospitalized stay in IWs, and the vast majority of IW patients come directly from ED. For this reason, we focus on the hospital sub-network consisting of the ED and the IW ("ED+IW").

Variety of factors contribute to the boarding of ED patients (e.g., Asplin et al. (2003)). In this chapter, we are interested in the timeliness of bed requests for an ED patient. Typically, in almost all EDs, the request for an IW bed and the subsequent allocation work are delayed until the IW admission decision for an ED patient is ascertained, which happens at the end of the patient's ED service. A typical ED patient spends several hours in the ED before a final disposition decision is made. What if the admission decision can be done earlier? Then the IW could start the preparation to admit earlier, and the waiting time for admission to IW can be reduced. Note that, this waiting time for admission to inpatient units is also known as the ED boarding time, which is the time elapsed between patient's admission decision and the time that patient is physically transferred to the designated inpatient unit.

Statistical models for predicting an ED patient's admission probability before the final disposition, possibly at triage, have been widely studied in the literature (e.g., Sun et al. (2011)), and have been shown to be highly reliable. Our goal in this chapter is, rather than

developing predictive models, to discuss both the actual ways and operational benefits for applying the admission prediction to the decision making. Qiu et al. (2015) investigated the threshold and timing of early bed request by using the admission probability prediction, and Peck et al. (2012) suggested that the summation of predicted probabilities could be used as the estimation of the inpatient beds demand.

In this chapter, we demonstrate the effectiveness of integrating the patient admission probability within the bed request process. Besides, we propose a framework that aggregates patient admission probabilities to improve the timeliness of inpatient transfer and reduce the ED boarding time. Unlike Qiu et al. (2015), we propose a framework that utilizes patient admission probabilities in an aggregated way, rather than only uses the probability for each individual patient one by one, to avoid the sensitivity of the prediction accuracy of an individual patient. Different from Peck et al. (2012), we employ the patient admission probabilities to guide bed request decisions, besides minimizing the expected costs associated with boarding delays and bed capacity wastage.

This chapter is organized as follows. We begin by reviewing the related literature in Section 4.2. In Section 4.4, we introduce the two policies proposed to guide the bed request process for ED patients by using the predicted IW admission probabilities. Then, we describe the simulated ED+IW subnetwork: we explain how a patient moves and how each proposed policy is implemented in the simulation model in Section 4.5; and show the empirical studies carried out to build a simulation model of high fidelity in Section 4.6. In Section 4.7, we discuss the simulation studies carried out to evaluate the effects of early bed request on hospital operations. We first validate the simulation model in Section 4.7.1, and it follows by the interpretation of the results of a number of simulation runs in Section 4.7.2. Finally, we verified in Section 4.8 that early requesting beds for ED patients who are likely to be hospitalized in the IW can reduce the overall ED boarding times to some extent; however, early requesting beds too aggressively might worsen the transfer performance; causing significantly increasing workload without any benefit.

4.2 Literature Review

Many articles studied hospital admission decisions using information available at triage, such as patients' demographic characteristics, arrival time and mode, clinical measures, complaints, severity level, and so on; with application of a variety of methodologies. Decisions can be related to classification or probability. For example, Leegon and Aronsky (2006), Li et al. (2009) and Sun et al. (2011) used different classification tools to identify whether or not an ED patient is going to be admitted. Peck et al. (2012) and Peck et al. (2013) predicted the probability that each individual ED patient is to be admitted, and Peck et al. (2012) proposed that the probabilities can be used as estimation of the inpatient bed needs in the future. Although many of these papers demonstrated that the admission prediction is promising and suggested that there are potential benefits, none of them investigated the benefits of using these predictions to guide the patient flow in general and the inpatient transfer in particular.

Peck et al. (2014) used the sum of admission probabilities to prioritize discharges in inpatient units. Qiu et al. (2015) proposed a model using the predicted admission probability for each individual ED patient to determine the optimal time to start the bed preparation. In this chapter, we propose two policies to guide the inpatient bed request decisions. Policy 1 evaluates each ED patient's admission probability against the admission probability threshold, similar to what Qiu et al. (2015) proposed. Policy 2 extends the idea of Peck et al. (2014): it considers all ED patients' admission probabilities in aggregate, in order to make bed request decisions to take care of all patients in the ED. Furthermore, to better match the bed capacity with the bed demand, Policy 2 uses a mathematical model to handle the costs associated with transfer delay and bed capacity wastage.

For building the simulation model to capture the patient flow within the ED+IW subnetwork to implement our proposed bed request policies, it is important to keep track of each patient along each step in ED, IW and the ED-to-IW transfer. Therefore, we need to estimate the time duration that an ED patient stay within each service and transfer segment. Many articles estimated parts of or the entire ED length of stay (LOS) using information on each individual patient's demographic and clinical information, hour-of-the-

day and day-of-the-week, and the departmental aggregated occupancy and staffing levels. For example, Sun et al. (2012) used quantile regression and Plambeck et al. (2014) applied fluid model estimators and linear regression to predict the waiting time before being seen by a doctor in the ED. Yoon et al. (2003) applied linear regression to estimate the total ED LOS.

Due to the lack of patient clinical information in our data, we need methods to estimate ED LOS given ED and hospital census levels. In particular, Rathlev et al. (2007) demonstrated that the daily mean ED LOS depends on the number of elective surgical admissions, number of ED admissions, and hospital occupancy. In general, Whitt (1999) showed that it is reliable to predict the waiting time of a customer in the queue using the number of customers ahead of that customer. In this chapter, we predict the length of time during each segment of ED LOS (receiving service in ED and waiting for transfer in ED) using patient's gender and age, day-of-the-week, hour-of-the-day, number of patients in the ED service, number of transfers, and hospital occupancy.

4.3 Description of IW Bed Request

We use Figure 4.1 to show the stylized process of how a bed request is triggered for an ED patient in typical hospital operations. When a patient starts the service in the ED, a few diagnoses and tests are carried out during this process, at the end of ED service, the disposition decision on whether or not he needs to be admitted into IW is made for him, based on corresponding results. If he is decided as an IW admission, then a bed request is put in, thereafter staff in IW will begin to allocate a bed and prepare it for his admission. During this period, the patient is boarding in the ED. When the allocation and preparation work is done, this patient is able to be physically transferred to IW. Therefore, the time used to allocate and prepare an IW bed is just his ED boarding time. We call the process of fulfilling a bed request as the *IW bed-allocation process*, and the corresponding time spent in this process as *IW bed-allocation time*. We denote the IW bed-allocation time and the ED boarding time of any given ED patient i as $BED_{ED\text{-}IW_i}$ and $EDBOARD_i$, respectively.

Then we have:

$$EDBOARD_i = BED_{ED\text{-}IW_i}. \tag{4.1}$$

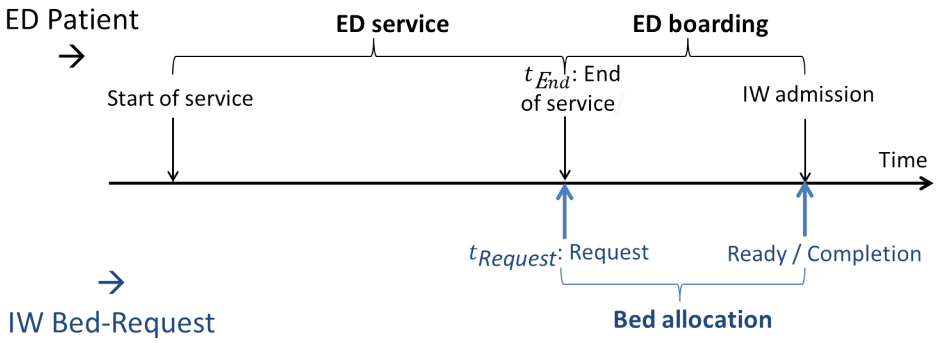


Figure 4.1: Regular Bed Request

What if the bed request is put in earlier, within the course of a patient's ED service? Then, the bed allocation process overlaps with the patient's ED service process. Thereby, patient boarding can be reduced, and the corresponding ED boarding time is

$$EDBOARD_i = \max\Big\{BED_{ED\text{-}IW_i} - \{t_{End} - t_{Request}\},\, 0\Big\}. \tag{4.2}$$

It is possible that the bed allocation is done before an ED patient finishes the ED service, when the bed allocation time is shorter than the difference between the end of ED service and the time of bed request, then a bed has already been done with the necessary allocation process for the patient's admission when he finishes the ED service. Therefore, the patient can be transferred immediately after the ED service, the ED boarding time is just zero. Under this case, the requested bed will have to wait for the transferred patient, since it finishes the bed-allocation before it is occupied by the transferred patient. This time period is named as the IW-bed idle time, and denoted as $BEDIDLE_i$ for the ith ED patient and can be expressed as follows:

$$BEDIDLE_i = \Big\{\{t_{End} - t_{Request}\} - BED_{ED\text{-}IW_i}\,,\, 0\Big\} \tag{4.3}$$

In the next, we propose actual ways of using the predicted admission probability to trigger the bed request before the end of the patient's ED service. For the rest of this

chapter, we name the request for an IW bed which is sent out at the end of the patient's ED service as *regular bed request* , and the bed request which is sent out based on the predicted admission probability before the end of the patient's ED service as *early bed request*.

4.4 Making Early IW Bed Requests: Two Policies

Throughout this chapter we make three important assumptions: First, each ED patient's true IW admission probability is known at the beginning of the patient's ED service. We use i to index each patient, and X_i to denote patient i's IW admission probability. The second assumption is that, the IW bed-allocation time for regular and early bed request has the same probability distribution, which will be discussed in Section 4.6. The third assumption is that, the timing of a bed request does not affect the ED patient's service time, and whether or not the patient will be admitted.

We propose two policies to guide hospitals in making bed request decision by taking advantage of the IW admission probability prediction, so as to lower the ED boarding times. More specifically, Policy 1 is at the patient-level in the sense that it early requests IW bed for each individual ED patient by taking into account the patient's IW admission probability; Policy 2 is at the system-level as it early requests IW beds as a batch based on IW admission probabilities for all patients who are currently in the ED and the outstanding early bed requests.

4.4.1 Policy 1: Early Requesting A Bed for Each Individual ED Patient

Policy 1 makes the early bed request decision for each individual ED patient based on the patient's IW admission probability. If the IW admission probability of any ED patient i (X_i) is greater than a pre-determined threshold p^* (i.e., $X_i > p^*$), then he is given an early bed request; otherwise, no early bed request action is triggered. In addition, we assume that all early bed request decisions are made at the beginning of patients' ED services.

At the end of the patient's ED service, the disposition decision on whether or not he is to be admitted into IW is made. If a decision is made to admit the patient and had already

been given an early bed request, then he will be transferred to the reserved bed in IW when the IW bed-allocation process is completed; if a decision is made to admit the patient but was not given an early bed request, then a regular bed request is placed for the patient immediately; if the patient is not admitted into IW (admitted into non-IW or discharged) but an early bed request had been made for the patient, then the early bed request is canceled immediately; if the patient is not admitted into IW (admitted into non-IW or discharged) and an early bed request was not made, then no further action is taken.

As assumed before, neither the IW bed-allocation time nor the ED service time depends on the time when bed request is put in. Then, no matter regular or early bed request is given for a patient, it would take the same amount of time to complete the IW bed-allocation process and the ED service, respectively. Therefore, the amount of ED service time can be saved from the transfer delay if a patient is given the early bed request at the beginning of the ED service. We denote the ED service time for any ED patient i as $EDSERVICE_i$. Then,

$$EDBOARD_i = \max\Big\{BED_{ED\text{-}IW_i} - EDSERVICE_i \times \mathbf{1}\{X_i > p^*\}\,,\, 0\Big\}. \tag{4.4}$$

As explained before, it is possible that the requested bed will have to wait for the transferred patient. As the early bed request is sent out at the beginning of the patient's ED service in Policy 1, the bed idle time for the ith ED patient can be expressed as:

$$BEDIDLE_i = \max\Big\{EDSERVICE_i \times \mathbf{1}\{X_i > p^*\} - BED_{ED\text{-}IW_i}\,,\, 0\Big\}. \tag{4.5}$$

The hospital wants the ED boarding times to be reduced, so it gives the incentive to early request a bed for a likely inpatient. However, it does not mean that the hospital would like to make the IW bed ready unnecessarily early and hold for an ED patient to finish the ED service. Therefore, it would be better to come up with a policy that can make a good balance. In addition, the early bed request might turn out to be unnecessary, in which case the workload is increased without any benefit. Setting the threshold of early bed request too high would mean very few early requests; and setting the threshold too low would

mean too many unnecessary requests. Thus, there is an "optimal" level for the threshold. Clearly, Policy 1 is sensitive to the choice of the threshold value. Hence, we propose Policy 2, which makes bed request decision by aggregating the IW admission probabilities of all ED patients, and takes into account the pros and cons of early bed request.

4.4.2 Policy 2: Early Requesting Beds As a Batch

Instead of using the IW admission probability prediction for each individual patient one by one to make the bed request decision, Policy 2 considers the predictions for all the patients who are currently staying in the ED. To strike the balance between the ED boarding time and the workload, Policy 2 uses the classic newsvendor model (e.g., Porteus (2002)) to determine the number of bed requests in the batch at the beginning of each hour.

At the beginning of each hour h, hospital sees ED_{census_h} patients in the ED, who are currently receiving service in the ED. Recall that, X_i is used to denote the IW admission probability of any ED patient i. Here, a new variable Y_i is introduced to indicate whether ED patient i is going to be admitted or not:

$$Y_i = \begin{cases} 1, & \text{w.p. } X_i, \\ 0, & \text{w.p. } 1 - X_i. \end{cases} \tag{4.6}$$

Then, $Y_i \sim \text{Bernoulli}\,(X_i)$. In addition, it is reasonable to assume that $Y_1, Y_2, \cdots, Y_n$ are independent. At the beginning of hour h, hospital sees ED_{census_h} patients in the ED, $S_h \triangleq \sum_{i=1}^{ED_{census_h}} Y_i$ patients are going to be admitted. S_h follows a Poisson-binomial distribution, with mean $\sum_{i=1}^{ED_{census_h}} X_i$ and variance $\sum_{i=1}^{ED_{census_h}} X_i(1 - X_i)$. The possible values of S_h are all non-negative integers from 0 to ED_{census_h}, namely $S_h \in \{0, 1, \cdots, ED_{census_h}\}$.

Let θ_h denote the number of bed requests at hour h, $\theta_h \in \{0, 1, 2, \cdots\}$. The hospital faces underage cost if requesting too few beds. On the other hand, if requesting too many beds, then the hospital faces the overage cost. Let c_p and c_b be the unit cost of boarding an ED patient and holding an IW bed, respectively. At the beginning of hour h, if θ_h beds

are requested correspondingly, then the mismatch cost is:

$$C(\theta_h) = \begin{cases} c_p(S_h - \theta_h), & \text{if } \theta_h < S_h, \\ 0, & \text{if } \theta_h = S_h, \\ c_b(\theta_h - S_h), & \text{if } \theta_h > S_h. \end{cases} \tag{4.7}$$

Therefore, the total expected mismatch cost of requesting θ_h beds for ED_{census_h} patients can be expressed as:

$$\begin{aligned} \mathrm{E}\left(C\left(\theta_h\right)\right) &= c_p \mathrm{E}\left(S_h - \theta_h\right)^+ + c_b \mathrm{E}\left(\theta_h - S_h\right)^+ \\ &= c_p \sum_{s=\theta_h+1}^{+\infty} \left(s - \theta_h\right) f_h(s) + c_b \sum_{s=0}^{\theta_h - 1} \left(\theta_h - s\right) f_h(s), \end{aligned} \tag{4.8}$$

where $f_h(s)$ is the probability mass function of S_h.

The hospital would like to lower the ED boarding times, but it can only be done at the cost of increasing the cost of unnecessary bed requests. To balance the two conflicting objectives, we can minimize the total expected waiting costs of ED patients and IW beds, namely,

$$\min_{\theta_h} \mathrm{E}\left(C\left(\theta_h\right)\right). \tag{4.9}$$

$$\mathrm{E}\left(C\left(\theta_h + 1\right)\right) - \mathrm{E}\left(C\left(\theta_h\right)\right) = -c_p + \left(c_p + c_b\right) \sum_{s=0}^{\theta_h} f_h(s) = -c_p + \left(c_p + c_b\right) F_h\left(\theta\right), \tag{4.10}$$

where F_h is the cumulative distribution function of S_h. $-c_p + \left(c_p + c_b\right) F_h\left(\theta_h\right)$ is an increasing function of θ_h, then the optimal solution to (4.9) satisfies

$$\begin{aligned} \theta_h^* &\triangleq \arg\min_{\theta_h} \mathrm{E}\left(C\left(\theta_h\right)\right) \\ &= \min\left\{\theta_h \in \mathbb{N} : F_h\left(\theta_h\right) \geq \frac{c_p}{c_p + c_b}\right\} \\ &= \min\left\{\theta_h \in \mathbb{N} : F_h\left(\theta_h\right) \geq \frac{1}{1+\gamma}\right\}, \end{aligned} \tag{4.11}$$

where $\gamma = c_b/c_p$ is the relative cost of holding one bed to boarding one patient, $\mathbb{N} = \{0, 1, 2, \cdots\}$. Equation (4.11) shows that θ_h^* is uniquely determined by the relative cost γ and the distribution of S_h.

θ_h^* is the optimal number of bed requests for all patients currently in the ED at the beginning of hour h. It is possible that some of them have already been counted when making the bed request decision in the previous hours. Thereby, we need to do some adjustments. If $REQUEST_h$ beds are under processing for bed requests of previous hours, then this number should be deducted from the bed request of the current hour h. Therefore, Policy 2 requests

$$\max\{\theta_h^* - REQUEST_h\,,\, 0\} \tag{4.12}$$

number of beds at the beginning of any hour h.

In Policy 2, the bed request action is taken once at the beginning of each hour, and in addition, it is not triggered by any specific patient directly. Each hour h, a batch of beds is requested to take care all patients in the ED, the batch size is determined by (4.12). Any bed from any batched bed request is sequentially held in the ready-bed-queue, after it complete the necessary bed-allocation process. Any ED patient sequentially joins the ready-patient-queue, after he is done with the ED service and also disposed as an IW admission. The first patient in the ready-patient-queue will be transfered to the first bed in the ready-bed-queue.

Before the end of this section, we introduce the approximation we consider for the Poisson-binomial distribution, when actually solving the optimal solution (4.11). F_h, the cumulative distribution function of Poisson-binomial distributed S_h can be expressed as follows:

$$F_h\,(k) = \sum_{l=0}^{k} \sum_{A \in M_l} \prod_{i \in A} X_i \prod_{i \in A^c} (1 - X_i), \quad k = 0, 1, \cdots, ED_{census_h}, \tag{4.13}$$

where A^c is the complement of A, and M_l is the set of all subsets of l integers that can be selected from $\{1, \cdots, ED_{census_h}\}$. For example, if $ED_{census_h} = 3$, then $M_2 = \{\{1, 2\}, \{1, 3\}, \{2, 3\}\}$. For most of the hours, the ED total census is greater than 40. Then, the computation seems difficulty. Hong (2013) demonstrated that normal cdf can approximate the Poisson-binomial cdf very well, especially when the success probabilities

follow a Beta distribution. In particular, the cdf of the Poisson-binomial distribution S_h is approximated by

$$F_h(k) \approx \Phi\left(\frac{k+0.5-\sum_{i=1}^{ED_{census_h}} X_i}{\left(\sum_{i=1}^{ED_{census_h}} X_i(1-X_i)\right)^{1/2}}\right), \quad k=0,1,\cdots,ED_{census_h}, \tag{4.14}$$

where Φ is the cdf of the standard normal distribution, $\sum_{i=1}^{ED_{census_h}} X_i$ and $\left(\sum_{i=1}^{ED_{census_h}} X_i(1-X_i)\right)^{1/2}$ are the mean and standard deviation of S_h, 0.5 is for the purpose of continuity correction.

As the cdf of normal is strictly increasing and continuous, then the optimal solution (4.11) can be approximated as follows:

$$\begin{aligned}
\theta_h^* &= \min\left\{\theta_h \in \mathbb{N} : F_h(\theta_h) \geq \frac{1}{1+\gamma}\right\} \\
&\approx \min\left\{\theta_h \in \mathbb{N} : \Phi\left(\frac{\theta_h+0.5-\sum_{i=1}^{ED_{census_h}} X_i}{\left(\sum_{i=1}^{ED_{census_h}} X_i(1-X_i)\right)^{1/2}}\right) \geq \frac{1}{1+\gamma}\right\} \\
&= \left\lceil \Phi^{-1}\left(\frac{1}{1+\gamma}\right) \cdot \left(\sum_{i=1}^{ED_{census_h}} X_i(1-X_i)\right)^{1/2} + \sum_{i=1}^{ED_{census_h}} X_i - 0.5 \right\rceil,
\end{aligned} \tag{4.15}$$

where Φ^{-1} is the inverse of standard normal cdf, $\lceil x \rceil$ is the smallest integer greater than or equal to x. Later on, in the simulation study, we will use the approximated solution (4.15) to replace the θ^* in the expression of (4.12).

4.5 Simulation Model

In this section, we describe the discrete-event simulation model built to capture the patient flow inside the ED+IW subnetwork and report results of a series of simulation studies conducted to investigate the potential effects of early bed requests on the ED boarding times. The simulation model is meant to capture the patient flow inside the ED+IW subnetwork at a relatively detailed level. The model keeps track of each patient's stay in the subnetwork, and it can be seen as a discrete-time model where time proceeds in the unit of one second. We first describe the 3-step procedure to complete a bed request in Section 4.5.1, then

discuss how each bed request policy is integrated within the patient flow in Section 4.5.2, finally explain how a patient goes through the simulation model in Section 4.5.3.

4.5.1 3-Step Bed-Request Procedure

In the following, we describe the 3-step procedure to complete any bed request for ED patients in the simulation model, where j is used to index an IW bed.

1. The IW bed-allocation time, denoted as $BED_{ED\text{-}IW_j} \in \mathbb{R}_+$, is randomly assigned using Model (4.20), when a bed request is sent out. The bed request joins the queue to seize the next vacant/flexible IW bed. A vacant IW bed is the one that is neither occupied by a patient, nor has already been seized by another bed request; a flexible IW bed is the one that has been processed for a canceled bed request.

2. When an IW bed j is seized, the bed-allocation process begins, which lasts for $BED_{ED\text{-}IW_j}$. Once this process is completed, IW bed j is ready to accommodate an ED-to-IW patient. We call this time point t_j^0.

3. IW bed j remains ready to admit an ED-to-IW patient, until there is an ED-to-IW patient transferred to it. Once the transfer occurs, the IW bed j is occupied. We denote this time point by t_j^1. The time difference between t_j^0 and t_j^1 is the time that IW bed j waits for an ED-to-IW patient, which is referred to as IW-bed idle time and denoted as $BEDIDLE_j$. Namely,

$$BEDIDLE_j = t_j^1 - t_j^0. \tag{4.16}$$

   The IW-bed idle times are collected and used as one of the performance measures in the following simulation study.

4.5.2 Formal Description of Bed-Request Policies

4.5.2.1 Policy 1

For each ED patient i, do the following.

1. At the beginning of ED service, if IW admission probability (X_i) is greater than the pre-specified threshold p^*, namely $X_i > p^*$, an early bed request is sent out, which follows the 3-step bed request procedure; otherwise no early bed request action is triggered.

2. At the end of ED service (s_i^0), the admission decision is made for patient i: *IW-admission* or *IW-denial*.

   - If *IW-admission* and
     - $X_i > p^*$, then patient i will be transferred to the IW bed seized by the early bed request when it is completed;
     - $X_i \leq p^*$, then a regular bed request is sent out, which follows the 3-step bed request procedure; patient i will be transferred to the IW bed seized by the regular bed request when it is completed.
   - If *IW-denial* and
     - $X_i > p^*$, then the early bed request is canceled as follows: if the early bed request is still waiting in the queue to seize an vacant/flexible IW bed, then it is removed from the queue; if there is an IW bed j seized and processing for the early bed request, then IW bed j still finishes the remaining IW-bed allocation process, but it is not reserved by the early bed request for patient i any more, and it becomes available to be seized by another early/regular bed request for an ED patient;
     - $X_i \leq p^*$, then no bed request action is taken for patient i.

4.5.2.2 Policy 2

At the beginning of any hour h, compute θ_h^* using the formula (4.15), and count the number of ongoing bed requests and record it as $REQUEST_h$. Then,

- If $\theta_h^* > REQUEST_h$, then send out $\theta_h^* - REQUEST_h$ number of bed requests, each following the 3-step bed request procedure. When a bed request is completed, the bed joins the ready-bed-queue to accommodate ED patients in the ready-patient-queue in the order they arrive.

- If $\theta_h^* \leq REQUEST_h$, then no bed request action is triggered in this hour.

4.5.3 Patient Route

In the simulation model, a patient can enter the system through the ED from outside, or via IW either from outside or hospital units other than ED or IW.

4.5.3.1 ED Patient

If a patient enters the system via ED from outside, who is named as "ED patient", then the patient goes through the following steps:.

1. Join the queue to seize the next vacant ED bed for admission. When an ED bed is seized, he is admitted into ED; the ED in-service census, denoted as $EDSERV_{census}$, increases by one. Upon admission, we randomly assign values of the following variables for any ED patient i:

   - gender, $GENDER_i \in \{\text{Male, Female}\}$ and age, $AGE_i \in \{0 : \text{age} < 60,\ 1 : \text{age} \geq 60\}$ (based on the proportions given in Table 4.2),
   - IW admission probability, $X_i \in (0, 1)$ (based on the beta distribution shown in Section 4.6.4),
   - non-IW admission probability, $q_i \in (0, 1)$ (based on the probability mentioned in Section 4.6.4),
   - ED service time, $EDSERVICE_i \in \mathbb{R}_+$ (based on Model (4.19)),
   - non-IW bed-allocation time, $BED_{ED\text{-}nonIW_i} \in \mathbb{R}_+$ (based on Model (4.21)).

2. Patient i stays in ED for the preassigned length of time, $EDSERVICE_i$, to finish the ED service. During this period, a bed request decision might be triggered depending on which policy is active as described in Section 4.5.2.

3. At the end of ED service (the time point is recorded as s_i^0), the IW admission decision is made based on X_i: *IW-admission* or *IW-nonadmission*. (Patient i has the probability of X_i to be decided as an IW-admission.) The ED in-service census ($EDSERV_{census}$) decreases by one.

- If *IW-admission*, the ED-to-IW census, denoted as $ED\text{-}IW_{census}$, increases by one. ED patient i waits in the ED until an IW bed is ready to accommodate him (waiting time can be zero). Then (the time point is recorded as s_i^1), ED patient i releases the occupied ED bed and is physically transferred to IW, the ED-to-IW census ($ED\text{-}IW_{census}$) decreases by one. The time elapsed between the end of ED service (s_i^0) and the time of IW admission (s_i^1) is just the ED boarding time for ED patient i, denoted as $EDBOARD_i$. Then,

$$EDBOARD_i = s_i^1 - s_i^0. \tag{4.17}$$

- If *IW-nonadmission*, the non-IW admission decision is made for ED patient i based on q_i: *nonIW-admission* or *nonIW-nonadmission*. Here, non-IW is used to denote all inpatient units other than IW.
  - If *nonIW-admission*, the ED-to-nonIW census, denoted as $ED\text{-}nonIW_{census}$, increases by one. ED patient i waits in the ED for the completion of the transfer process, which lasts for the preassigned length of time, $BED_{ED\text{-}nonIW_i}$. When the waiting time uses up, ED patient i is assumed to be transferred to non-IW, and releases the occupied ED bed; the ED-to-nonIW census ($ED\text{-}nonIW_{census}$) decreases by one. As we do not track what happens inside non-IW, ED patient i leaves the system.
  - If *nonIW-nonadmission*, ED patient i releases the ED bed and leaves the ED and the system immediately.

4.5.3.2 NonED-to-IW Patient

If a patient enters the system via IW from outside the hospital or hospital units other than ED or IW, who is named as an "nonED-to-IW patient", the following steps are taken. Note that nonED-to-IW patients are different from ED-to-IW patients.

1. Upon arrival, any nonED-to-IW patient i is randomly assigned the IW bed-allocation time, which is denoted as $BED_{nonED\text{-}IW_i}$ ($\in \mathbb{R}_+$), based on Model (4.22). The patient then joins the queue to wait for the next vacant IW bed. The time point is recorded as r_i^0; the nonED-to-IW census, denoted as $nonED\text{-}IW_{census}$, increases by one.

2. Once a vacant IW bed is seized, nonED-to-IW patient i waits for the IW bed to complete the bed-allocation process, which lasts for the preassigned time of $BED_{nonED\text{-}IW_i}$.

3. When the bed-allocation process is completed (the time point is recorded as r_i^1), patient i is admitted into IW; the nonED-to-IW census ($nonED\text{-}IW_{census}$) decreases by one. The time difference between entry (r_i^0) and admission (r_i^1) is named as the nonED-to-IW waiting time for nonED-to-IW patient i and denoted as $WAIT_{nonED\text{-}IW_i}$. Then,

$$WAIT_{nonED\text{-}IW_i} = r_i^1 - r_i^0. \tag{4.18}$$

4.5.3.3 IW Patient

When a patient is physically admitted into IW from either ED (ED-to-IW) or other than ED (nonED-to-IW), who is named as an "IW patient", following steps are taken.

1. Upon the admission of any IW patient i, she/he is randomly assigned values of the following variables; the IW census, denoted as IW_{census}, increases by one.

   - night-of-stay (# of nights), $IWNIGHT_i \in [0, 40]$
     (based on the empirical distribution from real hospital data, where summary statistics are given in Table 4.3),
   - time-of-discharge (hour-of-the-day), $IWDIS_i \in [0, 24]$
     (based on the empirical distribution shown in Figure 4.5).

2. Starting with admission, IW patient i spends the pre-assigned $IWNIGHT_i$ number of nights in IW in total. The remaining number of nights deceases by one at the end of each day (midnight).

3. When the remaining number of nights decreases to zero, IW patient i is to be discharged from IW the next day. The exact time-of-the-day when patient i is determined by $IWDIS_i$.

4. When IW patient i is discharged from the IW (leaves the system), the occupied IW bed is released and becomes vacant; the IW census (IW_{census}) decreases by one.

During a patient's stay in the system, the values of following variables are recorded: two time variables at the patient-level: ED boarding time and nonED-to-IW waiting time; one time variable at the bed-level: IW-bed idle time; five census variables: ED in-service census, ED-to-IW census, ED-to-nonIW census, nonED-to-IW census, and IW census. These variables will be used as performance measures to validate the simulation model and demonstrate the effects of early bed request on the ED boarding times in Section 4.7. In addition, the five census variables will also be used as inputs to the Model (4.19) - Model (4.22) in Section 4.6.

4.6 Specification of Simulation Model Parameters

To capture what happens throughout an individual patient's stay in the real hospital, all inputs to the simulation model including probability distributions and predictive models mentioned in the previous section, are estimated using data from the Rambam Hospital in Israel. Emergency care for patients at this hospital occurs in one of the eleven EDs. In our analysis, we will only focus on the main ED consisting of four major departments: Internal Medicine, Surgery, Traumatology and Orthopedic ED. As explained in Armony et al. (2011), other departments are located away from the main one, and are specialized for particular patients, such as Pediatrics or Ophthalmology. For brevity, we will refer to those four EDs as the ED. There are five IWs in the Rambam Hospital, which are responsible for the treatment of a wide range of internal conditions. During the first three months in 2007, one of the IWs was in charge of an additional 20-bed sub-ward, therefore we use the data from April 2007 to October 2007. All five IWs provide similar medical services, therefore we combine five IWs together into one with the aggregated capacity, which is referred to as the IW afterwards.

For each patient, we observed patient's gender and age at time of admission, time stamps for entry, exit and the first procedure time in inpatient units, and the index stamp to distinguish hospital units. From those stamps, we can come up with the length of stay within each stage of hospital operations, and hospital-level census data as well. Summaries

of key patient-level characteristics and hospital-level censuses are shown throughout this section.

4.6.1 Bed Capacity

In the Rambam hospital, the ED has 40 beds and it treats on average 248 patients per day. After checking the histogram of ED total census at the beginning of each hour in Figure 4.2, we realize that the ED holds more than 40 patients simultaneously about 40% of the time and can even exceed 100 patients. Besides, Armony et al. (2011) mentioned that the Rambam hospital has no rigid constraint on the ED capacity, beds are added in accordance to congestion levels. Then, we set the ED capacity at infinity in the simulation model, $ED_{cap} = +\infty$. There are five IWs in the Rambam hospital, which accommodate around 1000 patients monthly. Five IWs are responsible for the inpatient medical care to similar types of patients, then the IW bed is reasonably treated as interchangeable. In the simulation model, all IW beds are considered aggregately. Although the five IWs have 201 beds in record, the maximum number of patients in IW simultaneously is observed as 210, therefore set the IW capacity fixed at 210, $IW_{cap} = 210$.

Note that, the total ED census consists of three parts: ED in-service census, ED-to-IW census, and ED-to-nonIW census, therefore the summation cannot exceed the ED capacity, namely $EDSERV_{census} + ED\text{-}IW_{census} + ED\text{-}nonIW_{census} \leq ED_{cap}$; in addition, the IW census cannot exceed the IW capacity, namely $IW_{census} \leq IW_{cap}$.

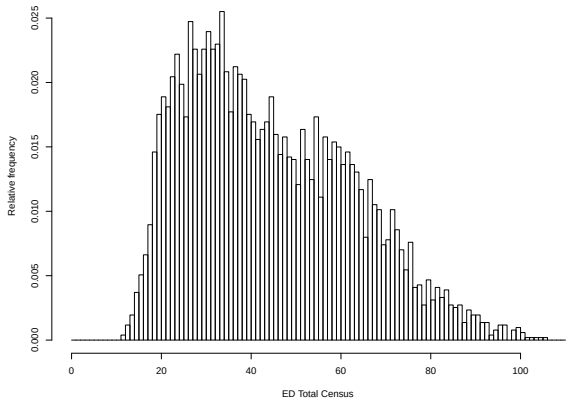
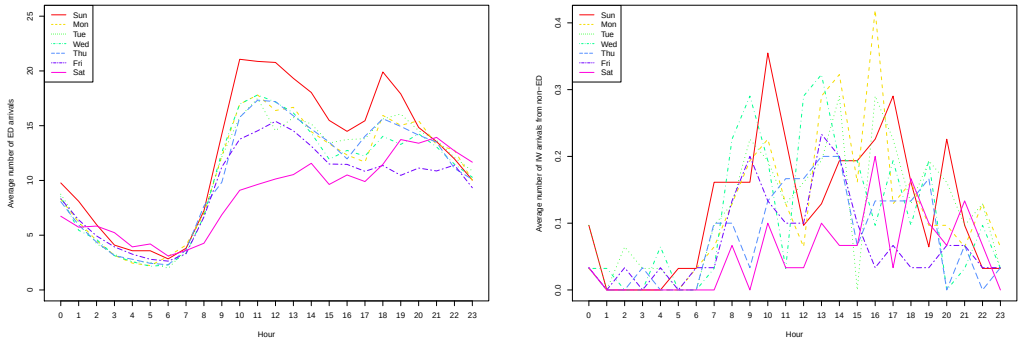


Figure 4.2: Histogram of ED Total Census

4.6.2 Arrival Rate

Then, we investigate the two arrival processes to the ED+IW subnetwork: arrival at ED directly from hospital outside and arrival at IW from other inpatient units (non-IW) ("nonED-to-IW") or hospital outside directly. Panel (a) and (b) in Figure 4.3 show the average number of ED arrivals and nonED-to-IW arrivals over each hour of the week, respectively. For either type of arrival, hour-to-hour patterns are similar throughout a week, while scales differ from day to day: weekdays (Sunday to Thursday) have larger numbers than weekends (Friday and Saturday). We model each arrival process within each hour of the week as a Poisson distribution with rate $\lambda^f_{w,h}$, $f \in \{1,2\}$, $w \in \{1,2,\cdots,7\}$, $h \in \{0,1,\cdots,23\}$. $f = 1$ and $f = 2$ stand for ED arrivals and nonED-to-IW arrivals, respectively. w and h are to distinguish days of the week and hours of the day, respectively. Values of $\lambda^f_{w,h}$, which are estimated from the average number of the corresponding type of arrival over the corresponding hour of the week from the real historical data, are listed in Table 4.1.



(a) ED (b) nonED-to-IW

Figure 4.3: Hourly Average Number of (a) ED Arrivals, and (b) nonED-to-IW Arrivals

4.6.3 Demographics

Table 4.2 lists demographics information for patients who arrive to ED from hospital outside. In the simulation model, age ($\{0 : \text{age} \leq 60, 1 : \text{age} > 60\}$) and gender ($\{0 :$ male, $1 :$ female$\}$) are randomly assigned for each new ED patient upon admission based on the following percentages.

Table 4.1: $\lambda^f_{w,h}$, $f \in \{1,2\}$, $w \in \{1,2,\cdots,7\}$, $h \in \{0,1,\cdots,23\}$

(a) ED ($f = 1$)

	Sunday ($w=1$)	Monday ($w=2$)	Tuesday ($w=3$)	Wednesday ($w=4$)	Thursday ($w=5$)	Friday ($w=6$)	Saturday ($w=7$)
$h=0$	9.77	8.29	8.74	8.52	8.07	8.33	6.73
$h=1$	8.10	6.19	6.19	5.45	5.93	6.40	5.73
$h=2$	5.97	4.52	4.23	4.61	4.37	4.93	5.83
$h=3$	4.10	3.16	3.26	3.13	3.13	3.93	5.23
$h=4$	3.58	2.45	2.55	2.58	2.80	3.27	3.93
$h=5$	3.58	2.23	2.55	2.16	2.43	2.80	4.20
$h=6$	2.84	3.06	1.97	2.29	2.30	2.63	3.10
$h=7$	3.81	4.10	3.68	3.35	3.67	3.30	3.60
$h=8$	7.23	6.97	7.06	6.71	7.67	6.67	4.27
$h=9$	14.23	12.23	12.16	12.58	9.83	11.23	6.87
$h=10$	21.06	16.97	15.77	16.97	15.77	13.77	9.10
$h=11$	20.87	17.81	17.48	17.81	17.33	14.53	9.63
$h=12$	20.77	16.39	14.55	17.16	17.20	15.40	10.13
$h=13$	19.32	16.68	15.74	16.16	15.90	14.53	10.53
$h=14$	18.03	14.45	15.16	14.32	14.70	13.13	11.57
$h=15$	15.48	13.35	13.42	11.94	13.50	11.50	9.63
$h=16$	14.48	12.32	13.77	12.74	11.97	11.47	10.50
$h=17$	15.45	11.68	13.81	12.23	14.03	10.83	9.90
$h=18$	19.90	15.94	15.61	14.03	15.63	11.37	11.47
$h=19$	17.90	15.03	16.06	13.29	14.93	10.47	13.73
$h=20$	14.84	15.45	14.55	14.26	14.13	11.13	13.40
$h=21$	13.55	13.26	13.68	13.06	13.50	10.87	13.93
$h=22$	11.97	12.77	12.10	11.48	11.10	11.40	12.70
$h=23$	10.03	10.00	10.87	10.29	10.10	9.33	11.67

(b) IW ($f = 2$)

	Sunday ($w=1$)	Monday ($w=2$)	Tuesday ($w=3$)	Wednesday ($w=4$)	Thursday ($w=5$)	Friday ($w=6$)	Saturday ($w=7$)
$h=0$	0.10	0.03	0.10	0.03	0.03	0.03	0.03
$h=1$	0.00	0.00	0.00	0.03	0.00	0.00	0.00
$h=2$	0.00	0.00	0.06	0.00	0.00	0.03	0.00
$h=3$	0.00	0.00	0.03	0.00	0.03	0.00	0.00
$h=4$	0.00	0.00	0.03	0.06	0.00	0.03	0.00
$h=5$	0.03	0.00	0.00	0.00	0.00	0.00	0.00
$h=6$	0.03	0.03	0.00	0.00	0.00	0.03	0.00
$h=7$	0.16	0.06	0.10	0.03	0.10	0.03	0.00
$h=8$	0.16	0.13	0.13	0.23	0.10	0.13	0.07
$h=9$	0.16	0.19	0.23	0.29	0.03	0.20	0.00
$h=10$	0.35	0.23	0.19	0.19	0.13	0.13	0.10
$h=11$	0.23	0.13	0.13	0.03	0.17	0.10	0.03
$h=12$	0.10	0.06	0.16	0.29	0.17	0.10	0.03
$h=13$	0.13	0.29	0.19	0.32	0.20	0.23	0.10
$h=14$	0.19	0.32	0.29	0.19	0.20	0.20	0.07
$h=15$	0.19	0.16	0.00	0.19	0.07	0.10	0.07
$h=16$	0.23	0.42	0.29	0.10	0.13	0.03	0.20
$h=17$	0.29	0.13	0.23	0.19	0.13	0.07	0.03
$h=18$	0.16	0.16	0.13	0.10	0.13	0.03	0.17
$h=19$	0.06	0.10	0.19	0.19	0.17	0.03	0.10
$h=20$	0.23	0.10	0.16	0.00	0.00	0.07	0.07
$h=21$	0.10	0.06	0.10	0.03	0.07	0.07	0.13
$h=22$	0.03	0.13	0.13	0.10	0.00	0.03	0.07
$h=23$	0.03	0.06	0.03	0.03	0.03	0.03	0.00

Table 4.2: Demographics

	ED	IW
Male	58%	55%
Age>60	22%	61%
Female	42%	45%
Age>60	31%	63%

4.6.4 Admission Probability

In the real hospital data, we only observe whether or not an individual ED patient is eventually admitted into IW, rather than the exact probability of the IW admission. Then, we need to find a reasonable distribution to model the IW admission probability of a random ED patient. Beta distribution is a suitable model for the random behavior of percentages. Besides, Peck et al. (2014) demonstrated that the admission probability can be modeled with a beta distribution using real hospital data. Therefore, we decide to use a beta distribution to estimate the IW admission probability for our hospital data either.

The mean of the beta distribution should be equal to the fraction of IW admission in the real hospital, 0.126; while the choice of variance is arbitrary as long as the density curve is right-skewed. In fact, the variance is chosen as 0.05 in the simulation study. The density curve plotted in Figure 4.4 confirms the rationality of choosing 0.5 as the variance. Meanwhile, whether or not an ED patient will be admitted should be independent with each other. Therefore, the ED patient's IW admission probability, X, is independently and identically distributed as $\text{Beta}\,(0.15, 1.05)$.

In our study hospital, there are 29.2% of ED patients admitted to inpatient units after the ED service: 12.6% are admitted to IW as mentioned before, and 16.6% are admitted to non-IW (all inpatient units other than IW). 16.6% of non-IW admissions is out of all ED patients, which corresponds to 19.0% out of all ED patients who are not admitted into IW. Therefore, in the simulation study, 19.0% of all ED patients who are not admitted into IW will be decided as the non-IW admissions.

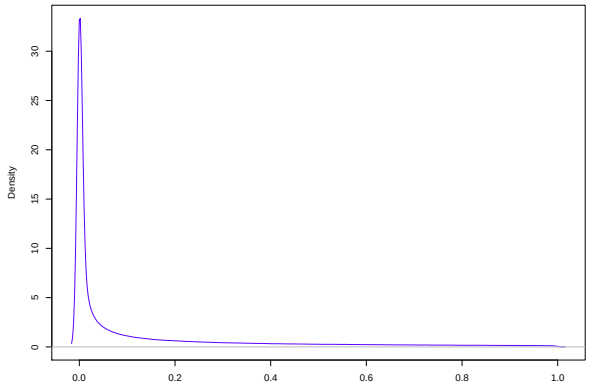


Figure 4.4: Probability Density Function of Beta $(0.15, 1.05)$

4.6.5 Service Times

To keep track of each patient's stay in the simulation model, we need to know the length of time within each step in the ED+IW sub-network. Table 4.3 lists summary statistics of seven patient-level time variables observed from the real data.

Table 4.3: Patient-level Summary Statistics

Variable	Mean	Standard deviation	First quantile	Median	Third quantile
ED total length-of-stay (hours)	4.4	4.0	1.6	3.2	5.7
ED service time (hours)	3.8	3.6	1.4	2.8	4.8
ED boarding time (hours)	2.7	2.9	1.0	1.9	3.4
ED-to-nonIW waiting time (hours)	2.1	3.0	0.4	1.0	2.3
NonED-to-IW waiting time (hours)	3.1	4.2	0.5	1.3	3.7
Night-of-stay in IW (nights)	4.9	5.7	2.0	3.0	6.0
Time-of-discharge from IW (hour-of-day)	15.2	2.9	14.4	15.0	16.5

Patients need to stay in IW for 4.8 nights, on average, from admission to discharge. A lot of things will happen during this long period, therefore it is hard to model the total length of time that a patient spends in IW based on the information obtained upon admission. From the plot of the average number of discharges from IW per hour over each day of the week in Figure 4.5, we observe that most IW discharges happen around 3pm on weekdays (Sunday to Thursday); the peak is shifted earlier to around 2pm and the scale is lowered a lot on Friday; the shape looks flat and the scale is the lowest throughout Saturday. Therefore, we model the total IW LOS by two separate parts: night-of-stay and time-of-discharge, on the

grounds of these exceptional hour-of-the-day and day-of-the-week discharge patterns. First, an integer sampled from the historical night-of-stay data is assigned as the night-of-stay for any IW patient upon admission. Then, a real number is sampled from the historical time-of-discharge data corresponding to the same day of the week as the day when any IW patient is to discharge (the remaining night-of-stay decreases to zero).

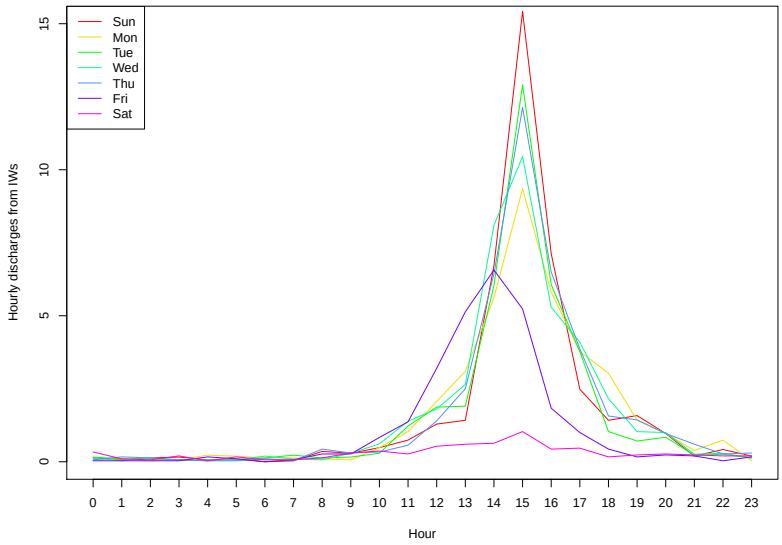


Figure 4.5: IW Discharges

Before the end of this section, we will describe the predictive models built for the following four time variables of any individual patient i: ED service time ($EDSERVICE_i$), IW bed-allocation time for ED-to-IW patients ($BED_{ED\text{-}IW_i}$), non-IW bed-allocation time for ED-to-nonIW patients ($BED_{ED\text{-}nonIW_i}$), and IW bed-allocation time for In-to-IW patients ($BED_{nonED\text{-}IW_i}$). For the last three bed-allocation times, we do not have direct data, instead we observe the actual time that a patient spends waiting in each transfer process (ED-to-IW / ED-to-nonIW / nonED-to-IW). Therefore, we use the observed ED boarding time, ED-to-nonIW waiting time and nonED-to-IW waiting time, to approximate the corresponding bed-allocation times.

The independent variables include: gender ($GENDER_i$), age (AGE_i), day-of-the-week (w_i), hour-of-the-day (h_i), and five hospital-level census variables, which are computed upon admission of patient i. Table 4.4 lists summary statistics for each of five hospital-level

census variables: ED in-service census ($EDSERV_{census}$), ED-to-IW census ($ED\text{-}IW_{census}$), ED-to-nonIW census ($ED\text{-}nonIW_{census}$), nonED-to-IW ($nonED\text{-}IW_{census}$), and IW census (IW_{census}).

Table 4.4: System-level Summary Statistics

Variable	Mean	Standard deviation	First quantile	Median	Third quantile
ED in-service census	38.0	16.7	24.0	35.0	50.0
ED-to-IW census	3.4	2.4	2.0	3.0	5.0
ED-to-nonIW census	3.4	2.2	2.0	3.0	5.0
NonED-to-IW census	0.2	0.5	0.0	0.0	0.0
IW census	167.2	14.2	157.0	168.0	178.0

First, we list all the independent variables we use to fit the linear regression model for each time variable as follows:

$$\begin{aligned}\log\{EDSERVICE_i\} = \alpha_0 &+ \alpha_1 w_i + \alpha_2 h_i + \alpha_3 AGE_i + \alpha_4 GENDER_i \\ &+ \alpha_5 EDSERV_{census_i} + \alpha_6 IW_{census_i} \\ &+ \alpha_7 ED\text{-}IW_{census_i} + \alpha_8 ED\text{-}nonIW_{census_i} + \epsilon_i,\end{aligned} \tag{4.19}$$

$$\begin{aligned}\log\{BED_{ED\text{-}IW_i}\} = \beta_0 &+ \beta_1 w_i + \beta_2 h_i + \beta_3 EDSERV_{census_i} + \beta_4 IW_{census_i} \\ &+ \beta_5 ED\text{-}IW_{census_i} + \beta_6 ED\text{-}nonIW_{census_i} \\ &+ \beta_7 nonED\text{-}IW_{census_i} + \epsilon_i,\end{aligned} \tag{4.20}$$

$$\begin{aligned}\log\{BED_{ED\text{-}nonIW_i}\} = \gamma_0 &+ \gamma_1 w_i + \gamma_2 h_i + \gamma_3 EDSERV_{census_i} \\ &+ \gamma_4 ED\text{-}IW_{census_i} + \gamma_5 ED\text{-}nonIW_{census_i} + \epsilon_i,\end{aligned} \tag{4.21}$$

and

$$\begin{aligned}\log\{BED_{nonED\text{-}IW_i}\} = \tau_0 &+ \tau_1 w_i + \tau_2 h_i + \tau_3 IW_{census_i} \\ &+ \tau_4 ED\text{-}IW_{census_i} + \tau_5 nonED\text{-}IW_{census_i} + \epsilon_i.\end{aligned} \tag{4.22}$$

Then, we estimate the coefficient of each independent variable using least squares. Table 4.5 lists coefficients for variables that are significant at the level of 0.1, where significances (Yes/No) instead of coefficients are indicated for both day-of-the-week and hour-of-the-day.

Table 4.5: Coefficients of Predictive Models

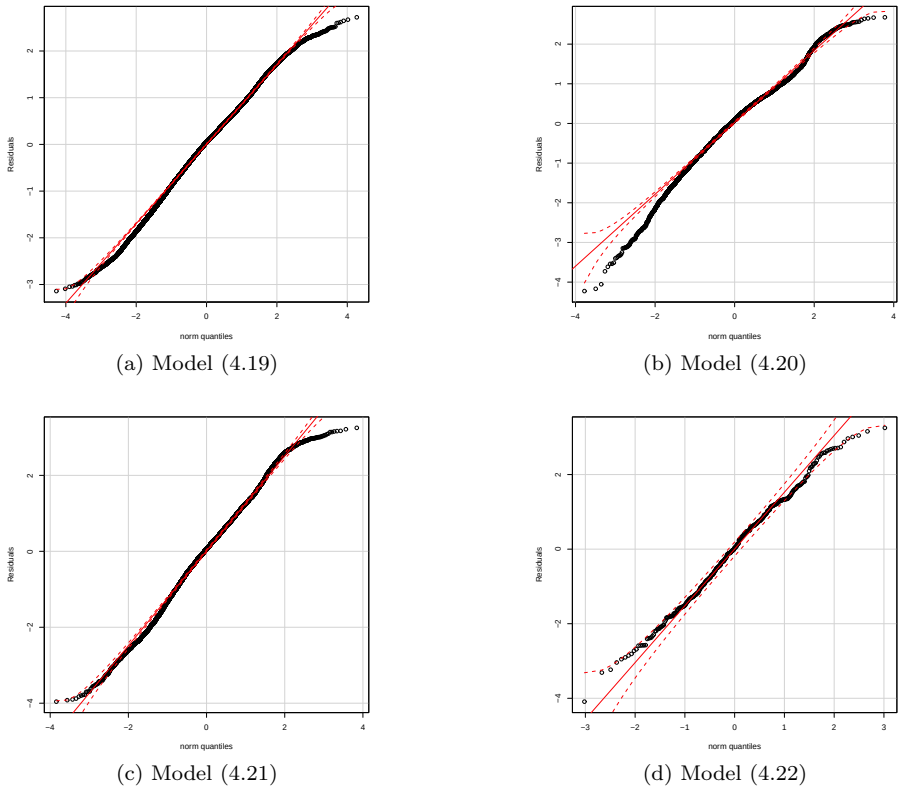
Variable	Model (4.19)	Model (4.20)	Model (4.21)	Model (4.22)
Intercept	0.4038 (0.0660)	-0.5845 (0.2149)	-0.2495 (0.0973)	-3.5364 (0.8583)
Day-of-week	Yes	Yes	Yes	No
Hour-of-day	Yes	Yes	Yes	No
Age	0.2961 (0.0090)	—	—	—
Gender	0.1272 (0.0079)	—	—	—
ED in-service census	0.0063 (0.0004)	0.0037 (0.0013)	No	—
IW census	0.0015 (0.0003)	0.0067 (0.0011)	—	0.0224 (0.0051)
ED-to-IW census	0.0049 (0.0018)	0.0368 (0.0010)	0.0120 (0.0066)	No
ED-to-nonIW census	0.0050 (0.0019)	-0.0147 (0.0059)	No	—
NonED-to-IW census	—	No	—	No
Adjusted R-squared	0.0496	0.0319	0.0084	0.0453
Variance of residuals	0.7759	0.9180	1.6486	1.9163

Notes. Standard errors are in parentheses.

We check whether each error term in the above models is normally distributed with mean as zero and variance as the variance of regression residuals, which are listed in Table 4.5, by depicting the following quantile-quantile plots in Figure 4.6. Except the error term in Model (4.22), all other error terms are far from the normal assumption. Then, when we need a error to do the prediction of any time variable, we consider the nonparametric method for Model (4.19) - (4.21) – sampling from the corresponding regression residuals; we still use the normal distribution $\mathrm{N}(0, 1.9163)$ to randomly generate errors for Model (4.22). Therefore, for each time variable, by directly feeding the actual value of each independent variable into the corresponding model, together with a sampled/generated error, we can get the time value on the log scale. Finally, we can take exponential to get the time value on the original scale.

4.7 Simulation Study

The ED+IW subnetwork is simulated as an non-terminating system. Without loss of generality, the first day in each simulation run is set as Sunday. For each run, we initialize the simulation model with the same non-empty ED and IW; and consecutively simulate for 4368 days, of which the first 728 days are deleted for the warm-up reason, and the remaining 3640 days are divided into 10 non-overlapping batches of equal length of 364 days (52 weeks). The warm-up period is deemed sufficiently long to populate the system



(a) Model (4.19)

(b) Model (4.20)

(c) Model (4.21)

(d) Model (4.22)

Figure 4.6: Quantile-Quantile Plots

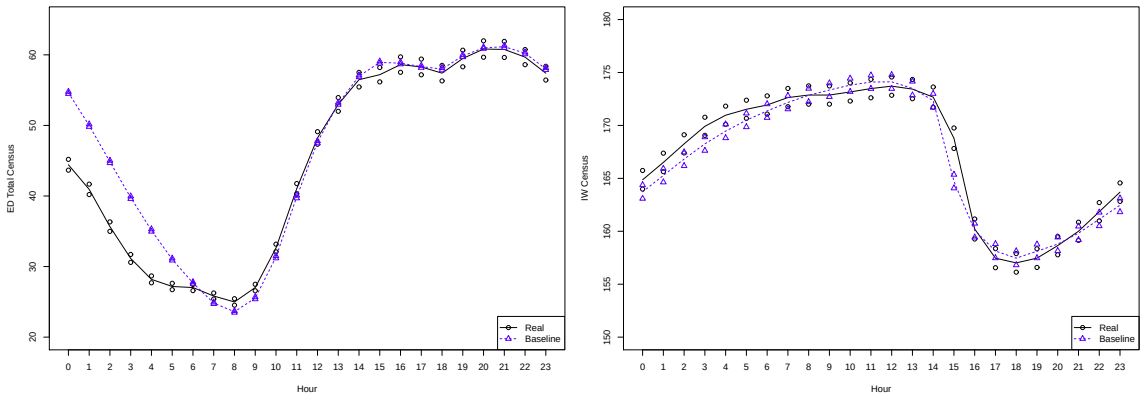
with enough censuses. The average performance measure are computed for each batch, including three patient-level time variables: ED boarding time, nonED-to-IW waiting time, and IW-bed idle time, and four system-level census variables: ED total census, IW census, ED-to-IW census, nonED-to-IW census.

For Policy 1, we test eight different values of the probability threshold ($p^* \in [0, 1]$), to see how the performance measure changes with respect to the choice of p^*. Each p^* makes the fraction of patients who are given the early bed request to be 0, 0.05, 0.1, 0.2, 0.4, 0.6, 0.7, or 1, namely $\mathrm{P}(X > p^*) \in \{0, 0.05, 0.1, 0.2, 0.4, 0.6, 0.7, 1\}$. $\mathrm{P}(X > p^*) = 0$ corresponds to the baseline scenario that no early bed request is implemented; $\mathrm{P}(X > p^*) = 1$ corresponds to the extreme case that early requesting a bed for any ED patient at the beginning of ED service. Because it is very challenging to estimate the real costs of boarding patients and also holding beds. For Policy 2, we test seven different values of the relative cost ($\gamma = c_b/c_p$) of holding one bed (c_b) to boarding one patient (c_p): $\gamma \in \{10, 5, 2, 1, \frac{1}{2}, \frac{1}{5}, \frac{1}{10}\}$.

We first validate that the simulation model can capture what happens in the real hospital, in Section 4.7.1. Then, we summarize all the simulation results and make conclusion how the early bed request affect the operations in the ED+IW subnetwork, in Section 4.7.2.

4.7.1 Validation of Simulation Model

We compare the real hospital data with the simulation results under the baseline scenario, when no early bed request is implemented. To make a thorough validation, we consider the following four system-level census variable: ED total census, IW census, ED-to-IW census, and nonED-to-IW census. To check whether the simulation model can capture the variability within a day, each census variable is collected at the beginning of each hour of a day, and averaged over the whole simulation period.



(a) ED Total Census

(b) IW Census

(c) ED-to-IW Census

(d) NonED-to-IW Census

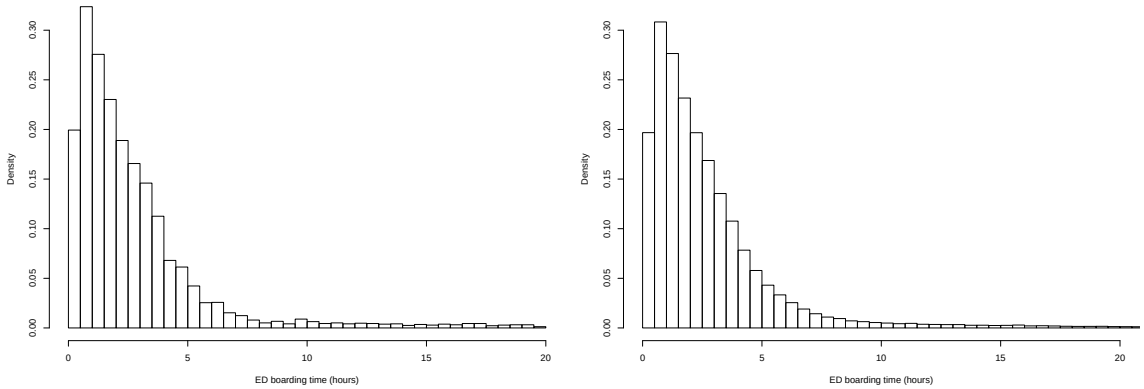
Figure 4.7: Model Validation: Hourly Census Variables

From Panel (a) of Figure 4.7, we observe that the simulation model can resemble the total number of patients in the ED during most hours of the day very well. However, in the early morning from 0 to 5, even though the simulation model can capture the decreasing trend, the simulation results show consistently higher numbers than the real hospital data. It indicates that either the arrival or the discharge process or both are not captured very well in the simulation model. After looking over the real hospital record, we realize that a lot of transfers occur exactly at 23:59, which can be visualized from Panel (a): 55 patients are in the ED service at 23:00, while only around 45 patients are at 0:00. Armony et al. (2011) mentioned that this unreasonable time of discharge was faked for the purpose of insurance reimbursement. In the simulation model, the discharges from ED are continuous, therefore it needs a few hours to get to the same level as the real data.

From Panel (b) of Figure 4.7, we observe that the curves of IW census throughout the day are quite similar in the simulation model and the real hospital. It suggests the arrival and discharge patterns in the IW are captured pretty well.

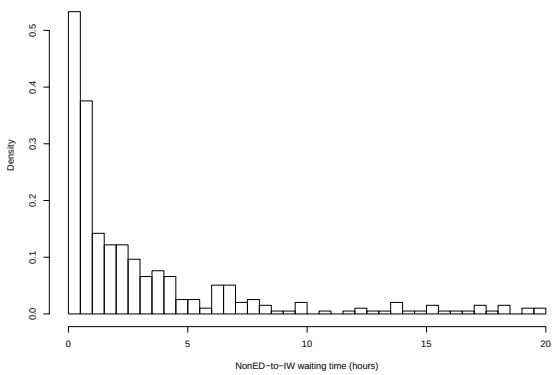
Panel (c) and Panel (d) of Figure 4.7 show the numbers of patients waiting in the ED-to-IW and nonED-to-IW transfer processes, respectively. Generally, the simulation model is able to capture the variability pattern throughout a day for each transfer process. However, the simulation results are consistently overestimated a little bit. The overestimation might be caused by the mechanism of how to make a transfer in the simulation model. First, an available IW bed needs to be seized, then the IW-bed allocation process is started on the seized bed. The times of IW-bed allocation are approximated by ED boarding times in the real hospital data. Therefore, the time waiting for an available IW bed is the reason to the overestimation. In addition, when we estimate the ED boarding times, we model on the log scale first and then take the exponential; a small larger on the log scale will be enlarged extremely after taking the exponential.

In Figure 4.8, we first compare the distribution of the ED boarding time constructed from the real hospital data with that constructed from the simulation results under the baseline scenario, then make a comparison in terms of the nonED-to-IW waiting time distribution. Note that, for each time variable, no data in the real hospital record is greater than 20 hours, however, less than 3% of data in the simulation results is greater than 20

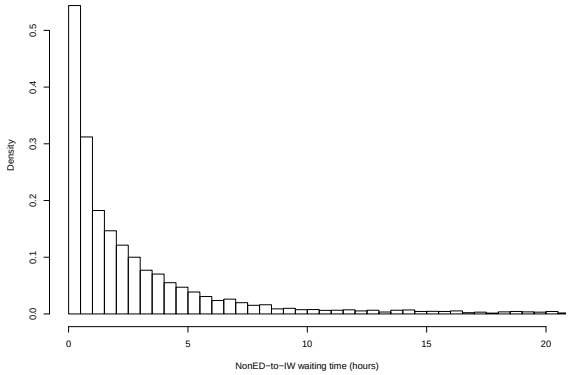


(a.1) Real Data

(a.2) Simulation



(b.1) Real Data



(b.2) Simulation

Figure 4.8: Model Validation: Distribution of ED Boarding Time and NonED-to-IW Waiting Time

hours. To make the comparison more clearly, those density bars corresponding to value beyond 20 hours are truncated for each distribution of the simulation results in Figure 4.8. Basically, the distribution of each time variable from the simulation results and the real data look similar. However, we find some discrepancies, when digging into details. First, the distribution from the simulation results for both variables seem smoother than that from the real data. That is because the size of data in the simulation results are more than 15 times larger than that in the real data. Second, the simulation results have lower density with small values compared with the real data. This finding is consistent with what found from the number of patients in each one of the two transfer processes before.

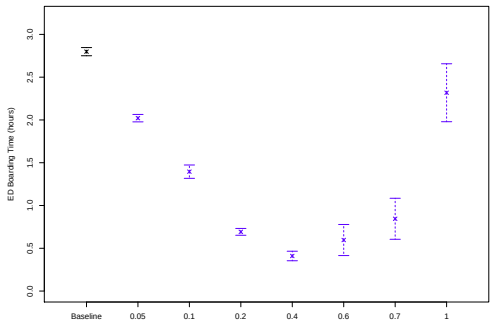
In conclusion, we verify that the simulation model is able to capture the variabilities in the real hospital operations with high fidelity. Therefore, it is meaningful to use the

simulation results derived from this simulation model to make conclusions and comparisons later on.

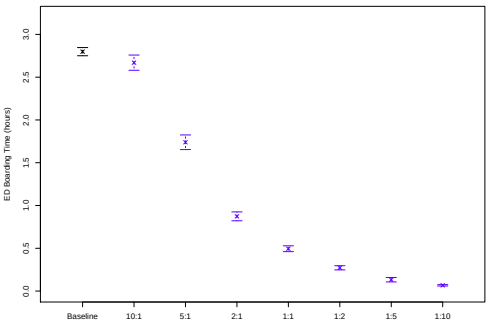
4.7.2 Simulation Results

In this section, we are using plots of 95% CIs of four key performance measures to show the pros and cons of the early bed request, and making comparisons between the two proposed bed request policies. In Figure 4.9, all black CIs belong to the baseline scenario. All blue CIs in the left four panels belong to one of the seven scenarios under Policy 1, which are arranged in the increasing order of the fraction of patients being given early bed requests, from 0.05 to 1. All blue CIs in the right four panels belong to one of the seven scenarios under Policy 2, which are arranged in the decreasing order of the relative cost, from $\frac{10}{1}$ to $\frac{1}{10}$. To be specific, firstly, the cost of boarding one patient is less than the cost of holding one bed, and the difference is getting smaller and smaller; then two costs are the same; in the end, the cost of boarding one patient is greater than the cost of holding one bed, and the difference is getting larger and larger. The label on the x-axis indicates the exact setting of each specific scenario. Note that, the labels for Policy 2 are in the format of $c_b : c_p$, e.g., 1:10 stands $\gamma = c_b/c_p = 1/10$ and means that the cost of boarding one patient is 10 times of the cost of holding one bed.

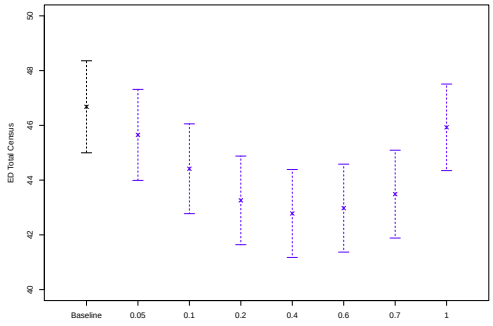
From both the ED boarding time and ED total census shown in the top four panels of Figure 4.9, we can see that all blue CIs are lower than the black CIs. Therefore, we find that early bed request is able to reduce both the ED boarding time and the ED occupancy, no matter which policy is implemented. Next, let's focus on the trend from left to right shown in the these four panels. Based on the arrangement of each policy, we should expect the downward trend for the ED boarding time, as more and more patients tend to be given early bed requests. In addition, the ED occupancy should be lowered at the same time, as the transfer delay is reduced. Policy 2 meets our expectation in both measures. However, for Policy 1, we observe that the ED boarding time decreases when $\mathrm{P}(X > p^*)$ is small, then starts to increase since $\mathrm{P}(X > p^*) = 0.4$. The same behavior is found for the ED occupancy. The reasons are: 1) when an appropriate proportion of patients are given early



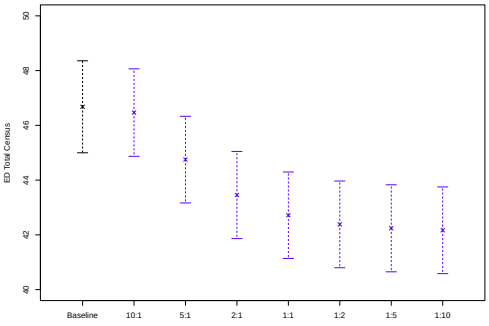
(a.1) ED Boarding Time



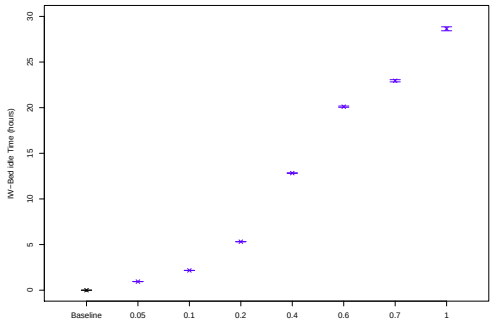
(a.2) ED Boarding Time



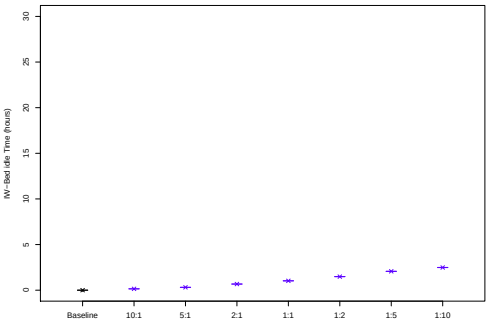
(b.1) ED Total Census



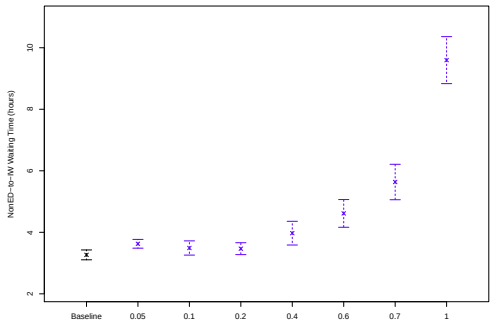
(b.2) ED Total Census



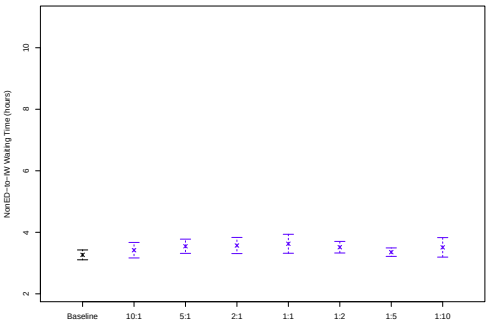
(c.1) IW-Bed Idle Time



(c.2) IW-Bed Idle Time



(d.1) NonED-to-IW Waiting Time



(d.2) NonED-to-IW Waiting Time

Figure 4.9: Simulation Results

bed requests, the inpatient flow process does benefit from the early bed request; 2) however, when the proportion of early bed requests are too large, the available bed capacity in IW is very low. Then, the time to allocate a bed for a new admission is prolonged, in addition, the boarding times of patients who are not given early bed requests are prolonged extremely. Therefore, the benefits of the early bed request are diluted.

The early bed request is not cost-free, as it needs some extra efforts. Next, we present the costs of the early bed request from two points: bed idle time, and waiting time for patients transferred from other than ED, as shown in the bottom four panels of Figure 4.9. The non-decreasing trend shown in each plot tells us that the better transfer performance comes along with longer bed idle time and waiting time for patients transferred from other than ED. For Policy 1, the increasing rate of bed idle time becomes higher and higher, as more and more early bed requests are given, and of which more and more are actually unnecessary as the expected number of admissions remains unchanged. The waiting time for patients transferred from other than ED starts to increase since $\mathrm{P}(X > p^*) = 0.4$, then becomes faster and faster, as too many IW beds are requested for transfers from ED, the available IW capacity for patients transferred from other than ED is very low. However, the increased amount for bed idle time in Policy 2 is much smaller than Policy 1. In addition, for the waiting time of transfers from other than ED, we do not even see significant increase for Policy 2.

Figure 4.10 shows the scatter plot of the mean ED boarding time (y-axis) and the mean IW-bed idle time (x-axis) under each scenario of each policy, where black dots are for Policy 1 and blue triangles are Policy 2. The U-shape of Policy 1 confirm our previous concern that it is too sensitive to the threshold: an aggressive threshold might hurt the overall system: higher costs with even longer boarding time. To make a comparison between two policies, we need to clearly compare how much each policy trades off for the same level of improvement in the patient boarding. The policy that achieves the same patient boarding performance with less trade-off would be better. Hence, the better policy would appear in the lower left of this figure, where locates all the scenarios of Policy 2. Therefore, we conclude that Policy 2 is consistently much better than Policy 1, it can achieve better service quality while causing less trade-off.

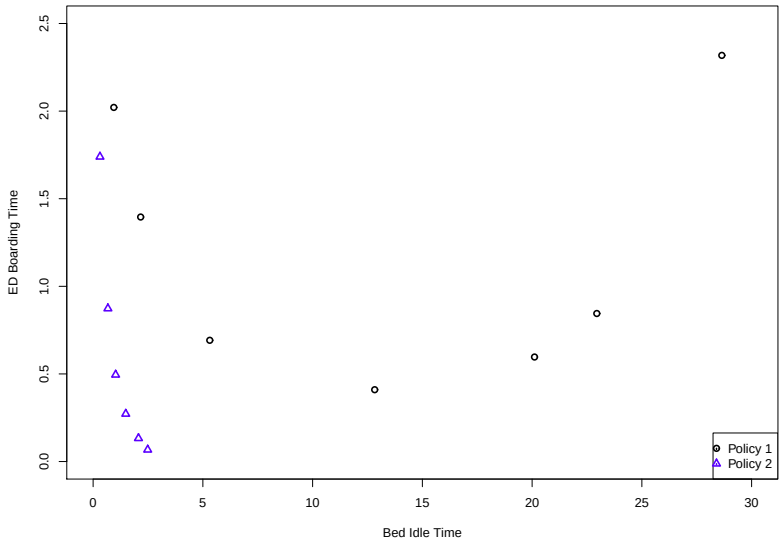


Figure 4.10: ED Boarding Time vs IW-Bed Idle Time

4.8 Conclusion and Discussion

In this chapter, we propose two strategies to integrate the prediction of the IW admission probability with the inpatient bed request process. We build a discrete-event simulation model to capture what happens within the ED+IW subnetwork using a lot of empirical studies of a real hospital data, and validate its rationality. We demonstrate that the early bed request can improve the boarding delay and reduce the ED occupancy as well. At the same time, we illustrate that the early bed request is not cost-free, the better performances come along with longer IW bed idle times and increased delays for transfers from other than ED. Furthermore, we show that it is essential to choose an appropriate probability threshold of the early bed request, and come up with a cost-sensitive model to strike a good balance between the pros and cons of the early bed request. Results from simulation studies are promising and suggest significant reduced ED patient boarding times and cost saving potential.

Though the conclusions come from the simulation model built upon an extensive empirical study of a single hospital, we believe that similar results can be found in other hospitals based on the similarity in many empirical observations between this hospital and others. These proposed policies can be generalized to other hospitals; besides, these insights can

help hospital managers choose among different policies to implement. In addition, based on the benefits and costs of the implementation, hospital administrators can choose the practical probability threshold and relative cost to implement these policies.

Our study has limitations in several aspects. First, the effectiveness of the proposed policies is evaluated in the simulation model built based on empirical studies from a single hospital. Thus, our findings might not be always generalizable to other hospital settings, e.g., the predictive models of the time variables might be fundamentally different from ours (Models (4.19) - (4.22)) in terms of the direction of each independent variable. Second, the better performance of Policy 2 is based on the assumption that all beds in IW have the same function and are interchangeable in this study. Had we included the bed specialization, it would be likely that the improvement resulted from Policy 2 is not that much.

BIBLIOGRAPHY

Abraham, J. and Reddy, M. C. (2010). Challenges to inter-departmental coordination of patient transfers: a workflow perspective. *International journal of medical informatics*, 79(2):112–122.

Armony, M., Israelit, S., Mandelbaum, A., Marmor, Y. N., Tseytlin, Y., and Yom-Tov, G. B. (2011). Patient flow in hospitals: A data-based queueing-science perspective. *Submitted to Stochastic Systems*, 20.

Asplin, B. R., Magid, D. J., Rhodes, K. V., Solberg, L. I., Lurie, N., and Camargo Jr, C. A. (2003). A conceptual model of emergency department crowding. *Annals of emergency medicine*, 42(2):173–180.

Audit Commission (2003). Acute hospital portfolio: bed managementreview of national findings. *London: Audit Commission*.

Bernstein, S. L., Aronsky, D., Duseja, R., Epstein, S., Handel, D., Hwang, U., McCarthy, M., John McConnell, K., Pines, J. M., Rathlev, N., et al. (2009). The effect of emergency department crowding on clinically oriented outcomes. *Academic Emergency Medicine*, 16(1):1–10.

Braunwald, E. (1998). Evolution of the management of acute myocardial infarction: a 20th century saga. *The Lancet*, 352(9142):1771–1774.

Brown, L., Gans, N., Mandelbaum, A., Sakov, A., Shen, H., Zeltyn, S., and Zhao, L. (2005). Statistical analysis of a telephone call center: A queueing-science perspective. *Journal of the American Statistical Association*, 100(469):36–50.

Carson, S. S., Stocking, C., Podsadecki, T., Christenson, J., Pohlman, A., MacRae, S., Jordan, J., Humphrey, H., Siegler, M., and Hall, J. (1996). Effects of organizational change in the medical intensive care unit of a teaching hospital: a comparison of 'open' and 'closed' formats. *Jama*, 276(4):322–328.

Chalfin, D., Cohen, I., and Lambrinos, J. (1995). The economics and cost-effectiveness of critical care medicine. *Intensive care medicine*, 21(11):952–961.

Crawford, E. A., Parikh, P. J., Kong, N., and Thakar, C. V. (2013). Analyzing discharge strategies during acute care a discrete-event simulation study. *Medical Decision Making*, page 0272989X13503500.

Elkin, K. and Rozenberg, N. (2007). Patients flow from the emergency department to the internal wards. *IE&M project, Technion (In Hebrew)*, 4(5.5):3.

Ghorra, S., Reinert, S. E., Cioffi, W., Buczko, G., and Simms, H. H. (1999). Analysis of the effect of conversion from open to closed surgical intensive care unit. *Annals of surgery*, 229(2):163.

Hochman, J. S., Sleeper, L. A., Webb, J. G., Sanborn, T. A., White, H. D., Talley, J. D., Buller, C. E., Jacobs, A. K., Slater, J. N., Col, J., et al. (1999). Early revascularization in acute myocardial infarction complicated by cardiogenic shock. *New England Journal of Medicine*, 341(9):625–634.

Hong, Y. (2013). On computing the distribution function for the poisson binomial distribution. *Computational Statistics & Data Analysis*, 59:41–51.

Hoot, N. R. and Aronsky, D. (2008). Systematic review of emergency department crowding: causes, effects, and solutions. *Annals of emergency medicine*, 52(2):126–136.

Howell, E., Bessman, E., Kravet, S., Kolodner, K., Marshall, R., and Wright, S. (2008). Active bed management by hospitalists and emergency department throughput. *Annals of internal medicine*, 149(11):804–810.

Institute of Medicine (2007). *Hospital-Based Emergency Care: At the Breaking Point*. The National Academies Press, Washington, DC.

Japsen, B. (2003). Hospital capacity debate heats up: Aging population means sharp rise in need, study says. *Chicago Tribune*, 17.

Jernberg, T., Johanson, P., Held, C., Svennblad, B., Lindbäck, J., Wallentin, L., et al. (2011). Association between adoption of evidence-based treatment and survival for patients with st-elevation myocardial infarction. *JAMA*, 305(16):1677–1684.

Katz, J. N., Shah, B. R., Volz, E. M., Horton, J. R., Shaw, L. K., Newby, L. K., Granger, C. B., Mark, D. B., Califf, R. M., and Becker, R. C. (2010). Evolution of the coronary care unit: Clinical characteristics and temporal trends in healthcare delivery and outcomes*. *Critical care medicine*, 38(2):375–381.

Katz, J. N., Turer, A. T., and Becker, R. C. (2007). Cardiology and the critical care crisis: a perspective. *Journal of the American College of Cardiology*, 49(12):1279–1282.

Killip, T. and Kimball, J. T. (1967). Treatment of myocardial infarction in a coronary care unit: a two year experience with 250 patients. *The American journal of cardiology*, 20(4):457–464.

Knaus, W. A., Draper, E. A., Wagner, D. P., and Zimmerman, J. E. (1985). Apache ii: a severity of disease classification system. *Critical care medicine*, 13(10):818–829.

Le Gall, J.-R., Lemeshow, S., and Saulnier, F. (1993). A new simplified acute physiology score (saps ii) based on a european/north american multicenter study. *Jama*, 270(24):2957–2963.

Leegon, J. and Aronsky, D. (2006). Impact of different training strategies on the accuracy of a bayesian network for predicting hospital admission. In *AMIA Annual Symposium Proceedings*, volume 2006, page 474. American Medical Informatics Association.

Li, J., Guo, L., and Handly, N. (2009). Hospital admission prediction using pre-hospital variables. In *Bioinformatics and Biomedicine, 2009. BIBM'09. IEEE International Conference on*, pages 283–286. IEEE.

Morrow, D. A., Fang, J. C., Fintel, D. J., Granger, C. B., Katz, J. N., Kushner, F. G., Kuvin, J. T., Lopez-Sendon, J., McAreavey, D., Nallamothu, B., et al. (2012). Evolution of critical care cardiology: Transformation of the cardiovascular intensive care unit and the emerging need for new medical staffing and training models a scientific statement from the american heart association. *Circulation*, 126(11):1408–1428.

Multz, A. S., Chalfin, D. B., Samson, I. M., Dantzker, D. R., Fein, A. M., Steinberg, H. N., Niederman, M. S., and Scharf, S. M. (1998). A "closed medical intensive care unit (micu) improves resource utilization when compared with an "open micu. *American journal of respiratory and critical care medicine*, 157(5):1468–1473.

National Center for Health Statistics (2013). Health, united states, 2012: With special feature on emergency care.

OGara, P. T., Adams III, J. E., Drazner, M. H., Newby, K., Scirica, B. M., and Sundt III, T. M. (2015). Cocats 4 task force 13: Training in critical care cardiology. *Journal of the American College of Cardiology*, 65(17):1877–1886.

O'Gara, P. T., Kushner, F. G., Ascheim, D. D., Casey, D. E., Chung, M. K., de Lemos, J. A., Ettinger, S. M., Fang, J. C., Fesmire, F. M., Franklin, B. A., et al. (2013). 2013 accf/aha guideline for the management of st-elevation myocardial infarction: a report of the american college of cardiology foundation/american heart association task force on practice guidelines. *Journal of the American College of Cardiology*, 61(4):e78–e140.

Olshaker, J. S. and Rathlev, N. K. (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.

OMalley, R. G., Olenchock, B., Bohula-May, E., Barnett, C., Fintel, D. J., Granger, C. B., Katz, J. N., Kontos, M. C., Kuvin, J. T., Murphy, S. A., et al. (2013). Organization and staffing practices in us cardiac intensive care units: a survey on behalf of the american heart association writing group on the evolution of critical care cardiology. *European Heart Journal: Acute Cardiovascular Care*, 2(1):3–8.

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

Peck, J. S., Benneyan, J. C., Nightingale, D. J., and Gaehde, S. A. (2014). Characterizing the value of predictive analytics in facilitating hospital patient flow. *IIE Transactions on Healthcare Systems Engineering*, 4(3):135–143.

Peck, J. S., Gaehde, S. A., Nightingale, D. J., Gelman, D. Y., Huckins, D. S., Lemons, M. F., Dickson, E. W., and Benneyan, J. C. (2013). Generalizability of a simple approach for predicting hospital admission from an emergency department. *Academic Emergency Medicine*, 20(11):1156–1163.

Plambeck, E., Bayati, M., Ang, E., Kwasnick, S., Aratow, M., et al. (2014). Forecasting emergency department wait times. Technical report.

Porteus, E. L. (2002). *Foundations of stochastic inventory theory*. Stanford University Press.

Powell, E. S., Khare, R. K., Venkatesh, A. K., Van Roo, B. D., Adams, J. G., and Reinhardt, G. (2012). The relationship between inpatient discharge timing and emergency department boarding. *The Journal of emergency medicine*, 42(2):186–196.

Qiu, S., Chinnam, R. B., Murat, A., Batarse, B., Neemuchwala, H., and Jordan, W. (2015). A cost sensitive inpatient bed reservation approach to reduce emergency department boarding times. *Health care management science*, 18(1):67–85.

Rathlev, N. K., Chessare, J., Olshaker, J., Obendorfer, D., Mehta, S. D., Rothenhaus, T., Crespo, S., Magauran, B., Davidson, K., Shemin, R., et al. (2007). Time series analysis of variables associated with daily mean emergency department length of stay. *Annals of emergency medicine*, 49(3):265–271.

Rubino, L., Stahl, L., and Chan, M. (2007). Innovative approach to the aims for improvement: emergency department patient throughput in an impacted urban setting. *The Journal of ambulatory care management*, 30(4):327–337.

Shi, P., Chou, M., Dai, J., Ding, D., and Sim, J. (2014). Models and insights for hospital inpatient operations: Time-dependent ed boarding time. *Management Science*.

Soumerai, S. and Avorn, J. (2001). Crossing the quality chasm: A new health care system for the 21st century.

Sun, B. C., Hsia, R. Y., Weiss, R. E., Zingmond, D., Liang, L.-J., Han, W., McCreath, H., and Asch, S. M. (2013). Effect of emergency department crowding on outcomes of admitted patients. *Annals of emergency medicine*, 61(6):605–611.

Sun, Y., Heng, B. H., Tay, S. Y., and Seow, E. (2011). Predicting hospital admissions at emergency department triage using routine administrative data. *Academic Emergency Medicine*, 18(8):844–850.

Sun, Y., Teow, K. L., Heng, B. H., Ooi, C. K., and Tay, S. Y. (2012). Real-time prediction of waiting time in the emergency department, using quantile regression. *Annals of emergency medicine*, 60(3):299–308.

Thompson, S., Nunez, M., Garfinkel, R., and Dean, M. D. (2009). Efficient short-term allocation and reallocation of patients to floors of a hospital during demand surges. *Operations Research*, 57(2):261–273.

Trzeciak, S. and Rivers, E. (2003). Emergency department overcrowding in the united states: an emerging threat to patient safety and public health. *Emergency medicine journal*, 20(5):402–405.

Vicellio, P., Schneider, S., Asplin, B., Blum, F., Broida, R., Bukata, W., Hill, M., Hoffenberg, S., and Welch, S. (2008). Emergency department crowding: High impact solutions. *Dallas, TX: American College of Emergency Physicians*.

Whitt, W. (1999). Predicting queueing delays. *Management Science*, 45(6):870–888.

Yoon, P., Steiner, I., and Reinhardt, G. (2003). Analysis of factors influencing length of stay in the emergency department. *Cjem*, 5(3):155–161.