1-1-2018

Proactive Coordination In Healthcare Service Systems Through Near Real-Time Analytics

Seung Yup Lee
Wayne State University,

PROACTIVE COORDINATION IN HEALTHCARE SERVICE SYSTEMS THROUGH NEAR REAL-TIME ANALYTICS

by

SEUNG YUP LEE

DISSERTATION

Submitted to the Graduate School

of Wayne State University,

Detroit, Michigan

in partial fulfillment of the requirements

for the degree of

DOCTOR OF PHILOSOPHY

2018

MAJOR: INDUSTRIAL ENGINEERING

Approved By:

____

Advisor Date

____

____

____

____

DEDICATION

I dedicate this dissertation to my family

who has stayed with me throughout the course of my doctoral studies.

Particularly to my thoughtful and sincere wife, Rim.

I must also thank my loving parents and sister and my terrific in-laws

who have helped so much and given me their fullest support, especially in prayer.

Finally, I dedicate my work to the friendship and memory of

my church members of Central Alliance Church of Detroit

who have supported and encouraged my family in Jesus's love.

ACKNOWLEDGEMENTS

In acknowledging my appreciation and thanks, I am most grateful to the Almighty God for His divine guidance and blessings, with which He has brought many wonderful people during the course of my research.

I owe special thanks to my academic advisors, Professors Ratna Babu Chinnam and Evrim Dalkiran, for their advice, encouragement, support and patience during my doctoral studies. Their guidance helped me in all facets of the research and writing of this thesis. I could not have imagined having better mentors for my doctoral research. I would also like to thank the entire faculty and staff of the Department of Industrial and Systems Engineering at Wayne State University, for the support and kindness they have shown me over the years.

My special gratitude also goes to Professors Alper Murat, Qingyu Yang at Wayne State University, Mark Van Oyen at the University of Michigan, and Doctor Seth Krupp at Henry Ford Health System for serving on my dissertation committee and for providing insightful advice and constructive criticism for my work. Their invaluable comments and instruction have broadened and deepened my knowledge of my research topic.

I would also like to thank Professors Ohbyung Kwon at Kyung Hee University, South Korea, and Kyoung-Yun Kim at Wayne State University for their encouragement and support on my academic career and faith.

Finally, I wish to thank my parents who have always provided endless love throughout my life with continuous prayer, and to my in-laws who have supported my family with faithful prayer. My intimate appreciation has been reserved to my wife, Rim, for her love, firm and steady support, and sincere prayers. Last but not the least, I would like to thank the little ones, Yeseong and Yejoon. Thank you.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1: INTRODUCTION

The United States (U.S.) healthcare system is the most expensive in the world. In 2016, the per capita healthcare expenditure in the U.S. was $10,348, more than eight times the world average [100], and hospital care forms more than 30% of aggregate U.S. healthcare spending [68]. However, it is worth noting that the Commonwealth Fund reported in 2014 that the U.S. has the least effective healthcare system among the 11 Organization for Economic Co-operation and Development (OECD) countries selected in the survey [35].

To improve the quality and safety of care, health information technology (HIT) is broadly adopted in hospitals, including EDs, along with lean [18, 13, 55, 4] and six-sigma programs [98, 62, 25, 42]. While electronic health record (EHR) systems form a critical data backbone for the facility, these IT systems are mostly applied for reactive management of care services and are lacking when they come to improving the real-time "operational intelligence" of service networks that promote efficiency and quality of operations in a proactive manner. We need systematic coordination methodologies and tools/platforms that can leverage real-time information combined with predictive analytics to enable proactive coordination, and in turn, enhance the operational intelligence of service networks.

In particular, we leverage operations research and predictive analytics techniques to develop proactive coordination mechanisms and decision methods to improve the operational efficiency of bed management service in the network spanning the emergency department (ED) to inpatient units (IUs) in hospitals, a key component of healthcare in most hospitals [44, 95]. The purpose of this study is to deepen our knowledge on proactive coordination empowered by predictive analytics in dynamic healthcare environments populated by clinically heterogeneous patients with individual information changing throughout ED caregiving processes. To enable proactive coordination for improved resource allocation and patient flow in the ED-IU network, we address two components of modeling/analysis tasks, i.e., the design of coordination mechanisms and the generation of future state information for ED patients, which become two main chapters of this dissertation.

In Chapter 2, we explore the benefits of early task initiation for the service network spanning the emergency department (ED) and inpatient units (IUs) within a hospital. In particular, we investigate the value of proactive inpatient bed request signals from the ED to reduce ED patient boarding. Using data from a major healthcare system, we show that the EDs suffer from severe crowding and boarding not necessarily due to high IU bed occupancy but due to poor coordination of IU bed management activity. The proposed proactive IU bed allocation scheme addresses this coordination requirement without requiring additional staff resources. While the modeling framework is designed based on the inclusion of two analytical requirements, i.e., ED disposition decision prediction and remaining ED length of stay (LoS) estimation, the framework also accounts for imperfect patient disposition predictions and multiple patient sources (besides ED) to IUs. We show that the ED-IU network setting can be modeled as a fork-join queueing system. Unlike typical fork-join queue structures that respond identically to a transition, the proposed system exhibits state-dependent transition behaviors as a function of the types of entities being processed in servers. We characterize the state sets and sequences to facilitate analytical tractability. The proposed proactive bed allocation strategy can lead to significant reductions in bed allocation delay for ED patients (up to $\sim 50\%$), while not increasing delays for other IU admission sources. We also demonstrate that benefits of proactive coordination can be attained even in the absence of highly accurate models for predicting ED patient dispositions. The insights from our models should give confidence to hospital managers in embracing proactive coordination and adaptive work flow technologies enabled by modern health IT systems.

In Chapter 3, we investigate the quantitative modeling that analyzes the patterns of decreasing uncertainty in ED patient disposition decision making throughout the course of ED caregiving processes. Proactive resource allocation (e.g., admissions, transport, environmental service etc.) based on a reliable computerized prediction of ED disposition decisions could help to significantly reduce boarding delay as demonstrated in Chapter 2. The classification task of ED disposition decision prediction can be evaluated as a hierarchical classification problem, while dealing with temporal evolution and buildup of clinical information throughout the ED caregiving processes. We focus on different time intervals within the ED course (registration, triage, first lab/imaging orders, and first lab/imaging results). The study took place at an academic urban level 1 trauma center with an annual census of 100,000. Data for the modeling was extracted from all ED visits between May 2014 and April 2016. Both a hierarchical disposition class structure and a progressive prediction modeling approach are introduced and combined to fully facilitate the operationalization of prediction results. Multinomial logistic regression models are built for carrying out the predictions under three different classification group structures: (1) discharge vs. admission, (2) discharge vs. observation unit vs. inpatient unit, and (3) discharge vs. observation unit vs. general practice unit vs. telemetry unit vs. intensive care unit. We characterize how the accumulation of clinical information for ED patients throughout the ED caregiving processes can help improve prediction results for the three-different class groups. Each class group can enable and contribute to unique proactive coordination strategies according to the obtained future state information and prediction quality, to enhance the quality of care and operational efficiency around the ED. In general, classification models that predict target disposition units can provide more actionable information than a binary admission prediction model. For example, prediction results from class group (3) would be more useful than those from class group (1), while making correct predictions becomes more challenging from class group (1) through (3). We also reveal that for different disposition classes, the prediction quality evolution behaves in its own unique way according to the gain of relevant information. Therefore, prediction and resource allocation and task assignment strategies can be tailored to suit the unique behavior of the progressive information accumulation for the different classes of patients as a function of their destination beyond the ED.

CHAPTER 2: PROACTIVE COORDINATION BETWEEN ED AND INPATIENT UNITS TO REDUCE PATIENT BOARDING

2.1 Introduction

Operations in healthcare facilities entail complex interactions between patients, care providers and resources. In order to improve the quality and safety of care, health information technology (HIT) is being broadly adopted in hospitals along with lean and six-sigma programs. By 2015, over 84% of hospitals adopted an EHR system in the U.S., an nine fold increase since 2008 [50]. While EHR systems form a critical data backbone for the facility, we need improved 'work-flow' coordination tools and platforms that can enhance real-time situational awareness and facilitate effective management of resources for enhanced and efficient care [56]. In this paper, we explore the benefits of proactive coordination methods in the form of early task initiation for the service network spanning the emergency department (ED) and associated inpatient units (IUs). The significance of ED-IU network is three-fold. First, ED is a major gateway to hospitals, accounting for more than 50% of inpatient admissions in the U.S [1]. Second, growing ED patient 'crowding' in recent years has been called an international crisis and has received significant public and academic attention [72, 20, 83, 52, 26]. Finally, ED crowding is known to result in adverse outcomes such as patient treatment delays and dissatisfaction [66, 64, 81], patient mortality [94, 17], patients leaving without receiving care [87, 75, 82], ambulance diversion [23, 91, 40], hospital financial losses [80, 71, 16], and harm to staff [54, 86]. Given that ED overcrowding affects all three aspects of hospital performance (clinical, operational and financial), developing effective coordination methods to streamline patient flow across the ED-IU network is critical.

The literature focusing on the means to alleviate ED crowding and improve patient flow has been growing. While Hopp and Lovejoy grouped the key management issues in the ED into three broad categories, i.e., long-term supply capacity expansion, variability control between supply and demand, and patient sequencing [53], counter to the general expectation that capacity expansion will reduce ED congestion, Han *et al.* and Mumma *et al.* have concluded that ED bed capacity expansion does not significantly influence patient crowding [48, 73]. Moreover, capacity expansion is often impractical due to financial and space constraints. Fortunately, other approaches that attempt to control variability between supply-demand and improve patient sequencing strategies have shown great promise [88].

Staffing optimization (physicians, nurses, lab techs) has been seen to be an important topic in improving ED patient length-of-stay (LoS) and census [3, 47]. Wiler *et al.* evaluated interventions, such as immediate bedding, bedside registration, advanced triage, physician at triage, and fast-track service lines, to streamline the front-end processing of ED patients [101]. Xu and Chan proposed that the prediction of ED patient arrivals can help proactively divert patients before ED becomes severely congested [102]. Eitel *et al.* discuss different methods for improving ED quality and flow, including demand management, critical pathways, and process mapping [41]. Saghafian *et al.* suggested an analysis framework for a patient 'streaming' strategy based on queueing network analysis to improve ED operations [89]. Shi et al. explored interactions between IU patient discharge practices and ED boarding [92]. Also, lean and six-sigma management methods have been exploited successfully for improving ED congestion [11]. Several studies have also focused on the use of discrete event simulation as a means for ED operations improvement [70, 65, 76].

A conceptual model partitioned the causes of ED crowding into three interdependent components: input, throughput, and output [8]. Meanwhile, research has shown that output factors, i.e. factors preventing timely transfer of emergency patients to inpatient beds, significantly contribute to ED crowding [97, 43, 92], the main investigation topic of this study. In particular, Abraham and Reddy show that ineffective inter-departmental interactions and information handoffs are the two predominant challenges that inhibit smooth patient transfers between departments [2].

Research has suggested that if the hospital admissions of ED patients can be predicted early during triage and communicated to different departments of a hospital, then necessary steps can be taken early to reduce transfer delays [79], the primary focus of this paper. Batt and Terwiesch also discuss a similar approach, early task initiation where certain downstream stage tasks can be initiated earlier than their normal start times by upstream stage servers within the ED [15]. However, unlike the approach that Batt and Terwiesch take, in our coordination strategy, the upstream stage (ED) guides the downstream stages (IUs and support services) to initiate their own tasks early, when deemed appropriate based on future state predictions.

In this chapter, we explore proactive coordination signals between ED and the inpatient units for patients likely to be admitted to the hospital beyond ED treatment, so as to reduce transfer and 'boarding' delays. Boarding refers to situations where patients to be admitted are held-up in the ED, past their ED treatment, utilizing critical and expensive resources while waiting for IU beds to be allocated and prepared. U.S. Center for Disease Control (CDC) reports that 62% of hospitals experience boarding patients for two hours or more at some point in the year [28]. The boarding time based on a survey of 1,195 U.S. EDs accounted for about 37% of the time an admitted patient spent in an ED on average [9], and the latest data from the Center for Medicare & Medicaid Services (CMS) report that patients in U.S. EDs are experiencing median boarding time of 2 hours and 16 minutes, with the longest delay of 4 hours and 26 minutes, in the District of Columbia [31]. It is also estimated that EDs experiencing severe boarding lose conservatively $15,500 a day of revenue compared to an average ED. This accounts for both direct losses from a significantly higher number of patients simply walking away or diverted to other hospitals due to long wait times as well as from lost admissions to IUs [10]. To bring more attention to ED crowding, the CMS is requiring hospitals in the U.S. to report length of boarding delay as an ED crowding measure since 2014 [33].

To reduce boarding through early task initiation, our coordination framework relies on prediction models that can estimate the ED patient's remaining length-of-stay (LoS) at different stages of the ED treatment process as well as models that can predict patient's likely admission to a specific IU, i.e., disposition decision prediction. There is a growing body of literature on ED LoS modeling [61, 27, 30] and admission predictions [78, 84, 14]. As for LoS predictions, [30] explored fifteen factors statistically associated with ED LoS and showed the predictive validity of a multivariate accelerated failure time model. As for disposition decision prediction, [14] have attained an accuracy of 91.23% and a sensitivity of 94.35% within an hour of patient arrival to the ED, while limiting the false-positive rate to 10%. Our analysis reveals that these performances are more than adequate to attain significant improvements in ED boarding under the proposed proactive IU bed allocation scheme.

From a tactical standpoint, Qiu *et al.* suggest a cost sensitive inpatient bed 'reservation' policy to reduce ED boarding time. The policy recommends an optimal bed reservation time slot based on a modified newsvendor model to minimize the cost of patient waiting and bed wastage [84]. They assume that the combined lead-time for IU bed preparation, assignment, and patient transfer is deterministic and known. In addition, their approach assumes independence in handling the work-flows for individual patients and does not explicitly account for ED-IU network level performance. We rely on a fork-join queueing network representation to effectively model the ED-IU network and analyze the performance of the proposed work-flow coordination scheme utilizing these predictive outcomes. The proposed scheme is practical and can be operationalized in hospitals without having to require significant changes to current service practices.

2.1.1 Contributions

This work allows us to make the following contributions:

1. We present a fork-join queue structure model for representing and analyzing the proposed proactive IU bed allocation scheme within an ED-IU network. This structure necessarily incorporates potential cancellation of bed preparation tasks that cause state-dependent behaviors. Instead of assuming that all servers and queues respond identically to all events (typical of fork-join queue studies), our original model entails more realistic transitions between states depending on system states. Our model successfully manages bed preparation effort and reduces the boarding time of ED patients by proactively preparing beds for the patients identified as likely to be admitted.

2. We characterize the benefits from employing the proposed patient flow coordination scheme across the ED-IU network through sensitivity analysis under specific combinations of bed request signal lead-time, patient admission rates, and bed preparation service rates. We show that under a reasonable ED-IU network setting with a bed preparation server having an exponentially distributed service time with a mean of 60 minutes, proactive IU bed request signals issued 90 minutes ahead of actual ED physician disposition decisions can attain on average 53% reduction in inpatient bed allocation delays. We also explicitly quantify the impact of disposition prediction capability on boarding

delay reduction through our analysis.

Key insights resulting from our study are the following: 1) EDs suffer from severe crowding and boarding in the afternoons and evenings not necessarily due to high IU bed occupancy but due to poor coordination of IU bed management activity, 2) There are significant portions of the day when IU patient discharge pattern does not influence the ED-IU network flows, allowing proactive IU bed allocations without having to explicitly consider IU occupancy, 3) Proactive bed request signals can lead to significant reductions in ED patient boarding, 4) Proactive bed preparation for ED patients does not compromise delays for other IU patient sources, 5) Benefits of proactive coordination can be attained even in the absence of highly accurate disposition prediction models, 6) Proposed coordination scheme effectively deals with Type I and Type II disposition prediction errors and does not increase the long-run utilization of bed preparation servers even under false positive predictions, and 7) Improved disposition prediction quality leads to greater benefits in boarding delay reduction as servers operate with higher utilization.

2.2 Clinical Setting & Observations

This section briefly introduces the operations across the ED-IU network in an academic urban level 1 trauma center by analyzing data gathered over multiple years (May 1, 2014 – December 15, 2016), constituting 243,745 ED visits and 41,942 admissions. Table 1 summarizes descriptive statistics of the operational facts around the ED-IU network in the hospital. This ED receives 10.6 patients per hour on average and has 108 beds in total (including 31 beds that can be placed in hallways during peak periods). Figure 1 illustrates the general patient process flow within the ED-IU network. After a certain level of observation and treatment for a patient, an ED physician decides whether the patient should be admitted for further care

Table 1: Summary Statistics of Supply and Demand around the ED-IU Network

(a) Capacity	Counts
Number of IUs / only GPUs	26 / 10
Number of beds in all IUs / only GPUs	758 / 360
Number of beds in ED rooms	77
Number of beds that ED hallways can accommodate	31

(b) Demand (on weekdays)	10%ile / Mean / 90%ile	
	Midnight to noon	Noon to midnight
ED bed occupancy	31.0 / 47.1 / 64.0	56.0 / 71.0 / 85.0
Number of boarded ED patients	3.0 / 8.5 / 15.0	6.0 / 12.7 / 21.0
Number of ED dispositions to IUs per hour	0.0 / 1.2 / 3.0	1.0 / 2.5 / 5.0
Number of total admissions to IUs per hour	1.0 / 3.2 / 6.0	3.0 / 5.9 / 9.0

or discharged. If admission is necessary, the physician determines the IU most clinically appropriate for the patient based on the diagnosis (labeled 'disposition' decision). A request is then sent to the bed management team for identifying a bed for the patient within the target IU. If there is a bed ready for use upon an admission, the bed allocation process is completed shortly. However, if a bed is still in the process of preparation, it takes longer to reach patient-bed assignment. Once a bed has been assigned, the patient is physically transferred to the IU.

Like most EDs, the ED in this hospital suffers from severe boarding delays. In the studied period, a typical admitted ED patient spent on the average 4 hours and 39 minutes receiving

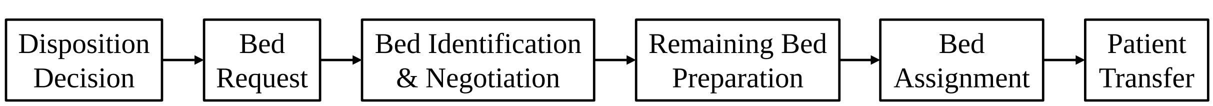


Figure 1: Typical Process for Admitting Patients from ED to IUs

Table 2: Process Intervals between Inpatient Unit Bed Request and Patient Departure from ED

Process intervals	Process duration (minutes)		
	10%ile	Mean	90%ile
ED patient arrival to ED disposition decision	109	278.7	482
Disposition decision to IU admission approval	17	75.8	163
Admission approval to IU bed assignment	6	102.2	269
Bed assignment to patient departure from ED	30	80.1	144

care within the ED and another 3 hours and 2 minutes undergoing additional boarding delay after admission approval. The facility is not an outlier in relation to boarding delays experienced by EDs across the U.S. [31]. Table 2 reports interval duration statistics for the different stages of the admission process. This hospital also utilizes an admission approval process from patient's health insurance provider, accounting for an additional 1 hour and 16 minutes of boarding delay on average. In this study, we target the reduction of admission approval to IU bed assignment delay (102.2 minutes long on average), which is labeled as bed allocation delay (BAD) in this paper.

The ED bed occupancy rate of a hospital has often been used as a simple yet reasonable measurement for ED crowding status [69, 63]. Similar to [6], Figure 2 reports the average ED census patterns for the case study hospital, i.e., occupancy rate, by time of day on weekdays. For instance, at 10AM, there have been 104 days where the ED occupancy was between 31 and 35 patients during the data collection period (686 weekdays), yielding a frequency of 0.15. Overall, the ED census can be categorized into four distinct stages according to its behavior: a) steady low census in the morning between 7–10AM; b) increasing census between 10AM–4PM; c) steady high census in the evening between 4–11PM; and d) decreasing census between 11PM–6AM. Obviously, ED census is a function of both ED arrival and departure rates and is influenced by ED service and boarding rates experienced at different times of the day. For example, during the decreasing census stage (d), the combined rate of IU admission and ED discharge exceeds the rate of ED arrivals. We can observe that there have not been many weekdays where the ED census exceeds 76 patients (full of regular beds or 70% bed occupancy, with $<$ 10% relative frequency) between 1AM–1PM. Indicating that patient boarding, even if present during this window, is not significant for ED has additional beds to serve incoming patients. However, from 2PM–Midnight, we see that ED reaches higher levels of occupancy, exhausting all beds on occasion and potentially suffering from the side effects of boarding. [103] and others suggest that the incidence of serious complications increases significantly for boarded patients as EDs reach high levels of occupancy.

The ED patient arrival rate is mostly outside the control of the hospital, and hence, is un-

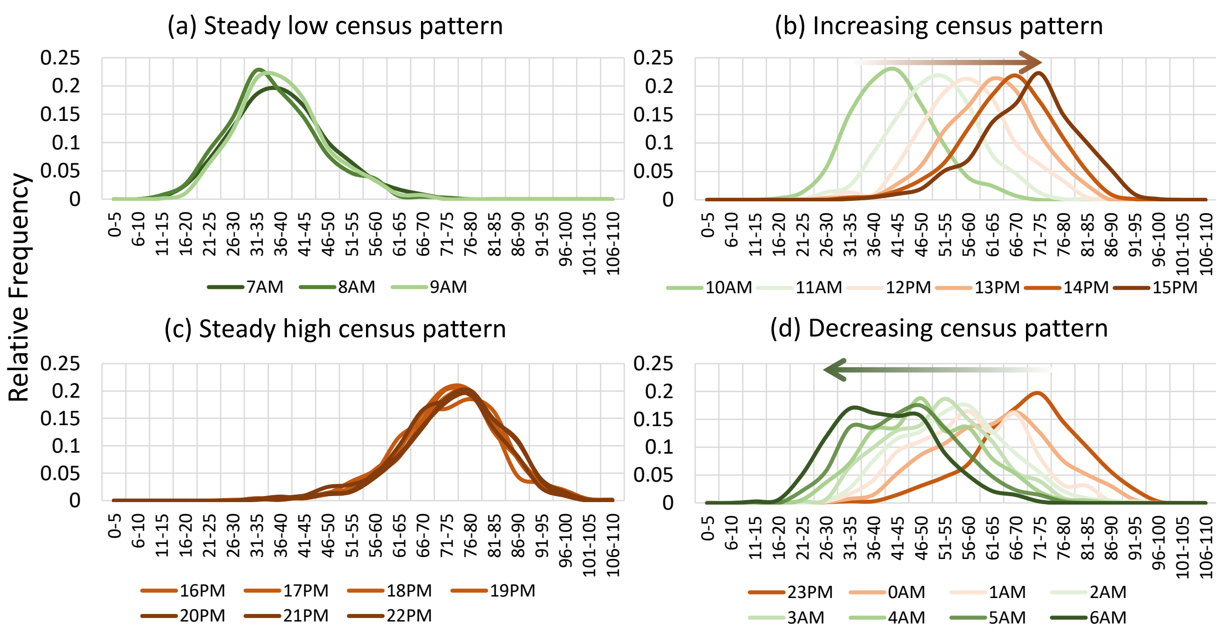


Figure 2: ED Patient Census by Hour of the Day during Weekdays

controllable. However, even if we were to take the ED service and the arrival rates as given, there are periods within a day where the boarded patient census is somewhat controllable due to increased outflow rate via proactive bed preparation (even without changing any patient treatment or observation procedures/protocols in the ED). Figure 3(a) reports the average number of ED patients that experience different levels of BAD at each hour of the day. The delays peak between noon–10PM, with a sharp drop as we increase the threshold for BAD (notice the large gaps between the lines during this time range). Note that during late evening and early morning hours, the number of patients experiencing delays does not drop easily even when we consider BAD limits reaching 10 hours. This can be partially attributed to lack of unoccupied beds in IUs for patient admission. Fortunately, as can be seen from Figure 2, the boarding delays during these periods do not lead to excessive ED crowding as there are far fewer ED patient arrivals compared to ED discharges during these hours. On the contrary, while many more patients are experiencing delays during the day and early evenings, relatively few patients are experiencing excessively long boarding delays. This is because IUs experience more discharges during the day, replenishing bed supply for incoming patients from the ED and elsewhere. Boarding delays during afternoons and evenings contribute to severe ED crowding and should be seen as being “more harmful" to ED operations. Fortunately, these boarding delays are not due to IU bed shortages and can be alleviated through proactive IU bed allocations and coordination, which is the primary focus of this manuscript. In summary, not all of the boarding delays contribute to overcrowding in the ED. Rather, it is the boardings that occur during a specific time range (between 2PM–midnight in our ED) that severely deteriorate overcrowding in the ED and can be controlled through more effective and proactive coordination. This pattern is typical of most hospitals across the U.S. as it is a common prac-

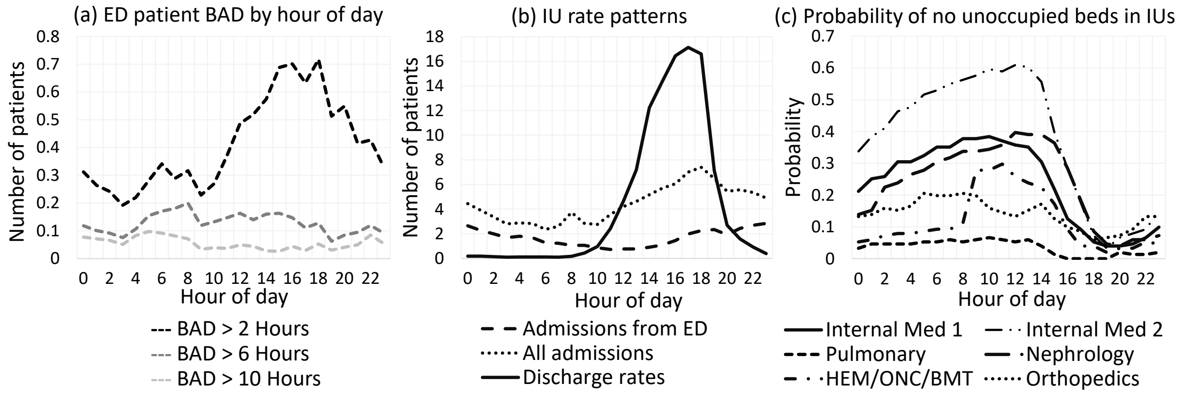


Figure 3: ED Patient Bed Allocation Delay (BAD) and IU rate and occupancy patterns by Hour of the Day

tice in the industry to discharge most of the IU patients around noon after being examined by providers during their “morning rounds" [99].

Figure 3(b) provides clear evidence for the availability of unoccupied IU beds during the time that the ED suffers from severe crowding. This is quite important for much of the extant literature often assumes that boarding is simply a result of IU bed shortage. Figure 3(b) clearly shows that from noon and beyond, the number of beds becoming unoccupied exceeds the number of beds required by admissions, supporting the arguments made based on Figure 3(a). The directors for the ED at this hospital also believe that this is representative of most EDs across the country. However, due to the uncertainty in admission rates to different units, it is difficult for the servers to prioritize their bed preparation tasks on their own. The added complication is that they are responsible for other duties besides bed cleaning. This is discussed in more detail in Section 2.3. Figure 3(c) provides strong evidence that there is an opportunity to improve bed allocation processes for many beds that become unoccupied after about 2PM. It plots the probability that there are no unoccupied beds in IUs. The term “no beds" is defined to represent the inpatient bed status where there is only 0 or 1 unoccupied beds in the IU. We conservatively included the 1 unoccupied bed case into the “no beds" category to account for any possibility that a bed is being held or blocked due to infection concerns or is temporally unavailable. The different IUs show their own occupancy trends but share the common feature after about 2PM, i.e., the increasing availability of beds. Moreover, considering routine overflow allowed among clinically similar IUs, the probability of unoccupied bed availability is even higher in reality.

2.3 Proactive Coordination of Inpatient Bed Allocation Process

In this chapter, we propose managing inpatient bed preparation servers in IUs through proactive bed request signals during periods of ED congestion so that bed allocation processes can be completed near “just-in-time" to ED patient disposition decisions to reduce boarding delays. Rest of the time, i.e., besides the peak discharge hours, bed cleaning/allocation processes revert to their normal routines where beds are cleaned as soon as they become unoccupied. Like most hospitals, the case study hospital operates environmental service (EVS) teams that are in charge of cleaning inpatient beds as well as area decontamination, hygiene management, and managing linen/paper and other supplies for the area. Generally, a group of servers are assigned to a set of adjacent IUs that share similar bed features (i.e., partially pooled). During the large volume of IU discharges in the early hours of the afternoon (see Figure 3(b)), servers lag behind in cleaning beds. Unable to clean beds as soon as they become unoccupied, they end up cleaning IU beds according to bed demand at the moment, generating queue lines and resulting in prolonged (more than 2 hours) BAD as shown in Figure 3(a).

While one might consider adding more EVS staff to cope with the sharp peak demand for bed cleaning that emerges in the afternoon, it has been shown that the annual cost associated with such a strategy could exceed several million dollars for the case study hospital and is not

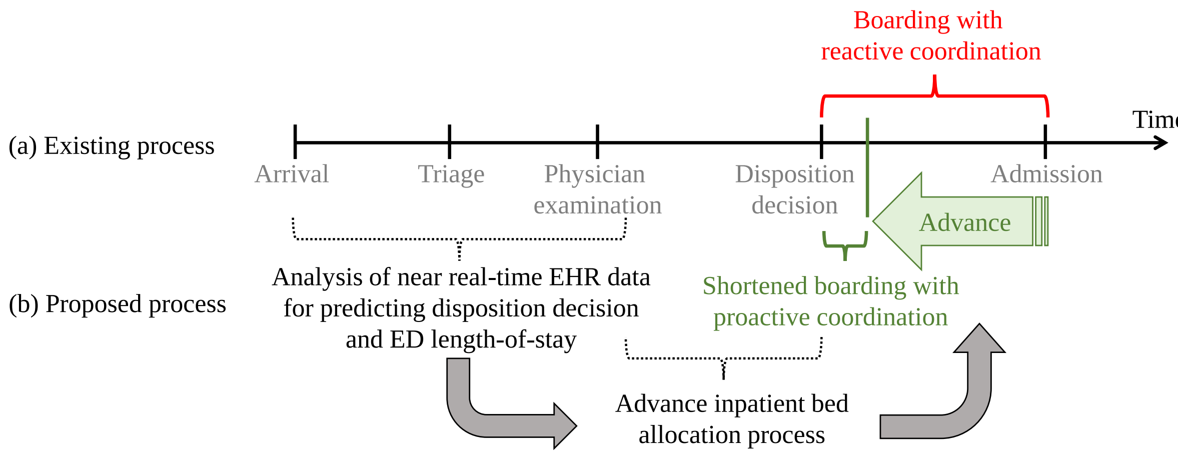


Figure 4: Proactive IU Bed Allocation Strategy for Reducing Boarding Delay

cost-effective. Instead, the requirement during this window with a large number of patient discharges is to proactively prioritize the cleaning of the beds to those IUs facing immediate or projected demand from the ED and other sources. We demonstrate that if servers could only receive advance bed request signals for likely admissions at each IU with adequate lead-time, they would be able to better plan and manage their work to drastically reduce ED patient boarding, and in turn, ED crowding. The hospital already employs communication tools and platforms (e.g., EPIC EHR system combined with pagers between EVS servers and bed managers) that can be leveraged for operationalizing these proactive signals.

Figure 4 illustrates our strategy for enabling proactive bed allocation by exploiting EHR data to predict future state information of ED patients. The key assumption is that once a patient enters an ED and starts undergoing triage, testing (laboratory work and imaging), and care, there is adequate and growing information for the patient within the EHR (including prior health history) to allow reliable prediction of ED disposition decision and remaining ED LoS well ahead of the final disposition decision, to proactively signal the relevant IU regarding an impending admission and need for a bed. As noted in Section 2.1, extant literature offers effective algorithms and techniques for addressing both requirements. In this section, we further investigate key interactions among the different entities within the ED-IU network so as to generate an effective and realistic proactive bed allocation scheme for reducing boarding delays.

2.3.1 Patient Flow Between ED and Inpatient Units

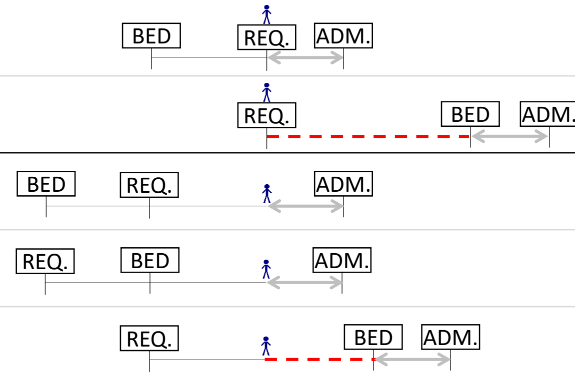
While an IU bed request signal is initiated along with the actual disposition decision of an ED patient within the traditional bed allocation scheme, with proactive bed allocation, a bed request signal is passed to an IU before the physician arrives at a disposition decision for the patient. The following list, in combination with Figure 5, summarizes the different scenarios that can be expected depending on the bed allocation strategies and situations.

1. *Reactive bed request with available beds*: A bed request is sent to an IU as soon as a disposition decision for a patient is made, and there is already a bed available for the patient. In this case, there is no bed allocation delay for the patient.

2. *Reactive bed request with no bed immediately available for allocation*: This is the most common but undesirable situation in the ED-IU network under the current bed allocation scheme. At worst, patient can experience a boarding delay equivalent to the full lead-time associated with bed preparation as well as waiting time in the queue. The red line in Figure 5(2) is representative of the length of these delays.

3. *Proactive bed request with available beds*: As depicted in Figure 5(3), a bed request is sent before patient's disposition decision. Right after the disposition decision, patient can be transferred to the IU and experiences no bed allocation delay as in scenario (1).

Scenarios		Event orders	Illustration
Reactive:	(1)	B – R(P) – A	
	(2)	R(P) – B – A	
Proactive:	(3)	B – R – P – A	
	(4)	R – B – P – A	
	(5)	R – P – B – A	



Event Notation: B – Bed becomes available; R – Bed is requested; P – Physician's disposition decision; A – Admission to IU complete

Figure 5: Delays in Bed Allocation Process Under Different Strategies and Event Orders

4. *Proactive bed request where bed is not readily available but will become available by disposition decision time*: In this scenario, a patient could experience significantly reduced boarding delay thanks to an advance bed request compared to the current bed allocation scheme.

5. *Proactive bed request where bed becomes available after final disposition decision*: This is another case where a positive impact of a proactive bed request is expected. While a bed is still not available when a disposition decision is made, since a bed request has been sent before the disposition decision, there is a good chance that boarding delay would be reduced.

2.3.2 Scheme of Proactive Inpatient Bed Allocation

The proactive inpatient bed allocation process scheme facilitates ED care providers to send an advance bed request signal to an IU or a bed preparation server based on the future state information predicted for the patient. Figure 6 depicts the underlying concurrent process of two different server systems of the scheme. The fundamental idea of the proposed scheme is that right after sending a bed request signal to an IU, the flow forks into two different routes, which are remaining ED processes (server s_1) and bed preparation/allocation processes (server s_2). A bed assignment is made at server s_3 when both a patient and a bed are ready at each queue at server s_3. More specifically, this scheme is proposed based on the following three main considerations:

1. *Impact of Bed Request Signal Lead-Time*: In this scheme, the ED LoS estimation model is continuously updating its predictions as ED patients go through more care steps. Given recent advances in LoS modeling techniques, we make a reasonable assumption that a model without bias is available, and we allow a rather large prediction coefficient of variation of 1. In particular, we assume that the remaining LoS follows an exponential distribution with a known mean (μ_1 time units). We allow IU bed managers to start bed allocation processes in a first-in-first-out (FIFO) manner as soon as they receive a proactive bed request from the ED. Under this setting, it is expected that the proposed scheme in Figure 6 can provide insights into the impact of the extent of proactivity in bed allocation, i.e., bed request signal lead-time, on boarding delay reduction. This in turn helps identify the optimal signal lead-time as a function of service rate of the bed preparation server s_2 and admission rate to the IU. Server system s_1 is a conceptual system that includes all ED processes remaining for a patient after sending an early bed request, i.e., representing bed request signal lead-time, and is not physically separated from the ED.

2. *Indistinguishable Beds within an IU*: While a hospital usually groups beds into multiple distinct IUs, the beds are homogeneous within any specific IU or over a set of IUs allow-

ing overflow. Moreover, any special equipment or accessories (e.g., oxygen) necessary for caring a patient are also common within an IU. So, there is no need to reserve a prepared IU bed for any particular ED patient within IUs that share same bed features. Rather, a prepared bed is ready to be occupied by any ED patient who comes first even if an early bed request is not sent for the patient. This makes the proposed unit-dedicated reservation scheme more flexible than patient-dedicated reservations that would be found inefficient in real hospital settings. Hence, the bed assignment process at server s_3 is modeled as a FIFO process.

3. *Significant Uncleaned Inpatient Beds:* As discussed earlier in this section, unlike the common expectation that discharged IU beds are immediately cleaned, many discharged IU beds remain uncleaned for few hours during the peak discharge period. It is mainly attributable to the surge of discharges within a short period (three to four hours). In turn, this practice makes bed preparation process to be reactive for many admissions, and the 'delayed bed turn around issue' has been a significant concern in the hospital. The bed preparation service time is shown to approximate an exponential distribution (see Section 2.8.1). The model takes μ_2 to represent the mean bed preparation time at server s_2.

In this model, there is no need to account for ED processes carried out before the advance bed request signal is initiated (no matter how complex and different they are for patients). Assuming that the set of ED processes prior to bed request signals are operating in a stable system, it is reasonable to assume that the bed request signals follow a Poisson process. The remaining ED processes are represented as a ∞-server system (server s_1). The justification for the infinite size of servers is that even if a patient would suffer from not receiving prompt

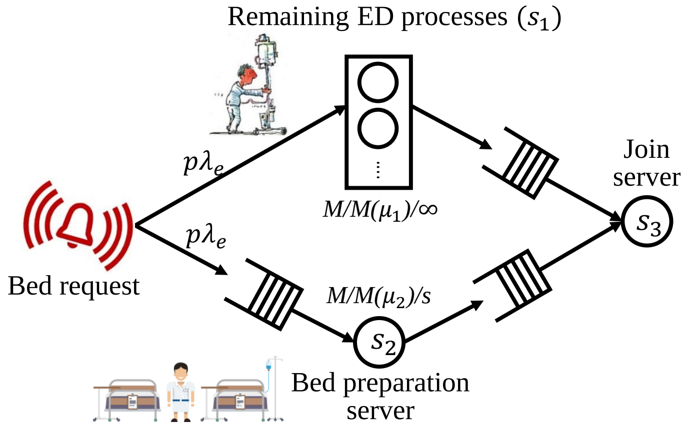


Figure 6: General Scheme for Proactive Inpatient Bed Request and Allocation

treatment due to the temporal lack of care providers, it can still be regarded as one being served with longer service time at server s_1, which can already be calculated by remaining LoS estimation models. So all of the patients in s_1 are regarded "being served".

There are s bed cleaning/preparation servers assigned to a certain section in the hospital. After finishing both services at s_1 and s_2, an ED patient with an actual disposition decision and a prepared bed are merged at server s_3. The joining process is viewed as the patient–bed assignment process. Due to difference in mean service times of the two servers, their configurations, and stochasticity, there would be gaps in the timings of the service completions at server s_1 and s_2. Let $N_1(t)$ denote the number of patients who have completed services at server s_1 and $N_2(t)$ represent the number of beds that have been prepared for patients at time t. Then, the lengths of the two queues at server s_3 can be specified as follows:

$$L_Q^{PT}(t) = [N_1(t) - N_2(t)]^+ \text{ and } L_Q^{BD}(t) = [N_2(t) - N_1(t)]^+ \, , \tag{2.1}$$

where L_Q^{PT} and L_Q^{BD} correspond to the size of the queues for patients and beds, respectively.

This network can be seen as an example of a fork-join queue structure. Within any type of a fork-join queue system, once one operation is finished, the outcome is stacked in a queue waiting for the other operations for the other components, and then the jobs synchronize before they leave for a next operation. However, the proposed network is unique in that the configuration of servers in the two server systems are different as well as their service times. Moreover, the baseline model (Figure 6) is modified significantly by incorporating multiple arrival sources and queue management rules, which make the network close to a real-world setting. One of the main objectives of this study is to model and analyze the expected values of $L_Q^{PT}(t)$ as $t \to \infty$, depending on several parameters.

2.4 Modeling Framework

This section introduces additional modeling details that further shape the baseline model to account for realities of ED-IU network operations. First, IUs serve patients not only being admitted from the ED but also patients arriving from other admission sources (including out-patients, transfers from other hospitals, and transfer from different departments within the hospital). Second, patient disposition predictions are error prone. Bed preparation/allocation operation rules should be able to effectively handle the unexpected status of the network resulted by errors. Relationships among the different types of patients and beds in the ED-IU network are translated into state-dependent transitions among states representing the network, and a continuous-time Markov chain is built to mimic the overall behavior of the network. We limit proactive IU bed allocations to ED patients alone according to the primary aim of this study, and future research can consider extending the framework to entertain proactive IU bed allocations for other sources.

2.4.1 Other Patient Sources for IU

For differentiating patients coming from different sources, let Type **E** and **O** denote ED patients and non-ED patients, respectively. Without loss of generality, we treat patients being admitted from all other sources besides ED as a single type of patients, i.e., **O**. This forms two demand types of patients for IUs, which are Type $\mathbf{E_p}$ and Type **O** patients, representing ED patients for whom inpatient bed request signals have been sent (i.e., patients *positively* predicted to be admitted to the specific IUs) and patients coming from other sources, respectively.

While Figure 6 presents the proposed proactive bed allocation scheme focusing only on ED patients, if not handled effectively, a bed readied for a Type **E** patient might be occupied by a Type **O** patient, compromising operational benefit. To guarantee the positive impact of the proactive strategy to the greatest extent possible to ED patients, we propose proactive bed allocation that *reserves IU beds* for Type **E** patients rather than non-dedicated bed allocations. By reserving beds for Type **E** patients, beds prepared in response to requests will be dedicated to ED patients and can only be occupied by ED patients. Moreover, we assume that Type **E** patients cannot take beds prepared for Type **O** patients either, to ensure that the bed reservation will not compromise the efficiency of bed allocation for Type **O** patients.

2.4.2 Accounting for Disposition Decision Prediction Errors

Under the proactive inpatient bed request and allocation scheme, each of the predictions becomes a source of errors caused from uncertainties in ED treatment and operations. The errors generated from remaining LoS estimation is partially represented with the exponentially distributed service time of s_1 server, which is remaining service time in the ED. Since coefficient of variation is unity for an exponentially distributed random variable, we implicitly model the

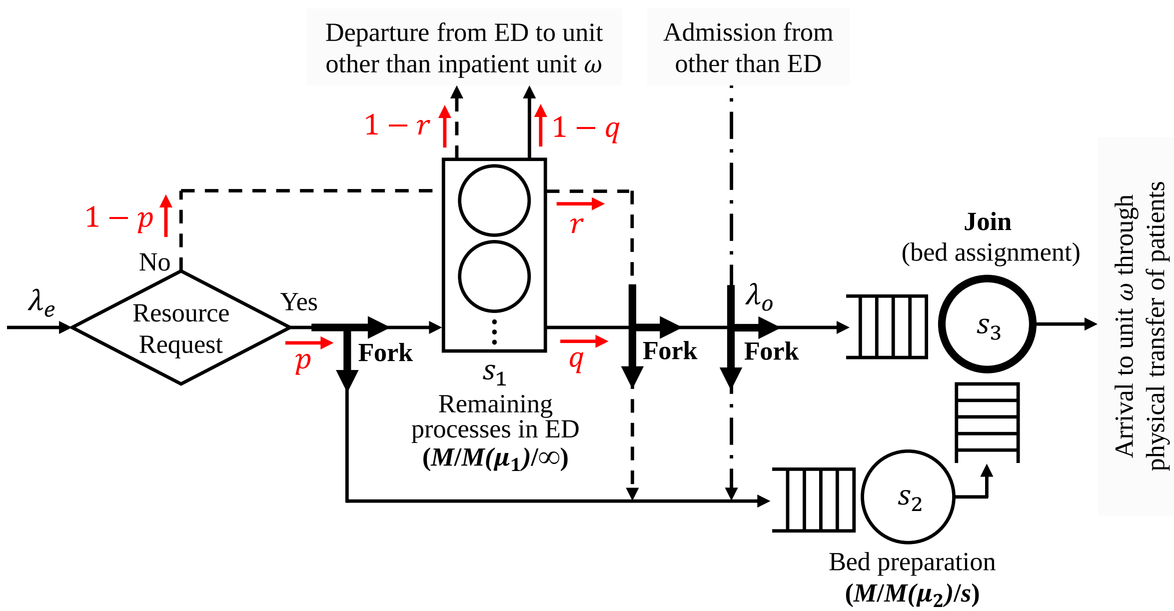


Figure 7: Fork-Join Queue Structure of Proactive Inpatient Bed Allocation.

estimated remaining service time at s_1 with an increasing variance as the mean increases. In other words, the sooner we estimate the remaining ED LoS for a patient upon arrival to ED, the higher the variance of the estimate. This is consistent with the observation that as patients go through more treatment/observation processes in the ED, the uncertainty of patients' future state information reduces because of the larger amount of information accumulated for patients.

The second type of error comes from disposition decision prediction. While the outcomes from remaining ED LoS estimation are continuous values, prediction results from disposition prediction are discrete, being a class. In particular, for a specific IU, say unit ω, the outcome of the classification will be either positive, meaning "will be admitted to ω" or negative, meaning "will not be admitted to ω". The outcome of the classification will be either correct or incorrect, and depending on actual disposition decisions, patients must be handled in different ways for reducing the total ED boarding delay. Figure 7 depicts the queueing network representation of the ED-IU network dealing with the classification errors as well as other admission sources. To fully specify all possible behaviors in the network, we introduce three probability parameters p, q, and r in Figure 7. p is the probability of sending a bed request to unit ω. There are two types of errors possible in the classification problem. First, false positive or Type-I error indicates the case where a classification model predicted that a patient will be sent to unit ω, while the patient is actually discharged or admitted to any other unit than ω, which is represented with probability $1-q$ in Figure 7. Second, false negative or Type-II error indicates the case where a patient is classified to be discharged or admitted to a unit other than ω, but the patient is actually being admitted to unit ω as displayed by probability r in Figure 7. The frequency of sending a bed request to unit ω is a function of probability thresholds in a disposition decision prediction model that acts as a filter for making positive prediction.

Unlike clinical classification problems, the evaluation of the two types of errors in operations should involve the investigation of the operational impact of them, rather than their nominal levels. Even the impacts are hard to be evaluated based on a prediction performance measure, e.g., the accuracy that is widely used to measure prediction performance. Let A denote the accuracy of disposition prediction and Z denote the admission probability from the ED to unit ω. Then, $A = pq + (1-p)(1-r)$ and $Z = pq + (1-p)r$ by definition. Since $0 \leq q \leq \min(1,\ Z/p)$,

$$1 - Z - p \ \leq A \leq \begin{cases} 1 + Z - p & \text{if } p \geq Z \\ 1 - Z + p & \text{otherwise,} \end{cases} \tag{2.2}$$

holds true. Even though the quality of bounds for accuracy are determined by the quality of

Table 3: Patient Types

Patient type	Source	Advance bed request	Disposition decision	Rate	Classification result
$\mathbf{E_P}$	ED	Yes	Not made yet	$p\lambda_e$	
$\mathbf{E_P^T}$	ED	Yes	Admitted	$pq\lambda_e$	True positive
$\mathbf{E_P^F}$	ED	Yes	Not admitted	$p(1-q)\lambda_e$	False positive (Type-I error)
$\mathbf{E_N}$	ED	No	Not made yet	$(1-p)\lambda_e$	
$\mathbf{E_N^T}$	ED	No	Not admitted	$(1-p)(1-r)\lambda_e$	True negative
$\mathbf{E_N^F}$	ED	No	Admitted	$(1-p)r\lambda_e$	False negative (Type-II error)
O	Non-ED	No	Admitted	λ_o	

classification models, the equation implies that the lower bound for accuracy of disposition prediction increases as the rate of early bed requests decreases $(p\lambda_e)$. In addition, the upper bound also increases as the rate of early bed requests decreases except for the case when $p < Z$. Considering that the value of Z is fairly small (e.g., in the studied hospital $Z < 0.1$ for any single IU), a naive policy primarily concerned about prediction accuracy can discourage early bed requests with smaller p, counter to our goal of relying on proactive requests to reducing boarding delays. Instead, the coordination policy should better account for the costs and operational outcomes associated with both Type-I and Type-II prediction errors. Through the patient types defined (introduced in Table 3), our model can quantify the operational impact based on a variety of prediction performance measures such as accuracy, diagnostic odds ratio, and F-score.

According to the proactive bed allocation network depicted in Figure 7, seven types of patients can exist as listed in Table 3. The arrival of ED patients to the network is modeled as a Poisson process with rate λ_e. A Type $\mathbf{E}$ patient turns into either a Type $\mathbf{E_P}$ patient with probability p or a Type $\mathbf{E_N}$ (*negatively* predicted) patient with probability $1-p$. Then, a Type $\mathbf{E_P}$ patient becomes either a Type $\mathbf{E_P^T}$ patient with rate q or a Type $\mathbf{E_P^F}$ (Type-I error) patient with rate $1-q$, depending on the final disposition decision for the patient. Also, depending on the actual outcomes, a Type $\mathbf{E_N}$ patient becomes either a Type $\mathbf{E_N^T}$ or $\mathbf{E_N^F}$ (Type-II error) patient. Then, only Type $\mathbf{E_P^T}$ and $\mathbf{E_N^F}$ patients head for server s_3. Since we assume indistinguishable inpatient beds, a bed reserved upon a bed request made for a Type $\mathbf{E_P^T}$ patient can be assigned to any Type $\mathbf{E_P^T}$ or $\mathbf{E_N^F}$ patients in a FIFO manner.

2.4.3 Representation of Bed Types

The errors in classification prediction not only generate different types of patients but also cause different types of beds. In particular, when a Type-I error (false positive) occurs, the network exhibits state-dependent transition behaviors. For instance, if an ED patient turns out to be false positive for whom a bed is already prepared and if there are no ED patients in the network, the bed should be handed over to non-ED patients waiting for beds. Therefore, this situation brings the need for defining an activity that cancels the bed reservation and releases the bed so that it can be occupied by anyone. Let us call this action 'release'. Thus, we need to clearly define different types of beds to fully describe interactions between multiple types of patients and beds under prediction uncertainty. We define five types of beds within the proposed proactive bed allocation scheme:

1. Type $\mathbf{E_B}$ beds represent those *being prepared* in server system s_2 for Type $\mathbf{E_P}$ or $\mathbf{E_P^T}$ patients. They include beds in the server queue.

2. Type $\mathbf{E_W}$ beds represent those for which bed preparation processes are already completed. These beds are *waiting* for Type $\mathbf{E_P^T}$ or $\mathbf{E_N^F}$ patients to occupy them.

3. Type $\mathbf{O_B}$ beds represent those that are being prepared in server system s_2 for Type $\mathbf{O}$ patients. They include beds in the server queue.

4. Type $\mathbf{R_B}$ beds represent released beds that had once been Type $\mathbf{E_B}$ beds and are still in process at server s_2. A Type-I error can change a Type $\mathbf{E_B}$ bed to be Type $\mathbf{R_B}$.

5. Type $\mathbf{R_W}$ beds represent either released beds that had once been Type $\mathbf{E_W}$ or $\mathbf{R_B}$ beds that have finished their bed preparation/allocation processes.

2.4.4 Transitions Defined within the ED-IU Network

We now introduce six transitions that can fully specify the behaviors of the trivariate Markov process as follows: 1) arrival of a Type $\mathbf{E_P}$ patient at rate $p\lambda_e$; 2) arrival of a Type $\mathbf{E_N^F}$ patient at rate $(1-q)r\lambda_e$; 3) arrival of a Type **O** patient at rate λ_o; 4) completion of remaining ED processes for a Type $\mathbf{E_P^T}$ patient at rate $\dfrac{iq}{\mu_1}$ when i Type $\mathbf{E_P}$ patients are processed in server s_1; 5) completion of remaining ED processes for a Type $\mathbf{E_P^F}$ patient at rate $\dfrac{i(1-q)}{\mu_1}$ when i Type $\mathbf{E_P}$ patients are processed in server s_1; and 6) completion of preparation/allocation of a bed at server s_2 at rate $\dfrac{1}{\mu_2}$. Instead of assuming that all servers and queues respond identically to the transitions (typical of fork-join queue structures), our model entails more realistic transitions between states depending on system states, which are presented in the next section.

2.4.5 Inpatient Bed Reservation Strategy

To guarantee the maximized positive influence of advance resource reservation in reducing boarding delay of ED patients, not compromising the bed allocation efficiency for Type **O** patients and the effort of server s_2, we explicitly delineate a set of bed reservation rules that govern bed reservation/release/cancellation for effective preparation/allocation of beds. The designed rules strategically handle different patient arrivals and bed preparation states. The rules are as follows:

1. An early bed request is made only for Type $\mathbf{E_P}$ patients.

2. ED and non-ED patients are served based on FIFO at the join server, s_3, but according to prepared bed types. An ED patient can only take a bed that is either reserved for ED patients or released, and a non-ED patient can only occupy a bed that is not reserved for ED patients.

3. Once a false positive occurs, then there are more beds that are either already reserved or being reserved than necessary. In this case, a bed is released or removed according to a bed preparation/allocation state at a given time.

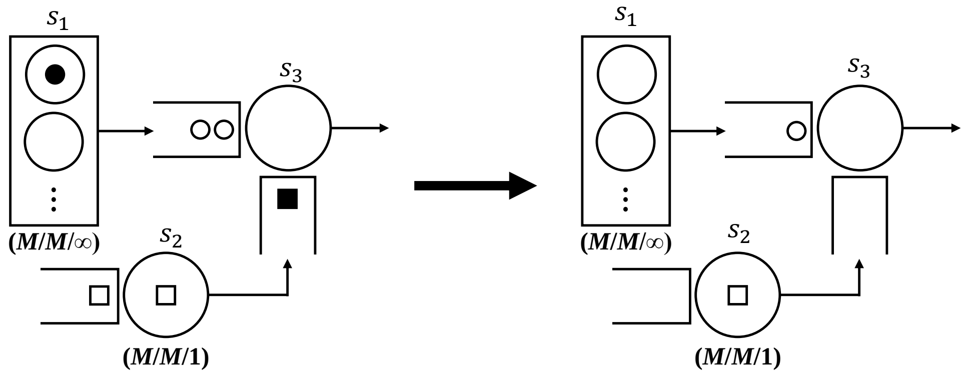
   (a) If all beds in the network are already prepared and ready, any one of them is released.

   (b) If there are beds waiting in the queue of server system s_2 to be prepared for ED patients, the last one is removed from the queue (LIFO).

   (c) If there is a Type $\mathbf{E_B}$ bed being currently prepared without any Type $\mathbf{E_B}$ bed in the queue of server s_2, the bed is released.

4. The type of a bed being prepared or already completed is decided towards the minimization of the number of beds to be prepared by releasing or removing the bed if it is allowed.

Consider Figure 8, where black and white circles represent ED patients and non-ED patients, respectively, and black and white squares represent beds prepared or being prepared for ED patients, and non-ED patients, respectively. Initially there are one Type $\mathbf{E_P}$ patient, two Type **O** patients, two Type $\mathbf{O_B}$ beds, and one Type $\mathbf{E_W}$ beds. As soon as a false positive case happens, the following events occur in order. As the Type $\mathbf{E_P^F}$ patient leaves the network, the



A false positive case happens when there are multiple beds being processed for non-ED patients

Figure 8: ED-IU Network Behaviors According to Bed Reservation Rules

Type $\mathbf{E_W}$ bed is released and becomes Type $\mathbf{R_W}$ (by rule (3–a)). Then, one Type **O** patient and the Type $\mathbf{R_W}$ bed is joined at server s_3 (by rule (2)), and the Type $\mathbf{O_B}$ bed in the queue of server s_2 is removed (by rule (4)). Eventually, a Type **O** patient and a Type $\mathbf{O_B}$ bed remain in the network.

All the rules above are designed based on actual bed allocation practices in industry but are reasonably modified in the direction of the most efficient way to reduce unnecessary waiting times and the wastage of effort to prepare beds. In fact, the rules guarantee the same amount of workload on bed preparation with the reactive processes.

2.4.6 Representation of ED-IU Network State Space

To obtain the steady-state probability distribution of the number of patients and beds of each type in the network, state space should include the following information due to state-dependent transition rules: the number of Type $\mathbf{E_P}$ patient, the sum of the numbers of Type $\mathbf{E_P^T}$ and Type $\mathbf{E_N^F}$ patients, the numbers of Type **O** patients, the number of Type $\mathbf{E_B}$, $\mathbf{E_W}$, $\mathbf{O_B}$ and $\mathbf{R_B}$ beds, and the permutation of the types of patients for which beds are being prepared in server s_2. Since this network incorporates the nine-dimensional state space with state-dependent transition rules, it is challenging to analytically solve the steady-state probability distribution of the network. One alternative is using simulation, which may result in limited insights. Instead, we propose representing the network as a trivariate Markov process by imposing constraints that can aggregate and categorize the whole state space into manageable subsets of state space.

It is worth noting that the properties of Burke's output theorem enable reduction of the complexity of the problem [22]. First, the arrival processes of Type $\mathbf{E_P}$ and $\mathbf{E_N}$ patients are independent by the property of Poisson processes. Therefore, the departure process of Type $\mathbf{E_N}$ patients at server system s_1 can also be analyzed independently. Moreover, the number of Type $\mathbf{E_N}$ patients in server s_1 is not of interest as we pay attention to queue lengths at server s_2 and s_3. Only when a Type $\mathbf{E_N}$ patient turns into $\mathbf{E_N^F}$, entering to server s_3, it could affect the queue lengths at server s_2 and server s_3 based on the proactive bed reservation rules. According to Burke's output theorem, the departure of Type $\mathbf{E_N}$ patients from server system s_1 in the steady state is also a Poisson process having rate $(1-p)\lambda_e$. Among the departures with rate $(1-p)\lambda_e$, $(1-p)r\lambda_e$ of them proceed to server s_3, which now become Type $\mathbf{E_N^F}$ patients. Therefore, unlike Type $\mathbf{E_P}$ patients, without explicitly tracking the behavior of Type $\mathbf{E_N}$ patients in server system s_1, we can achieve a complete model to obtain analytical solutions for the steady state probability distribution of queue lengths at server s_3.

The proposed queuing network is represented as a continuous-time Markov process on the state space $\{(i,\ \boldsymbol{j},\ k) : 0 \le i \le N,\ 0 \le |\boldsymbol{j}| \le N,\ -N \le k \le N\}$, where i is the number of Type $\mathbf{E_P}$ patients, $\boldsymbol{j}$ is a string that stores the composition and order of Type $\mathbf{E_B}$ and Type $\mathbf{O_B}$ beds in server s_2, consisting of elements in a patient type set $PT = \{e, o\}$, and k indicates the number of either ED patients or reserved beds queued at server s_3. For $k > 0$, k represents the number of ED patients waiting for Type $\mathbf{E_W}$ beds, whereas for $k < 0$, $|k|$ is the number of Type $\mathbf{E_W}$ or Type $\mathbf{R_W}$ beds waiting to be occupied by Type $\mathbf{E_P^T}$, Type $\mathbf{E_N^F}$ or Type $\mathbf{O}$ patients. To be specific, Type $\mathbf{E_W}$ beds can be taken only by Type $\mathbf{E_P^T}$ or Type $\mathbf{E_N^F}$ patients, while Type $\mathbf{R_W}$ beds can be assigned to any patients. Let $|\boldsymbol{j}|$ denote the number of elements in a string $\boldsymbol{j}$, where $|\boldsymbol{j_e}|$ and $|\boldsymbol{j_o}|$ represent the numbers of e elements and o elements in $\boldsymbol{j}$, respectively. We do not need an element to represent Type $\mathbf{R_B}$ beds in string $\boldsymbol{j}$ since the number of Type $\mathbf{R_B}$ beds can be inferred from i and k.

Due to the complexity of the system and the string-based representation of state vectors, simple lexicographic ordering still does not work. For this reason, we introduce methods that can categorize the whole state space into subsets of states so that transitions can be aggregated and organized between different subsets of states. We define (1) 'sets' of states, i.e., $H(n)$, based on the number of bed requests remaining in the network, (2) 'blocks' based on the sign of k within a set, (3) 'groups' based on the availability of different types of beds within a block, and (4) 'sequences' based on transition patterns across groups (refer to Sections 2.5 and 2.8.2). The patterns in transitions in turn enable to find an analytical solution to the steady state distribution by solving global balance equations. Without loss of generality, we focus on the state space for the $M/M/\infty - M/M/1$ network case for clarity, and the rules can be extended and applied to $M/M/\infty - M/M/s$ cases for any s.

To study the structure of the state space, we first define variables that represent the number of entities under consideration. For instance, we simply let E_P^T and E_B denote the number of Type $\mathbf{E_P^T}$ patients and Type $\mathbf{E_B}$ beds in the network, respectively. Therefore, $E_P^T = k - E_N^F$ when $k > 0$, and $E_B = |k| - R_B$ when $k < 0$. To more succinctly describe the network conditions, we make sure that $L_Q^{PT} = E_P^T + E_N^F + O$ and $L_Q^{BD} = E_W + R_W$ to indicate the number of patients and beds that are queued at server s_3 at any given time based on the definition of $L_Q^{PT}(t)$ and $L_Q^{BD}(t)$ introduced in Section 2.3. A set of possible states satisfies the following conditional equation and inequality based on the previous discussion on the state space representation:

$$|\boldsymbol{j_e}| - i \begin{cases} = \; k & \text{if } R_B = 0 \wedge R_W = 0 \\ > \; k & \text{otherwise} \end{cases} \tag{2.3}$$

where the condition $R_B = 0 \wedge R_W = 0$ indicate the case that there are no released beds in the network at all. Hence, the number of patients who need beds and the number of beds being prepared or already prepared are balanced, i.e., $E_P + E_P^T + E_N^F + O = E_B + E_W + O_B$, making $i + k = |\boldsymbol{j_e}|$ hold true, where $i + k$ represents the number of ED patients that require beds to be prepared additionally, and $|\boldsymbol{j_e}|$ is the number of beds being currently prepared, corresponding to the $i + k$ bed requests yet to be met. On the other hand, if there are released beds at server s_2 and s_3, there are more beds than necessary (the second case of Equation 2.3). Because of Type $\mathbf{R_B}$ and Type $\mathbf{R_W}$ beds existing among $|\boldsymbol{j_e}|$ number of e elements in $\boldsymbol{j}$, $i + k < |\boldsymbol{j_e}|$ is true. Note that this case can happen only when $k \leq 0$, since when $k > 0$, released beds cannot exist in the network. When the second case of Equation 2.3 is met, $|i + k| + |\boldsymbol{j_e}|$ number of released beds, i.e., $|i + k|$ units of Type $\mathbf{R_B}$ and $|\boldsymbol{j_e}|$ units of Type $\mathbf{R_W}$ beds, exist in the network. We note that in the $M/M/\infty - M/M/s$ network, the maximum possible number of Type $\mathbf{R_B}$ beds is s, which are being prepared. In other words, if a Type $\mathbf{E_P}$ patient turns to be a Type $\mathbf{E_P^F}$ patient and $|\boldsymbol{j_e}| > s$, then the last e element in $\boldsymbol{j}$ is removed as stated by rule (3–b). Server s_2 can “cancel" the beds in queue once a false positive occurs since bed allocation processes for those beds have not begun yet. This rule can be applied to $M/M/\infty - M/M/s$ for any s. The way to indicate the number of Type $\mathbf{O}$ patients and Type $\mathbf{O_B}$ beds is straightforward. It is always true that $|\boldsymbol{j_o}| = O_B = O$. The first equality holds true by definition, and the second equality is valid since the bed request for Type **O** patients is made only when the patients actually arrive to the network given no prediction ability for the demand of Type **O** patients. Hence, there is no need to specify the number of Type **O** patients in the network.

Now we define set $H(n)$ to indicate all states that have n bed requests remaining in the network. This process helps organize the whole state space so that transition patterns can be explored by further categorizing the states. For example both $\big(1,\ \langle e \rangle,\ 0\big)$ and $\big(0,\ \langle\ \rangle,\ -1\big)$ are in set $H(1)$, since in both states there is a single bed request remaining in the network whether it is already prepared or not. Any states that satisfy the following equation can be categorized into set $H(n)$:

$$|\boldsymbol{j}| + |\min(k,\ 0)| = n. \tag{2.4}$$

Equation 2.4 ties up the states that contain n bed requests. With Equations 2.3 and 2.4, we are able to represent the complete set of states and classify the states in each set according to the transitional behaviors. Sections 2.5 and 2.8.2 provide a detailed explanation of the state space.

2.5 State Groups and Transition Matrix Blocks

In this section, we discuss the partitioning of the generator matrix for the processes within the proposed fork-join network into tractable blocks according to the state space defined in Section 2.4.6. Partitioning greatly reduces the complexity of the transition behaviors and transforms the matrix into separable blocks with common patterns that are manageable. Toward this end, let $S_\nu(X)$ denote the set containing all permutations with repetitions that choose ν elements from a set X. For the number of permutations, we have $|S_\nu(X)|{=}|X|^\nu$. We categorize the state space into two main blocks for each set n: $H_{k^-}(n)$ and $H_{k^+}(n), \forall n \geq 1$, where n is the number of remaining bed requests in the ED-IU network, as follows:

1. $H_{k^-}(n) = \{(i, \boldsymbol{j}, k) : Eq.(2.3) - (2.4), k < 0, \boldsymbol{j} \in S_\nu(PT)\} :\Leftrightarrow E_P^T + E_N^F = 0 \wedge E_W > 0 \vee R_W > 0 \vee R_B > 0, \ \ \forall n \geq 1.$

2. $H_{k^+}(n) = \{(i, \boldsymbol{j}, k) : Eq.(2.3) - (2.4), k \geq 0, \boldsymbol{j} \in S_\nu(PT)\} :\Leftrightarrow E_P^T + E_N^F \geq 0 \wedge E_W = 0 \wedge R_W = 0 \wedge R_B = 0, \ \ \forall n \geq 1.$

The blocks $H_{k^-}(n)$ include all states in which there are beds already prepared and waiting for Type $\mathbf{E_P^T}$ and Type $\mathbf{E_N^F}$ patients. On the other hand, $H_{k^+}(n)$ represents the states for which there are no beds waiting for patients, but there may be patients waiting for beds. The main reason for partitioning the state space based on k is that the state space representations bifurcate depending on the existence of Type $\mathbf{R_B}$ and Type $\mathbf{R_W}$ beds. As described in Figure 8, false positive predictions can generate Type $\mathbf{R_B}$ and Type $\mathbf{R_W}$ beds in the system, for which bed-patient assignment procedures are different than those for other beds, i.e., other beds are 'reserved' and taken only by corresponding type patients, but released beds can be occupied in a FIFO manner regardless of their original types. Moreover, states in each block share the same transition patterns over n. Therefore, we arrange the states according to the number of remaining bed requests in the system, and partition the states into $H_{k^-}(n)$ and $H_{k^+}(n)$ in each set. The partitioned generator matrix **Q** is given in (2.5).

Matrix **Q** in (2.5) involves 12 types of blocks, labeled from **A** to **L**. A state in block $H_{k^-}(n)$ can transition into a state in blocks from $H_{k^+}(n-2)$ to $H_{k^-}(n+1)$, whereas a state in block $H_{k^+}(n)$ can transition into a state in blocks $H_{k^-}(n-1)$ through $H_{k^+}(n+1)$. For instance, states in block $H_{k^-}(n+1)$ are transitioned to states in block $H_{k^+}(n-1)$, following transition patterns in $\mathbf{A_{(n+1)(n-1)}}$, and there are no direct transitions between $H_{k^-}(n+1)$ and $H_{k^-}(n-1)$. Each type of blocks share unique state patterns, enabling a tractable representation for the exponentially increasing state space. Hence, we compute transition rate matrix and solve global balance equations to obtain steady-state distribution for all states, regardless of the size of $\mathbf{Q}$.

$$\mathbf{Q} = \begin{array}{c} \\ H(0) \\ \vdots \\ H_{k^-}(n-1) \\ H_{k^+}(n-1) \\ H_{k^-}(n) \\ H_{k^+}(n) \\ H_{k^-}(n+1) \\ H_{k^+}(n+1) \\ \vdots \\ H_{k^+}(N) \end{array} \begin{array}{c} \begin{array}{cccccccccc} H(0) & \ldots & H_{k^-}(n-1) & H_{k^+}(n-1) & H_{k^-}(n) & H_{k^+}(n) & H_{k^-}(n+1) & H_{k^+}(n+1) & \ldots & H_{k^+}(N) \end{array} \\ \left[\begin{array}{cccccccccc} & & & & & & & & & \\ & \ddots & \ddots & \ddots & & & & & & \\ & \ldots & \mathbf{D_{(n-1)(n-1)}} & \mathbf{E_{(n-1)(n-1)}} & \mathbf{F_{(n-1)(n)}} & & & & & \\ & \ldots & \mathbf{I_{(n-1)(n-1)}} & \mathbf{J_{(n-1)(n-1)}} & \mathbf{K_{(n-1)(n)}} & \mathbf{L_{(n-1)(n)}} & & & & \\ & \ldots & \mathbf{B_{(n)(n-1)}} & \mathbf{C_{(n)(n-1)}} & \mathbf{D_{(n)(n)}} & \mathbf{E_{(n)(n)}} & \mathbf{F_{(n)(n+1)}} & & & \\ & \ldots & \mathbf{G_{(n)(n-1)}} & \mathbf{H_{(n)(n-1)}} & \mathbf{I_{(n)(n)}} & \mathbf{J_{(n)(n)}} & \mathbf{K_{(n)(n+1)}} & \mathbf{L_{(n)(n+1)}} & & \\ & & & \mathbf{A_{(n+1)(n-1)}} & \mathbf{B_{(n+1)(n)}} & \mathbf{C_{(n+1)(n)}} & \mathbf{D_{(n+1)(n+1)}} & \mathbf{E_{(n+1)(n+1)}} & \ldots & \\ & & & & \mathbf{G_{(n+1)(n)}} & \mathbf{H_{(n+1)(n)}} & \mathbf{I_{(n+1)(n+1)}} & \mathbf{J_{(n+1)(n+1)}} & \ldots & \\ & & & & & & & & \ddots & \ddots \\ & & & & & & & & \ddots & \mathbf{J_{(N)(N)}} \end{array} \right] \end{array}. \tag{2.5}$$

We further partition each block to fully specify the state space and transition patterns. Depending on the availability of prepared beds and their types, $H_{k^-}(n), \forall n$ can be categorized into n+2 groups.

1. $H^1_{k^-}(n) = \{(i, \boldsymbol{j}, k) : 0 \leq i < n, \boldsymbol{j} = , k = -n\}, \forall n \geq 1$, for which we have $E_P = i \wedge L^{PT}_Q = E_B = O_B = R_B = 0 \wedge E_W = i \wedge R_W = n - i, \forall n \geq 1$. The states in group $H^1_{k^-}(n)$ have more than necessary beds for ED patients who have sent bed requests, and redundant reserved beds are released. There is not any non-ED patients, and no more beds need to be prepared.

2. $H^2_{k^-}(n) = \{(i, \boldsymbol{j}, k) : 0 \leq i < n, \boldsymbol{j} = \langle e \rangle,\ k = -n+1\}, \forall n \geq 1$, where we have $E_P = i \wedge L_Q^{PT} = E_B = O_B = 0 \wedge R_B = 1 \wedge E_W = n-1 \wedge R_W = n-i-1$. The states in group $H^2_{k^-}(n)$ have also more than necessary beds for ED patients who have sent bed requests, and redundant reserved beds are released. There is not any non-ED patients. The bed for which the preparation process already started, is released.

3. $H^{3+\nu}_{k^-}(n) = \{(i, \boldsymbol{j}, k) : i = |\boldsymbol{j_e}| + |k|, \boldsymbol{j} \in S_\nu(PT), k = -n+\nu\}$ for $n \geq 2$, $\boldsymbol{j} \in S_n(PT)$ and $0 \leq \nu \leq n-1$, for which we have $E_P = i \wedge L_Q^{PT} = O = n - E_P \wedge E_B = |\boldsymbol{j_e}| \wedge O_B = k \wedge R_B = 0 \wedge E_W = n - \nu \wedge R_W = 0$. The states in group $H^{3+\nu}_{k^-}(n)$ have ν beds being prepared for ED and non-ED patients in the order of bed request arrivals. The number of patients who have sent bed requests equals to the sum of the number of beds being prepared and already reserved for each type of patients. Hence, there is no redundant beds.

Groups $H^1_{k^-}(n), H^2_{k^-}(n)$, and $H^{3+\nu}_{k^-}(n), \forall \nu \leq n-1$ satisfy the requirements of block $H_{k^-}(n)$ and cover all states in block $H_{k^-}(n)$ for each n. For instance, the states in each group of $H_{k^-}(4)$ are as follows:

- $H^1_{k^-}(4) = \{(0, , -4), (1, , -4), (2, , -4), (3, , -4)\}$.
- $H^2_{k^-}(4) = \{(0, \langle e \rangle, -3), (1, \langle e \rangle, -3), (2, \langle e \rangle, -3), (3, \langle e \rangle, -3)\}$.
- $H^3_{k^-}(4) = \{(4, , -4)\}$.
- $H^4_{k^-}(4) = \{(4, \langle e \rangle, -3), (3, \langle o \rangle, -3)\}$.
- $H^5_{k^-}(4) = \{(4, \langle ee \rangle, -2), (3, \langle oe \rangle, -2), \big(3, \langle eo \rangle, -2\big), \big(2, \langle oo \rangle, -2\big)\}$.

- $H^{6}_{k^-}(4) = \{(4, \langle eee \rangle, -1),\ (3, \langle oee \rangle, -1),\ (3, \langle eoe \rangle, -1),\ (2, \langle ooe \rangle, -1), (3, \langle eeo \rangle, -1),$
$(2, \langle oeo \rangle, -1), (2, \langle eoo \rangle, -1),\ (1, \langle ooo \rangle, -1)\}.$

Considering the permutations in $\boldsymbol{j}$ for groups $H^{3+\nu}_{k^-}(n)$, the number of states in block $H_{k^-}(n)$ is calculated using the following formula:

$$\Upsilon(n) = \sum_{j=1}^{n} \sum_{i=1}^{j} \frac{(j-1)!}{(j-i)!(i-1)!} + 2n = \sum_{j=1}^{n} 2^{j-1} + 2n, \quad \forall n \in \mathbb{Z}_+, \qquad (2.6)$$
$$\Upsilon(0) = 1.$$

The blocks $H_{k^+}(n), \forall n$ are partitioned into $|PT|^n$ groups as defined as follows:

$$H^{\tau}_{k^+}(n) = \{(i,\ \boldsymbol{j},\ k) : i = |\boldsymbol{j_e}| - k,\ \boldsymbol{j} \in S_n(PT),\ 0 \leq k \leq |\boldsymbol{j_e}|\}, \forall n \geq 1 \text{ and } \forall \boldsymbol{j} \in S_n(PT),$$

where we have $E_P = i \wedge E^T_P + E^F_N = k \wedge E_B = |\boldsymbol{j_e}| \wedge O = O_B = |\boldsymbol{j_o}| = n - |\boldsymbol{j_e}| \wedge R_B = 0 \wedge L^{BD}_Q = 0$, and where τ is an index corresponding to the elements of $S_n(PT)$. For example, we have $S_2(PT) = \{\langle ee \rangle, \langle eo \rangle, \langle oe \rangle, \langle oo \rangle\}$ and $\tau \in \{1, 2, 3, 4\}$, for $n = 2$. The groups $H^{\tau}_{k^+}(n), \forall \tau$ and $\forall n$ involve states that have τ^{th} element of set $S_n(PT)$ as the ordered composition of beds being prepared in server s_2. By definition of $H_{k^+}(n)$, R_B and R_W are zero. The states in each group in block $H_{k^+}(2)$ are as follows:

- $H^{1}_{k^+}(2) = \{(2, \langle ee \rangle, 0),\ (1, \langle ee \rangle, 1),\ (0, \langle ee \rangle, 2)\}.$
- $H^{2}_{k^+}(2) = \{(0, \langle eo \rangle, 1),\ (1, \langle eo \rangle, 0)\}.$
- $H^{3}_{k^+}(2) = \{(0, \langle oe \rangle, 1),\ (1, \langle oe \rangle, 0)\}.$
- $H^{4}_{k^+}(2) = \{(0, \langle oo \rangle, 0)\}.$

Based on the pattern of state expansion, the number of states in block $H_{k+}(n)$ is calculated with the following formulation:

$$\Phi(n) = \sum_{j=1}^{n+1} \sum_{i=1}^{j} \frac{n!}{(n-i+1)!(i-1)!,} \quad \forall n \in \mathbb{Z}_{\geq 0}. \tag{2.7}$$

The summation of $\Upsilon(n)$ and $\Phi(n)$ gives the total number of states in blocks $H_{k^-}(n)$ and $H_{k^+}(n)$. The total number of states in the network that allows N bed requests is formulated in (2.9).

$$X(n) = \Upsilon(n) + \Phi(n), \quad \forall n \in \mathbb{Z}_{\geq 0}. \tag{2.8}$$

$$\Sigma(N) = \sum_{n=0}^{N} X(n). \tag{2.9}$$

As revealed in (2.6)–(2.9), the state space for n grows exponentially (approximately $\sim 2.33^n$), and the total number of states for N reaches to $\sum_{n=1}^{N} 2.33^n$. More detailed description of the transition patterns between different state sequences is given in 2.8.2.

Once the transition rate matrix is obtained, we identify a proper truncation set of generator $\mathbf{Q}$, which is called the truncation parameter [46], to solve the global balance equations and ensure a reasonable accuracy. The decision on the truncation set is made by finding the minimum n that stabilizes changes in $\mathbb{E}(E_W)$, $\mathbb{E}(O)$, and $\mathbb{E}(L_Q^{BD})$.

2.5.1 Multiple Inpatient Unit Case

In general, admissions from the ED to IUs are interdependent, where the sum of admission rates to individual IUs equals to the total rate of admissions to the hospital. Moreover, a Type $\mathbf{E_P^F}$ patient at an IU can be admitted to another IU as a Type $\mathbf{E_N^F}$ patient. Thanks to Burke's output theorem [22], the case of multiple IUs receiving proactive bed requests from the ED can also be modeled and solved analytically through decomposition. Let Ω denote a set containing all dispositions. The following equations express the relationships between different units, and they are used to model the bed allocation operations over multiple IUs:

$$\begin{aligned}
&\text{(a)} \quad \sum_{\omega\in\Omega} p_\omega = \sum_{\omega\in\Omega} Z_\omega = 1, \\
&\text{(b)} \quad p_\omega q_\omega + r_\omega(1-p_\omega) = Z_\omega \quad \forall\omega\in\Omega, \\
&\text{(c)} \quad \sum_{\omega\in\Omega} p_\omega(1-q_\omega) = \sum_{\omega\in\Omega} r_\omega(1-p_\omega),
\end{aligned} \tag{2.10}$$

where p_ω represents the probability of sending a bed request or discharge signal to an arbitrary unit ω, while Z_ω symbolizes the percentage of Type **E** patients who are actually sent to ω after ED treatment. Similarly, q_ω and r_ω at unit ω correspond to q and r in the single IU case, respectively. Equation 2.10(a) states that a disposition decision prediction is made for every ED patient and that each ED patient is assigned to its actual single disposition destination at the end of ED treatment. Equation 2.10(b) is a general form of the equation of Z in Section 2.4.2. Especially, when the rate of proactive signals sent to ω is the same with the rate of actual disposition to ω, Equation 2.10(b) becomes equivalent to $p_\omega(1-q_\omega) = r_\omega(1-p_\omega)$ for all dispositions in Ω, indicating the local equality between patient inflow generated by false negatives and patient outflow caused by false positives at each disposition in Ω. Equations 2.10(a)–(b) guarantee the validity of the global equality (Equation 2.10(c)) for patient flow over all dispositions even without the local equality since Equation 2.10(a) always holds true. In other words, Equation 2.10(c) requires that the total amount of Type I errors and that of Type II errors should be always the same over set Ω. In addition, even though there are multiple IUs through which patients move according to their prediction results, the patient flow in any

ED-IU network is still feed-forward since a patient would not visit the same server or queue more than one time during their admission processes. Therefore, Equation 2.11 holds true for the independence of queue lengths:

$$P(\omega_1 = \xi_1, \omega_2 = \xi_2, \ldots, \omega_\delta = \xi_\delta) = \prod_{i=1}^{\delta} P_i(\omega_i) \tag{2.11}$$

where we assume δ dispositions in total, and ξ_i represents a queue length at ω_i.

2.6 Analysis of Model

In this section, we discuss the effectiveness of the proposed proactive bed allocation approach. Our primary performance measure is waiting time of the both types of patients. We analyze the system under two main scenarios: Ideal and Realistic. Under the ideal scenario, we assume that IU beds are dedicated to ED patients and ED disposition decisions are perfect. Hence, ED patients are the sole beneficiary, and the results report the maximum potential improvement for ED patients. The purpose of studying the ideal case is to explicitly and concisely show the fundamental impact of the proposed bed allocation strategy. Whereas, under the realistic scenario, we assume that disposition predictions for ED patients are error prone, and the IU serves both ED and non-ED patients. We also quantify the impact of prediction quality under various operational settings.

2.6.1 Ideal: IU Dedicated to ED Patients with Perfect Disposition Decision Prediction

For the ideal scenario, we assume that the arrival rate to the ED is 10 patients/hour and the admission rate to IU (ω) is 0.4 patients/hour, mimicking the activity at the studied hospital. For various bed request signal lead-times and bed preparation service times, the expected bed allocation delay is displayed in Figure 9. We observe that when the proactive bed allocation

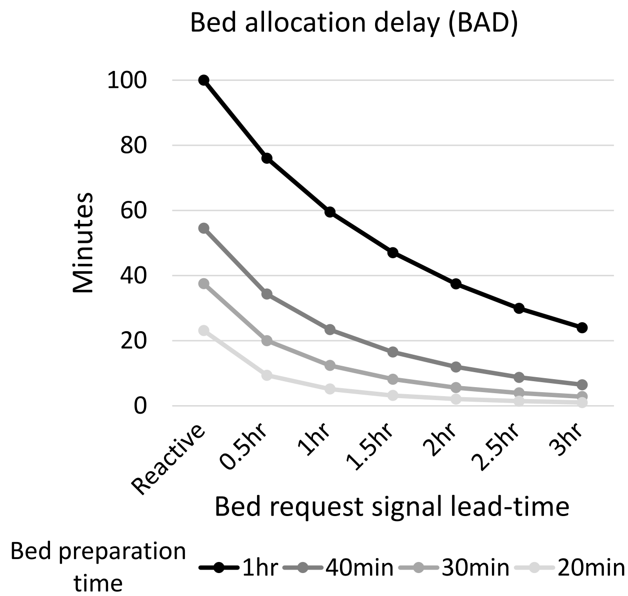


Figure 9: Impact of Proactive Bed Requests for ED patients under a Dedicated IU.

scheme is implemented with a bed request signal lead-time of 1.5 hours for the case where average bed preparation service time is 1 hour, the boarding delay can be reduced from 100 minutes under the reactive strategy to 47 minutes (a 53% reduction). As another example, an IU with 40 minutes of bed preparation service time subject to a bed request signal lead-time of 1 hour experiences the same boarding delay (22 minutes) as an IU operating with just 20 minutes of bed preparation service time. These are significant reductions in BAD enabled through proactive bed request signals. As shown in Figure 9, applying proactive coordination to a busier ED-IU network leads to larger benefits in boarding delay reduction.

2.6.2 Realistic: ED and Non-ED Patient Admissions with Imperfect Disposition Decision Prediction

We conduct our analysis under the setting where the average bed preparation time is 1 hour, and there are the equal number of admissions to the IU from both the ED and outside

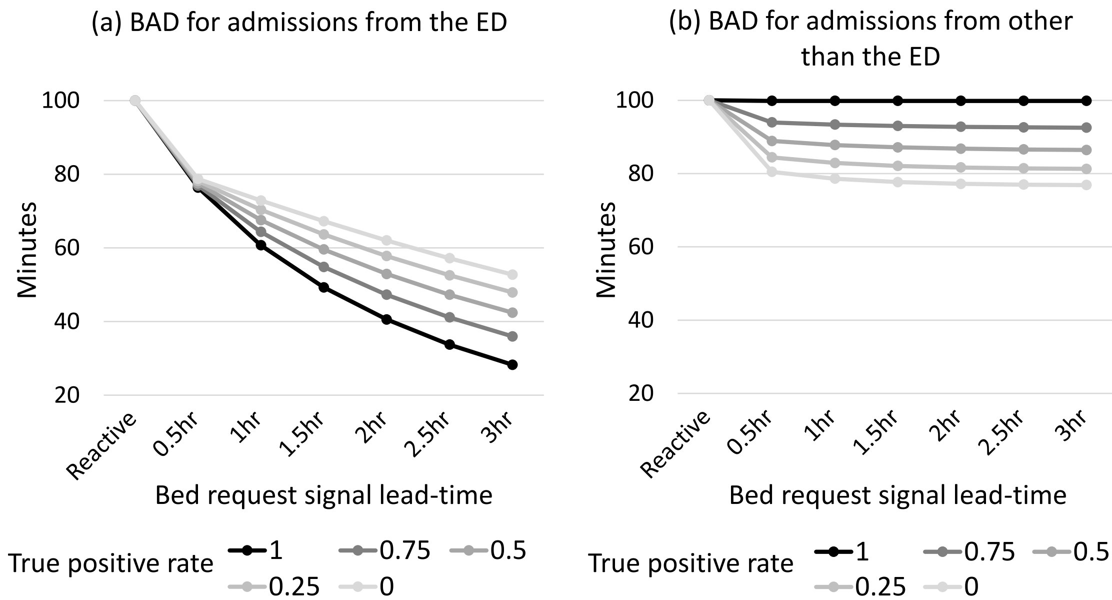


Figure 10: Impact of Proactive Bed Requests in Realistic Settings

ED (0.2 admissions/hour from each, totaling 0.4 admissions/hour) on average to study the effect of erroneous disposition decision predictions on the performance measures. We assume that the frequency of proactive bed requests is the same with the actual admission rate from the ED, i.e., $Z = p$, even though individual predictions could be wrong. Figure 10(a) and (b) display BADs for ED and non-ED admissions respectively, under various true positive rates (q in Figure 7). As the true positive rate gets smaller, the impact of early bed allocation for ED patients decreases, where the value of prediction quality (or the cost of errors) can be approximated by the reduction in the BADs in Figure 10(a). In the experimental setting (the same rate of admissions from the ED and non-ED), the released beds are taken half of the time by Type **O** patients with no delay, resulting in overall BAD reduction for Type **O** patients (Figure 10(b)). Hence, BAD for Type **O** patients under the proposed proactive bed allocation approach is bounded by the delay in the reactive case. Moreover, the delay for Type **O** patients is reduced as a true positive rate decreases and bed request signal lead-time increases, but the amount of reduction quickly saturates as bed request signal lead-time increases. It is because the chance for Type **O** patients to take Type $\mathbf{R_b}$ or $\mathbf{R_w}$ beds rarely depends on how early those beds are available (due to making no reservations for Type **O** patients).

2.6.3 Impact of Quality of Disposition Decision Prediction on BAD Reduction

Figure 11(a) demonstrates the effect of prediction quality on the bed allocation delay reduction for various utilization rates ρ_2 $\left(= \dfrac{Z\lambda_e + \lambda_o}{\frac{1}{\mu_2}} \right)$. The plots display the differences in bed allocation delays for perfect prediction $(q = 1)$, and completely erroneous prediction $(q = 0)$, i.e., $BAD_{q=1} - BAD_{q=0}$, assuming $Z = p$. We observe that with a lower ρ_2, the impact of prediction quality saturates faster as bed request signal lead-time gets longer, i.e., not much benefit from sending bed requests very early. Better individual prediction quality has a greater influence even with a longer bed request signal lead-time when the ED-IU network operates under higher utilization.

Figure 11(b) shows the impact of proactive bed preparation on ED and non-ED patients

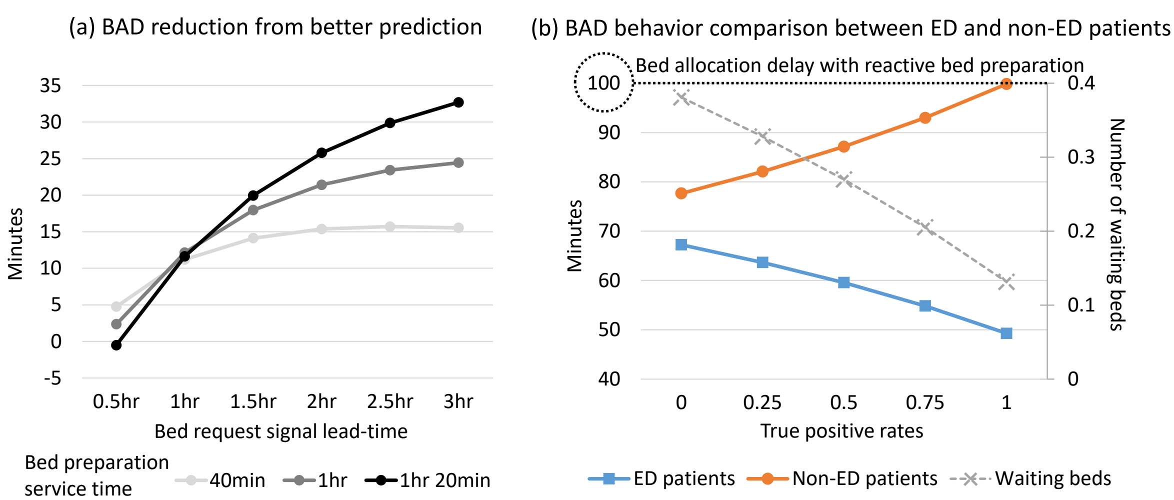


Figure 11: Influence of Disposition Decision Prediction Quality on BAD Reduction

as well as the number of waiting beds for bed preparation service time of 1 hour and bed signal lead-time of 1.5 hours, depending on different levels of disposition prediction quality. We assume an equal number of admissions (0.2 admissions/hour) from the ED and outside ED. For the imperfect disposition prediction cases, ED and non-ED patients compete for available beds. While non-ED patients benefit from false bed allocations, the number of beds waiting for patients increases as the prediction quality decreases.

2.6.4 Timeliness of Proactive Bed Requests with Progressive Prediction Results

Finally, we present the most realistic cases (inspired by analyzing data from the case study hospital) to see how the BAD for ED patients can vary depending on patient disposition prediction quality that evolves progressively throughout the ED care giving cycle. In particular, we identify four essential discrete care giving stages in the ED, i.e., triage, ordering a first set of lab tests, receiving the results of the first set of lab tests, and the final disposition decision. While there can be more steps between any two contiguous stages, these four intervals are fairly well-established and common for most ED patients. The care giving stages, the average time spent at each stage, as well as the results (admission prediction, precision and sensitivity) of the predictive models are displayed in Table 4, where the Logistic regression predictive models are built using the EHR data from the studied hospital. Therefore, unlike the previous analyses, the rates of proactive bed requests are determined from the actual prediction models, which turn out to be strictly less than the admission rates under all cases, denoted p/Z. All other parameters remain the same with those in Section 2.6.2. As a patient goes through more ED care giving stages, there is growing clinical information to improve the reliability of disposition decision prediction, which leads to increasing performance as reported in Table 4.

However, a delayed/postponed bed request signal can also compromise the timeliness of

Table 4: Patient Disposition Prediction Quality Evolution and Its Impact on BAD Reduction

(a) IU ω_1	Triage	First lab orders	Results of first lab orders	Disposition decision
Bed request signal lead-time	250min	200min	145min	0min
p/Z	0.50	0.78	0.93	1.00
q (precision)	0.24	0.31	0.42	1.00
Sensitivity	0.12	0.24	0.39	1.00
BAD for ED patients	63min	53min	**52min**	100min

(b) IU ω_2	Triage	First lab orders	Results of first lab orders	Disposition decision
Bed request signal lead-time	250min	200min	145min	0min
p/Z	0.81	0.82	0.84	1.00
q (precision)	0.49	0.52	0.56	1.00
Sensitivity	0.40	0.43	0.47	1.00
BAD for ED patients	**42min**	47min	53min	100min

bed requests. Table 4 reveals how this trade-off can actually affect BAD reduction, comparing two distinct IUs that have different trajectories of prediction quality progress. IU ω_1 represents a telemetry unit (TU) that does not seem to have distinctive clinically actionable information at triage (in vital signs, chief complaint, and so on) to make effective disposition predictions, but waiting for additional clinical information from downstream stages leads to improved predictive capability for TU patients. As shown in Table 4(a), due to the clear improvement in prediction quality during the ED care giving cycle, making bed request decisions later can actually reduce overall BAD in IU ω_1, overcoming the negative influence of postponed decisions. On the contrary, as shown in Table 4(b), IU ω_2 represents an intensive care unit (ICU) for which the patients have far more distinctive features right at triage to generate fairly good prediction results within about 20 minutes of patient arrival to the ED. In addition, the information gained during downstream care giving stages is not significant, and there is no benefit in delaying proactive bed request signals for these patients by waiting to see what lab work is being ordered or waiting for their results.

2.7 Discussion and Conclusion

To remedy the growing overcrowding being witnessed in EDs and the negative consequences associated with it, we propose a novel early task initiation scheme that facilitates proactive IU bed allocations for ED patients projected to be admitted into the hospital. The proposed scheme does not require any major modifications in care services both in the ED and the IUs or their support services. Our results suggest that making proactive bed reservations for ED patients can significantly reduce bed allocation delay for ED patients (and in turn ED crowding) without compromising the bed allocation efficiency for non-ED patients. The results from our study led the studied hospital to implement parts of the proposed early task initiation strategy within its ED-IU network. In the initial phase of the implementation, the hospital started to send bed requests before the completion of admission approvals for a representative IU. The hospital is also planning to fully implement the proposed strategy for IU bed allocation.

The proposed proactive bed allocation approach can be studied further for operationalization. First, as can be seen from Figure 2, the ED-IU network is a dynamic system with varying patient arrival and discharge rates, requiring modifications to our model setting parameters. While a simulation study can be pursued based on the detailed coordination strategy proposed in this study, for all practical purposes, it is safe to assume that the entire ED-IU network system can be effectively characterized into a relatively small number of distinct steady system states (each spanning one to several hours). Once characterized, the ED should operate the advance bed request signals based on the optimal policy that corresponds to the parameters representing the matching system state.

Second, the proposed coordination scheme can enable different types of process modifications and improvements in the ED-IU network. For instance, when bed preparation servers are shared as a completely pooled server system over all IUs, a hospital can strategically deploy its servers according to the predicted IU bed demand. For bed preparation servers that cover IUs requiring different bed or equipment types, deployment of the servers becomes a prioritization problem. In this setting, proactive bed allocation schemes can be utilized for proactively directing the servers to the most appropriate IUs. While different settings exist, this study provides core ideas, rigorous representation, and operational impact analysis for proactive bed allocation schemes.

Finally, even though this study focuses on a health care service network setting, it adds to the general body of literature investigating service systems that can benefit from proactive resource coordination utilizing prediction outcomes (e.g., just-in-time logistics, manufacturing, and project management). The implementation of early task initiation in complex service systems should become a promising area of scientific research and exploration for both industry and academia.

2.8 Appendix

This section provides additional discussion and details surrounding the proposed modeling methodology and case study. First, we demonstrate that the bed preparation server time does indeed follow an exponential distribution at the studied hospital. Second, we discuss means for identification and organization of state groups and transition matrix blocks for obtaining analytical solutions in large scale ED-IU network settings. Finally, we introduce a matrix that represents the state sequences within each state group.

2.8.1 Exponentially Distributed Bed Preparation Service Time

As noted in Section 2.2, BAD (i.e., bed allocation delay) denotes the total waiting time in the bed preparation server (s_2). Although the main task processed at server s_2 is bed preparation, BAD also includes the time to communicate with the bed preparation servers to deploy them, travel time for servers to move to the assigned bed location, and time to finish duties associated with any prior assignments (i.e., bed cleaning requests might queue up). We assume that the bed preparation server (server s_2) takes exponentially distributed service time with mean μ_2. For justification, Figure 12 reports the cumulative distributions of inter-departure times from server s_2 for two representative inpatient units (internal medicine and pulmonary units) at the target hospital on weekdays for two 2-hour intervals during severe ED census and boarding. The inter-departure plots indeed resemble exponential distributions albeit with varying rates across units. Note that the departure process in a stationary $M/G/s$ queuing system approaches a Poisson process as $s \to \infty$. In particular, as shown by [38] and [39], an $M/G/s$ (for any s ≥ 1) queue has a Poisson departure process if and only if it is a steady-state queue satisfying that the service times are exponential $(G = M)$ when the service discipline is FIFO. Hence, it is reasonable to assume that the bed preparation service times are exponentially distributed.

2.8.2 Representation of State Sequences

In Section 2.5, n+2 and 2^n groups of states are introduced for blocks $H_{k^-}(n)$ and $H_{k^+}(n), \forall n$, respectively. Then, states within each group are sorted in a lexicographic order for i, the first element on the state space, if $|\boldsymbol{j_e}| = |\boldsymbol{j_o}| = 0$, i.e., there is no need to order according to a string

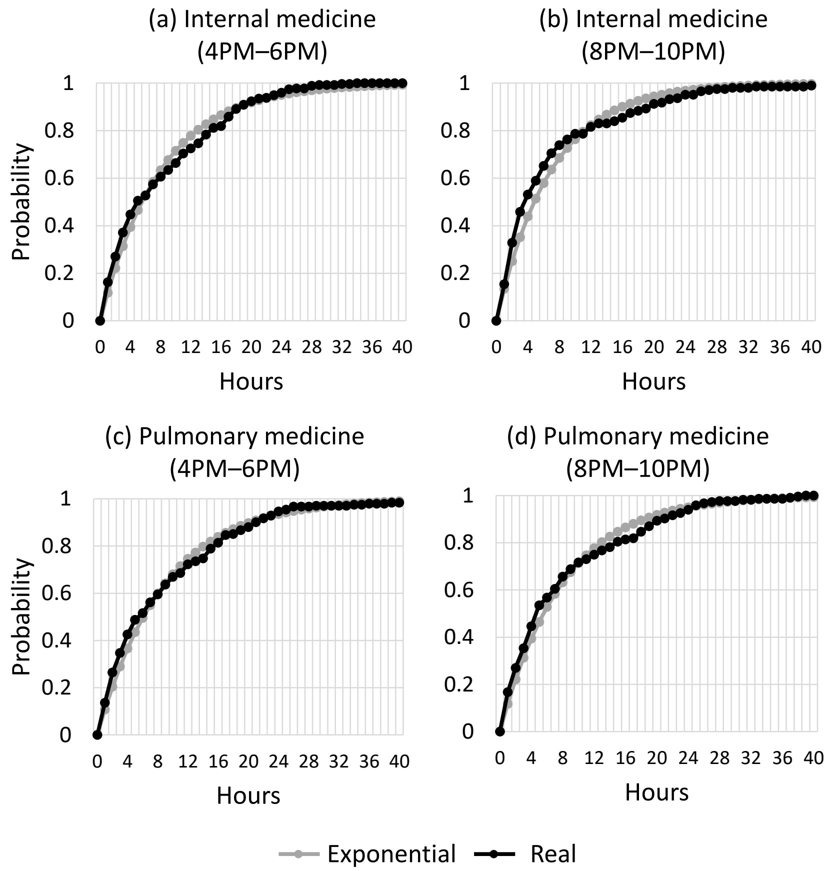


Figure 12: Cumulative Distributions of Inter-departure Time at Server s_2 and Exponential Distribution for Select Inpatient Units

$\boldsymbol{j}$. However, when $|\boldsymbol{j_e}| \neq 0 \vee |\boldsymbol{j_o}| \neq 0$, states are first ordered in a colexicographic order for $\boldsymbol{j}$, assuming $e < o$. For instance, elements in $S_2(PT)$ are ordered using the colexicographic ordering as follows: $ee < oe < eo < oo$. Moreover, states that have the same $\boldsymbol{j}$, are further ordered anti-lexicographically for i. The ordering methods are implemented such that all possible state transitions are conveniently tracked through the regular sequence of states. In particular, the colexicographic order applied for $\boldsymbol{j}$ decides the transition patterns that are related to the number of beds being prepared for Type $\mathbf{E_P}$, $\mathbf{E_P^T}$, and $\mathbf{E_N^F}$ patients. Since states in each group are first ordered based on the elements in $\boldsymbol{j}$ due to their great influence in the network behavior, transition patterns also reflect the order. Considering $S_2(PT) = \{\langle ee\rangle, \langle oe\rangle, \langle eo\rangle, \langle oo\rangle\}$, the number of e elements in each string are 2, 1, 1, 0, respectively. The pattern of the numbers of e elements in strings in any set $S_n(PT)$ can be derived from the colexicographic order. Thus, the 'sequence of the number of e elements in element strings' is given by (0), (1, 0), (2, 1, 1, 0), (3, 2, 2, 1, 2, 1, 1, 0), $\ldots$, and can be represented by the following matrix Ψ:

$$\mathbf{\Psi} = \begin{array}{c} (\alpha,\, n) \\ 0 \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \\ 7 \\ 8 \\ \vdots \end{array} \begin{array}{c} \begin{array}{ccccc} 0 & 1 & 2 & 3 & \ldots \end{array} \\ \begin{bmatrix} 0 & 0 & 0 & 0 & \ldots \\ 0 & 1 & 2 & 3 & \ldots \\ 0 & 0 & 1 & 2 & \ldots \\ 0 & 0 & 1 & 2 & \ldots \\ 0 & 0 & 0 & 1 & \ldots \\ 0 & 0 & 0 & 2 & \ldots \\ 0 & 0 & 0 & 1 & \ldots \\ 0 & 0 & 0 & 1 & \ldots \\ 0 & 0 & 0 & 0 & \ldots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix} \end{array}. \tag{2.12}$$

The first row is added as a dummy for simpler expression of state patterns.

As introduced in Section 2.4.4, the state transitions are categorized into six groups. Below, we list all the state transitions grouped by the transition type. Each sequence corresponds to only one of the groups and subgroups exclusively over increasing n, where the functions $\Upsilon(n)$ and $\Sigma(n)$ are introduced in (2.6) and (2.9), respectively.

Table 5: Transitions for the arrival of a Type $\mathbf{E_P}$ patient at rate $p\lambda_e$.

	Index of the initial state	Index of the final state	
1	$\Sigma(n-1)+n-1$	$\Sigma(n-1)+n$	$\forall n \geq 1$
2	$\Sigma(n-2)+n-1$	$\Sigma(n-1)+2n+1$	$\forall n \geq 2$
3	$\Sigma(n-1)+2n$	$\Sigma(n-1)+2n+1$	$\forall n \geq 2$
4	$\Sigma(n-1)+m$	$\Sigma(n-1)+m+1$	$\forall n \geq 2,\ 0 \leq m \leq n-2$
5	$\Sigma(n-1)+2n-1$	$\Sigma(n-1)+2n$	$\forall n \geq 2$
6	$\Sigma(n-1)+n+m+1$	$\Sigma(n-1)+n+m+2$	$\forall n \geq 3,\ 0 \leq m \leq n-3$
7	$\Sigma(n-2)+2(n-1)+\sum_{\alpha=1}^{m+1} 2^{\alpha-1}+l$	$\Sigma(n-1)+2n+\sum_{\alpha=1}^{m+2} 2^{\alpha-1}+l$	$\forall n \geq 3,\ 0 \leq m \leq n-3,\ 0 \leq l \leq 2^{m+1}-1$
8	$\Sigma(n-2)+\Upsilon(n-1)+\sum_{\alpha=0}^{m}\Psi(\alpha,n)+l$	$\Sigma(n-1)+\Upsilon(n)+m+\sum_{\alpha=0}^{m}\Psi(\alpha,n)+l$	$\forall n \geq 1, 0 \leq m \leq 2^{n-1}-1, 0 \leq l \leq \Psi(m+1,n)-1$

Table 6: Transitions for the arrival of a Type $\mathbf{E_N^F}$ patient at rate $(1-q)r\lambda_e$.

	Index of the initial state	Index of the final state	
1	$\Sigma(n)$	$\Sigma(n-1)$	$\forall n \geq 1$
2	$\Sigma(n)+n$	$\Sigma(n-1)+n$	$\forall n \geq 1$
3	$\Sigma(n-1)+n$	$\Sigma(n-1)+2n+1$	$\forall n \geq 2$
4	$\Sigma(n)+2(n+1)$	$\Sigma(n-1)+2n+1$	$\forall n \geq 2$
5	$\Sigma(n)+m+1$	$\Sigma(n-1)+m+1$	$\forall n \geq 2,\ 0 \leq m \leq n-2$
6	$\Sigma(n)+n+2$	$\Sigma(n-1)+n+1$	$\forall n \geq 2$
7	$\Sigma(n)+2n+1$	$\Sigma(n-1)+2n$	$\forall n \geq 2$
8	$\Sigma(n)+n+m+3$	$\Sigma(n-1)+n+m+2$	$\forall n \geq 3,\ 0 \leq m \leq n-3$
9	$\Sigma(n-1)+2n+\sum_{\alpha=1}^{m+1} 2^{\alpha-1}+l$	$\Sigma(n-1)+2n+\sum_{\alpha=1}^{m+2} 2^{\alpha-1}+l$	$\forall n \geq 3,\ 0 \leq m \leq n-3,\ 0 \leq l \leq 2^{m+1}-1$
10	$\Sigma(n-2)+\Upsilon(n-1)+\sum_{\alpha=0}^{m}\Psi(\alpha,n)+l$	$\Sigma(n-1)+\Upsilon(n)+m+\sum_{\alpha=0}^{m}\Psi(\alpha,n)+l+1$	$\forall n \geq 1, 0 \leq m \leq 2^{n-1}-1, 0 \leq l \leq \Psi(m+1,n)-1$
11	$\Sigma(n-1)+2n+\sum_{\alpha=1}^{n-1}2^{\alpha-1}+m$	$\Sigma(n-1)+\Upsilon(n)+m+\sum_{\alpha=0}^{m}\Psi(\alpha,n)$	$\forall n \geq 2,\ 0 \leq m \leq 2^{n-1}-1$

Table 7: Transitions for the arrival of a Type $\mathbf{O}$ patient at rate λ_o.

	Index of the initial state	Index of the final state	
1	$\Sigma(n)$	$\Sigma(n-1)$	$\forall n \geq 1$
2	$\Sigma(n)+n$	$\Sigma(n-1)+n$	$\forall n \geq 1$
3	$\Sigma(n)+m+1$	$\Sigma(n-1)+m+1$	$\forall n \geq 2,\ 0 \leq m \leq n-2$
4	$\Sigma(n)+n+2$	$\Sigma(n-1)+n+1$	$\forall n \geq 2$
5	$\Sigma(n-2)+n-1$	$\Sigma(n-1)+2n+2$	$\forall n \geq 3$
6	$\Sigma(n-1)+2n$	$\Sigma(n-1)+2n+2$	$\forall n \geq 3$
7	$\Sigma(n)+2n+1$	$\Sigma(n-1)+2n$	$\forall n \geq 2$
8	$\Sigma(n)+n+m+3$	$\Sigma(n-1)+n+m+2$	$\forall n \geq 3,\ 0 \leq m \leq n-3$
9	$\Sigma(n-2)+2(n-1)+\sum_{\alpha=1}^{m+1} 2^{\alpha-1}+l$	$\Sigma(n-1)+2n+\sum_{\alpha=1}^{m+2} 2^{\alpha-1}+2^{m+1}+l$	$\forall n \geq 3,\ 0 \leq m \leq n-3,\ 0 \leq l \leq 2^{m+1}-1$
10	$\Sigma(n-2)+\Upsilon(n-1)+m$	$\Sigma(n-1)+\Upsilon(n)+2^{n-1}+\sum_{\alpha=1}^{2^{n-1}} \Psi(\alpha,\ n)+m$	$\forall n \geq 1,\ 0 \leq m \leq \Phi(n-1)-1$

Table 8: Transitions for the completion of remaining ED processes for a Type $\mathbf{E_P^T}$ patient at rate $iq\mu$ when i Type $\mathbf{E_P}$ patients are processed in server s_1.

	Index of the initial state	Index of the final state	
1	$\Sigma(n)+1$	$\Sigma(n-1)$	$\forall n \geq 1$
2	$\Sigma(n)+n+1$	$\Sigma(n-1)+n$	$\forall n \geq 1$
3	$\Sigma(n)+2(n+1)+1$	$\Sigma(n-1)+2n+1$	$\forall n \geq 2$
4	$\Sigma(n)+m+2$	$\Sigma(n-1)+m+1$	$\forall n \geq 2,\ 0 \leq m \leq n-2$
5	$\Sigma(n)+n+3$	$\Sigma(n-1)+n+1$	$\forall n \geq 2$
6	$\Sigma(n)+2(n+1)+2$	$\Sigma(n-1)+2n+2$	$\forall n \geq 3$
7	$\Sigma(n)+2n+2$	$\Sigma(n-1)+2n$	$\forall n \geq 2$
8	$\Sigma(n)+n+m+4$	$\Sigma(n-1)+n+m+2$	$\forall n \geq 3,\ 0 \leq m \leq n-3$
9	$\Sigma(n)+2(n+1)+\sum_{\alpha=1}^{m+2} 2^{\alpha-1}+l$	$\Sigma(n-1)+2n+\sum_{\alpha=1}^{m+2} 2^{\alpha-1}+l$	$\forall n \geq 3,\ 0 \leq m \leq n-3,\ 0 \leq l \leq 2^{m+2}-1$
10	$\Sigma(n-1)+\Upsilon(n)+m+\sum_{\alpha=0}^{m}\Psi(\alpha,n)+l$	$\Sigma(n-1)+\Upsilon(n)+m+\sum_{\alpha=0}^{m}\Psi(\alpha,n)+l+1$	$\forall n \geq 1, 0 \leq m \leq 2^n-2, 0 \leq l \leq \Psi(m+1,n)-1$
11	$\Sigma(n)+2(n+1)+\sum_{\alpha=1}^{n} 2^{\alpha-1}+m$	$\Sigma(n-1)+\Upsilon(n)+m+\sum_{\alpha=0}^{m}\Psi(\alpha,\ n)$	$\forall n \geq 1,\ 0 \leq m \leq 2^n-1$

Table 9: Transitions for the completion of the preparation/allocation of a bed at server s_2 at rate θ.

	Index of the initial state	Index of the final state	
1	$\Sigma(n-1)+2n+1$	$\Sigma(n-1)+n$	$\forall n \geq 1$
2	$\Sigma(n)+2(n+1)+2$	$\Sigma(n-1)+n$	$\forall n \geq 1$
3	$\Sigma(n-1)+2n+3$	$\Sigma(n-1)+2n+1$	$\forall n \geq 2$
4	$\Sigma(n)+2(n+1)+4$	$\Sigma(n-1)+2n+1$	$\forall n \geq 2$
5	$\Sigma(n-1)+n+m+2$	$\Sigma(n-1)+m+1$	$\forall n \geq 2,\ 0 \leq m \leq n-2$
6	$\Sigma(n-1)+2n+5$	$\Sigma(n-1)+2n+2$	$\forall n \geq 3$
7	$\Sigma(n)+2(n+1)+6$	$\Sigma(n-1)+2n+2$	$\forall n \geq 3$
8	$\Sigma(n-1)+\Upsilon(n)+2\sum_{\alpha=0}^{m}\Psi(\alpha,\ n)+m$	$\Sigma(n-1)+2n+\sum_{\alpha=1}^{n-1}2^{\alpha-1}+m$	$\forall n \geq 3,\ 0 \leq m \leq 2^{n-1}-1$
9	$\Sigma(n-1)+2n+\sum_{\alpha=1}^{m+3}2^{\alpha-1}+2l$	$\Sigma(n-1)+2n+\sum_{\alpha=1}^{m+2}2^{\alpha-1}+l$	$\forall n \geq 4,\ 0 \leq m \leq n-4,\ 0 \leq l \leq 2^{m+2}-1$
10	$\Sigma(n)+\Upsilon(n+1)+2\sum_{\alpha=0}^{m}\Psi(\alpha,\ n+1)$ $+m+k\Psi(m+1,\ n+1)+l+1$	$\Sigma(n-1)+\Upsilon(n)+\sum_{\alpha=0}^{m}\Psi(\alpha,\ n+1)+l$	$k \in \{0,\ 1\},\ \forall n \geq 1,$ $0 \leq m \leq 2^n-1,\ 0 \leq l \leq \Psi(m+1,\ n)$

Table 10: Transitions for the completion of remaining ED processes for a Type $\mathbf{E}_{\mathbf{P}}^{\mathbf{F}}$ patient at rate $i(1-q)\mu$ when i Type $\mathbf{E}_{\mathbf{P}}$ patients are processed in server s_1.

	Index of the initial state	Index of the final state	
1	$\Sigma(n-1)+1$	$\Sigma(n-1)$	$\forall n \geq 1$
2	$\Sigma(n)+2(n+1)+3$	$\Sigma(n-1)+2n+1$	$\forall n \geq 2$
3	$\Sigma(n-1)+m+2$	$\Sigma(n-1)+m+1$	$\forall n \geq 2,\ 0 \leq m \leq n-2$
4	$\Sigma(n-1)+n+2$	$\Sigma(n-1)+n+1$	$\forall n \geq 2$
5	$\Sigma(n)+2(n+1)+4$	$\Sigma(n-1)+2n+2$	$\forall n \geq 3$
6	$\Sigma(n)+2(n+1)+5$	$\Sigma(n-1)+2n+2$	$\forall n \geq 3$
7	$\Sigma(n+1)+2(n+2)+6$	$\Sigma(n-1)+2n+2$	$\forall n \geq 3$
8	$\Sigma(n-1)+2n+1$	$\Sigma(n-1)+2n$	$\forall n \geq 2$
9	$\Sigma(n)+2(n+2)$	$\Sigma(n-1)+2n$	$\forall n \geq 2$
10	$\Sigma(n-1)+n+m+3$	$\Sigma(n-1)+n+m+2$	$\forall n \geq 3,\ 0 \leq m \leq n-3$
11	$\Sigma(n)+2(n+1)+\sum_{\alpha=1}^{4} 2^{\alpha-1}-2^{k+1}+\sum_{\alpha=2}^{m+2} 2^{\alpha+1}+l-8$	$\Sigma(n)+2n+\sum_{\alpha=1}^{3} 2^{\alpha-1}-k+\sum_{\alpha=1}^{m+1} 2^{\alpha+1}+l-5$	$\forall k \in \{0,1\},\ n \geq 3,\ 0 \leq m \leq n-3,\ 0 \leq l \leq k$
12	$\Sigma(n)+2(n+1)+\sum_{\alpha=1}^{k+1} 2^{\alpha-1}+\sum_{\alpha=k}^{m+k} 2^{\alpha+1}-2^{k+1}+l$	$\Sigma(n-1)+2n+\sum_{\alpha=1}^{k} 2^{\alpha-1}+\sum_{k-1}^{m+k-1} 2^{\alpha+1}+2^{k}+l$	$\forall k \geq 2,\ n \geq k+1,\ 0 \leq m \leq n-k-1,\ 0 \leq l \leq 2^{k}-1$
13	$\Sigma(n+1)+2(n+2)+\sum_{\alpha=1}^{m+3} 2^{\alpha-1}-1$	$\Sigma(n-1)+2n+\sum_{\alpha=1}^{m+2} 2^{\alpha-1}-1$	$\forall n \geq 2,\ 0 \leq m \leq n-2$
14	$\Sigma(n+1)-3$	$\Sigma(n)-1$	$\forall n \geq 1$
15	$\Sigma(n)+\Upsilon(n+1)+\sum_{\alpha=0}^{2^{n+1}-2^{k}} \Psi(\alpha,n+2)$ $+\sum_{\alpha=0}^{m} \Psi(\alpha,\ k+1)+l$	$\Sigma(n-1)+\Upsilon(n)+\sum_{\alpha=0}^{2^{n}-2^{k-1}} \Psi(\alpha,n+1)$ $+\sum_{\alpha=0}^{m} \Psi(\alpha,\ k)+l$	$2 \leq k \leq N,\ \forall n \geq k-1,$ $0 \leq m \leq 2^{k-1}-1,\ 0 \leq l \leq \Psi(m+1,k-1)$
16	$\Sigma(n+1)+\Upsilon(n+2)-1$	$\Sigma(n)-1$	$\forall n \geq 1$

CHAPTER 3: PREDICTING EMERGENCY DEPARTMENT DISPOSITION DECISION FOR EARLY RESOURCE ALLOCATION

3.1 Introduction

While many studies have identified overcrowding as one ot the main issues in the ED management [66, 52, 54, 80], boarding has been identified as a factor contributing to ED crowding [77, 29, 12]. To alleviate crowding in the ED, various approaches that streamline the admission process from the ED have been suggested including the adoption of fast-track units, advanced patient triage strategy, and the implementation of six-sigma and lean programs [60, 57, 19, 92, 85, 34, 89, 90, 7, 67, 36, 51]. The latest data from the Centers for Medicare & Medicaid Services (CMS) shows median boarding times in across states and territories in U.S. in 2016 as Figure 13, with a median boarding time of 2 hours and 16 minutes [31].

The motivation of this study is the modeling of ED disposition decision prediction that can lead to the proactive coordination of resource preparation, e.g., allocation and management of inpatient beds. Predictive analytics is certainly receiving a great deal of attention in recent years for improving healthcare service quality [5].In the context of ED operations management, while there is good progress with ED patient admission prediction modeling research

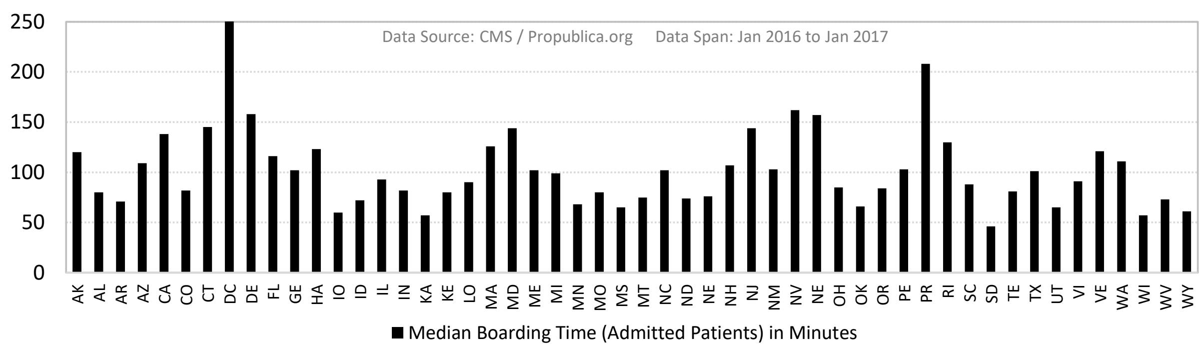


Figure 13: ED Patient Census by Hour of Day during Weekdays

[78, 14, 96, 21], the models are lacking in their resolution in order to allow operationalization of prediction outcomes in real-world settings. Just predicting that a patient will be "admitted" (i.e., binary classification: admission vs. discharge) will not necessarily allow full proactive coordination of resource across the ED-IU network for streamlined patient admission and flow. In this study, we put emphasis on the operationalization of prediction information by modeling predictions that reflect the more realistic hierarchical structure in actual disposition decisions. In general, IUs can be divided into several units that usually do not share the same resource features across each other. Hence, patient disposition decision prediction models need to identify the likely inpatient care unit for admission for proper proactive coordination. Generally, inpatient care can be categorized into three main types based on the intensity of required care [45]: general practice care, telemetry/stepdown care, and intensive care. In turn, the three different care levels define three main types of inpatient units, i.e., general practice unit (GPU), telemetry unit (TU, also known as stepdown unit), and intensive care unit (ICU). Furthermore, our prediction modeling strategy incorporates the progressive nature of ED care processes, where more clinical information is revealed and accumulated for an ED patient as he/she goes through more ED processes. This study is a response to the need for better coordination between the ED and IUs as well as a growing attention to predictive medicine, especially relevant to emergency medicine.

3.2 Materials and Methods

3.2.1 Experimental Setting

The study is based on Electronic Health Record (EHR) data collected at the ED of an academic urban trauma center in SE-Michigan. The data items were collected for the period from May 2014 to April 2016 and covers 184,895 patient visits. After accounting for abnormal departures, including AMA (against medical advice), dismissed patients, patients died in the ED, LWCS (leaving without completing services), and transfers to other facilities, there were 175,500 observations. After removing patients who went to units that are not regarded as regular IUs and are difficult to be categorized as one of the IUs, e.g., catheterization lab, peri-operative unit, and post-anesthesia care unit, 172,809 patients remained in the dataset, which corresponds to 93.5% of the total visits. As noted earlier, this hospital includes three main types of IUs: GPU, TU, and ICU. Each main unit is further categorized based on specialties, which include general internal medicine GPU, pulmonary GPU, surgical ICU, medical ICU, etc. While a disposition decision can specify the most proper specialty unit for a patient, if all beds in that unit are in use, the patient is generally transferred to the second most proper inpatient unit and so on (i.e., overflow). In such cases, bed managers decide on the alternative unit that can accommodate the patient within the same main unit (e.g., assigning the general medicine GPU for a pulmonary GPU patient), then the disposition decision is updated accordingly. Since the main goal of this research is to facilitate proactive coordination to reduce ED patient boarding times, we consider the overflow scheme so that the classification results can help inpatient units initiate their bed management processes earlier. It is worth noticing that the features of inpatient beds and accessories are common over a main unit, and this in turn usually constrains overflow to happen within the same main unit. Besides these three IUs, we also consider the observation unit (OU). OUs are increasingly used as a short-stay clinical decision unit for ED patients who require further observation to make the final disposition decision. Usually the final disposition decisions for OU patients are made within 24 hours after they enter the OU. Even though the OU is not a part of regular IUs in most hospitals, it plays a significant role to control demand to IUs, and the number of OU patients is not negligible (around 6% in the target hospital). Therefore, we include the OU class as one of disposition decisions in this study.

Considering how the five classes (the three IU classes, OU class and discharge class) are defined, we can notice that the classification model should be able to discriminate the clinical 'intensity' of ED patients, which makes the classification task challenging. Moreover, since disposition decisions cannot always be completely objective, it is important to check how effectively machine-based prediction models can assign ED patients into the five classes by exploiting clinical, demographic, and operational data. In addition, different classification structures, e.g., admission-discharge binary classification and five-class classification, can provide various levels of utility for proactive coordination. For instance, while the five-class classification would be helpful for managing resources that require unit-specific coordination, the binary classification can still provide useful information for proactive resource coordination, including a more efficient hospital resource allocation for expediting discharge processes and better ED service for patients predicted to be admitted. Therefore, distinct levels of outcome granularity need to be investigated to assess the value of ED disposition decision prediction.

3.2.2 Disposition Class Structure

We define three distinct levels of classification group structures for modeling disposition decisions. At the first level (C1), the outcome of disposition decision classification is the binary admission decision, i.e., admission vs. discharge. The OU class is included in the admission class. The C1 group structure has been adopted in most of the ED disposition decision prediction modeling research to date [78, 14, 96, 21]. Since it is a relatively simple binary classification task, it could produce the most accurate results for those two classes. While the use of prediction outcomes from this binary classification scheme would mostly be limited within the ED (i.e., the way to provide care to two different patient types), the prediction outcomes can also be used when coordinating resources that are common and shared over all units (e.g., admission approvals/processing). At the second level (C2), the admission class at C1 is further segmented into two sub-classes, i.e., IU vs. OU admission. While the IU class is regarded as official admission, OU patients may not be considered as admitted patients depending on the hospitals even though OU patients could have features that are clinically similar to IU patients. Rather, the OU treatment is often regarded as an 'extended ED care', and patient transfers to OUs can be different from regular IUs in many hospitals. Finally, at the most granular level (C3), the inpatient admission class at C2 is further categorized into three main IU classes, i.e., ICU, TU, and GPU. We believe that the C3 group is where the most significant operational benefit can be derived for coordinating patient flow across the ED-IU network. Most hospitals in the U.S. physically separate ICU, TU, and GPU, pooling resources (e.g., environmental services) within each unit exclusively. We can thus symbolically represent each class group as follows: C1={admission, discharge}, C2={IU, OU, discharge}, and C3={ICU, TU, GPU, OU, discharge}. Figure 14 depicts the overall structure of the studied classification problem structure. It is expected that as we increase the granularity of prediction (i.e., from C1 through C3), the prediction problem becomes more challenging.

We acknowledge that each unit at C3 can be further classified. For instance, in the studied hospital, there are twelve distinct units based on specialty of care in GPU. They include internal medicine, nephrology, obstetrics/gynecology, orthopedic surgery, general surgery, neurology, transplant, hospice, hospitalist, pulmonary, and hematology/oncology/bone marrow transplant. However, a patient going to any one of these units can often be admitted to any other unit in the GPU based on common overflow policies and resource availability. Moreover,

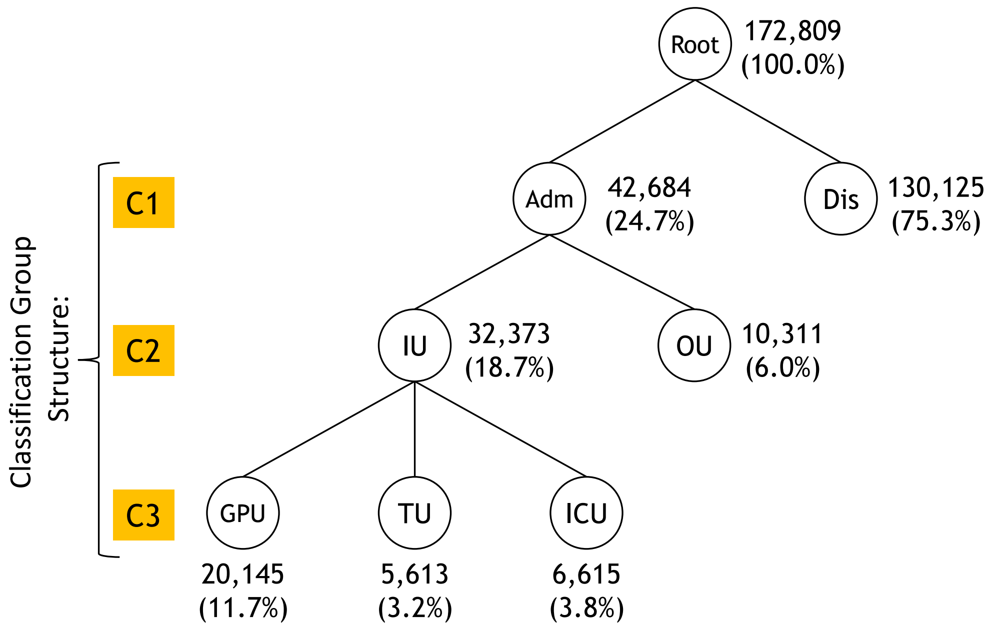


Figure 14: Hierarchical Structure of ED Disposition Decision Along with Number and Percentage of Patients

about half the time, physicians in the hospital make disposition decisions at the main IU level without further specifying any specialty. Therefore, in this study, we focus on classifying those three five classes (i.e., GPU, TU, ICU, OU, and discharge) at C3. This focus was also chosen due to its higher utility of predictions.

3.2.3 ED Patient Disposition Classification Strategy

It is worth introducing the proposed classification strategy briefly with few associated concepts from data science. The ED disposition prediction inherently presents a 'hierarchical' classification structure (as shown in Figure 14). The disposition decision classification is a mandatory-leaf classification problem in that the most proper class is always found at the lowest level (overall structure resembles a tree, and the classes of interest are the leaf nodes).

For instance, an IU patient at C2 should be one of the GPU, TU, or ICU patients at C3. Also, it is tree-structured in that no two nodes share a common child node. For instance, any two patients that are classified as belonging to the IU class and the OU class respectively at C2 can never be classified as belonging to the same class at C3. To tackle hierarchical classification problems, there are multiple possible approaches, but the various modeling approaches can be categorized into two main categories, which are the 'big-bang' approach and 'top-down' approach [93]. While the top-down approach starts its classification task from the parent node and uses the obtained prediction outcomes for classification at its child nodes, the big-bang approach classifies the most proper class for the full problem with a single model. We take notice that the ED disposition classification problem is an 'imbalanced' classification problem in that most patients are sent home or discharged (75.3% in the target hospital). Most classification methods tend to suffer in handling imbalanced classification problems. Given the hierarchical structure of ED disposition and the class membership imbalance at C1 level, the top-down approach would incur serious challenges at the down-stream levels (C2 and C3). This is due to the tendency of classification models to favor classes with dominant membership (discharge class in our case) and possibly produce many false negative predictions at C1 that can propagate down the hierarchy. Especially when the problem is of a mandatory-leaf node structure, this issue cannot be easily handled (i.e., false negative cases will spread throughout the entire levels). Therefore, we choose to adopt the big-bang approach for predicting disposition decisions. Moreover, since each prediction class level could bring about their own operational benefits, we also build classification models at each level. In summary, we model and analyze the mandatory-leaf node, tree-structured ED disposition classification problem with the big-bang classifier per level approach.

3.2.4 Temporal Aspects of ED Disposition Prediction

We consider not only the hierarchical structure of inpatient units but also the decreasing diagnostic uncertainty patterns characterized at ED care stages. To study how disposition decision uncertainty reduces as ED care processes proceed, we identify four different ED care stages for each patient: 'ED arrival' (T1), 'triage complete' (T2), 'provider encounter' (T3), and 'first lab/imaging results returned' (T4). At T1, ED patients come with some basic information such as arrival time/mode, health history (including prior ED visit history), and demographics. Then patients go through triage processes where patients' vital signs and chief complaints are recorded (T2). T2 is when most admission decision prediction models have been built in the literature [96, 24, 37, 59, 49]. At T3, ED care providers examine patients and issue orders of lab/imaging tests. We incorporate the first set of lab/imaging test items ordered by care providers at T3 (upon examining the results from these tests, care providers can order additional tests downstream within the ED care cycle, which are outside the scope of data employed for the T3 setting). The chief objective of this research study being 'proactive' coordination within the ED-IU network, there is no point in making accurate disposition predictions using information that arrives too late in the ED care cycle when the final disposition decisions is already available. T4 indicates the time when the results for the first set of lab/imaging orders are fully reported. Table 11 describes the time spans between these different ED care stages in the study hospital. Of note median door to disposition is four and a half hours at the study hospital.

Table 11: Summary Statistics of Supply and Demand around the ED-IU Network

ED care stages	1Q / median / 3Q (in minutes)
T1 to T2	11.3 / 17.6 / 29.0
T2 to T3	19.5 / 47.5 / 92.8
T3 to T4	35.0 / 55.0 / 90.0

3.2.5 Data

Tables 12 and 13 summarize data items, including univariate statistics for demographics, acuity level, chief complaints (five example complaints), vital signs, and bivariate statistics for lab test results and imaging test items (five example items respectively). We ran t-tests for the vital signs and lab test results. As stated earlier, we focus on the first set of lab/imaging tests ordered at the first encounter with care providers. Due to the size of the feature set and different class levels (i.e., C1, C2, and C3), we do not present analysis results for all the variables. Rather, we selectively present the multivariate analysis statistics for some key features that are shown to be statistically associated with outcome classes at C3 in Section 3.3. The entire dataset is split into two parts, first 60% of the patient visits for training and the rest for testing. We only report prediction results obtained from the testing dataset in Section 3.3.

3.2.6 Feature Engineering

As expected, we observed that most of the data items collected are categorical variables. Demographic and operational information, including gender, arrival time (hour), arrival method, and insurance plan, are inherently categorical. While there is no definite rule to categorize age into groups, we use the age categorization rule applied by the CMS for personal health care spending study [32]. For variables that have too many categories, such as arrival method and

Table 12: Univariate Analysis of the Variables Used in the Model

Variable	Median/mean/ count/proportion	(Q1, Q3)	Comments
Age (median)	45.0	(27.0, 60.0)	
Number of arrivals per hour (median count)			
12AM∼4AM	5.0	(3.0, 7.0)	8.6%
4AM∼8AM	4.0	(3.0, 5.0)	6.8%
8AM∼12PM	13.0	(10.0, 16.0)	21.1%
12PM∼4PM	15.0	(12.0, 17.3)	24.5%
4PM∼8PM	14.0	(11.0, 16.0)	22.3%
8PM∼12AM	10.0	(8.0, 12.0)	16.7%
Gender (proportion)			
Female	54.9%		
Male	45.1%		
Insurance plan			285 different plans
Patient arrival modes			62 different modes
Any prior ED visit history to ED (within last 30 days)	23.1%		0.4 prior visits/patient on average
Acuity levels (proportion)			
ESI level-1	2.0%		
ESI level-2	36.3%		
ESI level-3	54.4%		
ESI level-4	6.3%		
ESI level-5	0.5%		
Chief complaints (proportion, 5 example complaints)			
Abdominal pain	10.0%		
Shortness of breath	5.5%		
Chest pain	5.5%		
Back pain	3.6%		
Headache	3.3%		

ESI, emergency severity index.

Table 13: Bivariate Analysis of the Variables Used in the Model

	Discharge		Admission		
Variable	Mean/ proportion	(Q1, Q3)	Mean/ proportion	(Q1, Q3)	Comments
Vital signs (mean)					p–Value
Temperature (°F)	98.3		98.4		> 0.5
Pulse (rpm)	86.4		89.5		< 0.001
Systolic BP (mm Hg)	133.4		137.1		< 0.001
Diastolic BP (mm Hg)	78.5		79.1		< 0.001
Respirations (rpm)	18.6		19.7		< 0.001
Pulse oximetry (%)	98.2		97.2		< 0.001
Results of lab tests ordered at first encounter with doctor (mean, ordered proportion)					p–Value
Troponin I (ng/mL)	0.04 8.6%	(0.04, 0.04)	0.8 43.6%	(0.04, 0.14)	<0.001
BNP (pg/mL)	177.9 3.3%	(13.0, 110.0)	538.5 8.9%	(38.0, 633.0)	<0.001
PT/INR/PTT (sec)	15.2 14.1%	(13.2, 14.6)	18.0 18.7%	(14.1, 18.7)	<0.001
Lactate blood (mmol/L)	1.7 8.6%	(1.1, 2.1)	2.2 43.6%	(1.0, 2.4)	<0.001
CBC /w differential	8.0 32.3%	(5.8, 9.6)	10.0 73.8%	(6.2, 11.9)	<0.001
Imaging text orders after first encounter with doctor (ordered proportion)					
Chest x-ray	3.4%		25.8%		
CT head	3.2%		11.0%		
CT abdomen pelvis	0.9%		1.9%		
Acute abdominal series	1.8%		3.1%		
CT pulmonary embolism	0.2%		0.7%		

BNP, brain natriuretic peptide; PT/INR/PTT, prothrombin time/international normalized ratio/partial thromboplastin time; CBC, complete blood count.

insurance plan, the categories are condensed based on both frequency and meaning of categories. All of the demographic and operational information is immediately available at the arrival of patients. Through natural language processing (NLP), chief complaints and associated text were encoded as vectors. For vital signs, we categorized the numerical variables based on National Institutes of Health (NIH) National Library of Medicine [74].There are 357 types of lab/imaging tests that were ordered at the first encounter with care providers in total within this patient set. Lab test results are initially entered as numerical values, but the HIT system also provides result flags that automatically categorize numerical lab values based on preset category boundaries. For instance, for magnesium testing, the system categorizes its values into five groups: low panic $\leq 0.9 <$ low $\leq 1.8 <$ normal $\leq 2.2 <$ high $\leq 4.0 <$ high panic. Since the categories are clinically valid, when result flags are available, we choose to use the categories rather than numerical values for easing the burden on the classification model. For lab tests that do not have the result flags, supervised feature discretization techniques are applied to categorize the numerical values (in particular a Chi-square based algorithm named ChiM [58] is mainly used). Unlike the lab tests, the results for radiology tests were not available electronically in a form that can be readily coded. Therefore, for radiology tests, we only encompass the types of tests ordered into the prediction models.

3.2.7 Modeling

We applied various classification modeling techniques that include multinomial logistic regression, artificial neural networks, support vector machines, and random forests. Among these approaches, a well-established approach, multinomial logistic regression produced the best prediction performance overall. Therefore, in this study, we focus on reporting prediction results generated by multinomial logistic regression models.

Table 14: Comparing Prediction Model Performance at Different Care Points for C1

C1 class		T1	T2	T3	T4
Accuracy		77.9%	84.6%	85.5%	87.1%
Admission (24.7%)	Sensitivity	30.2%	56.3%	58.7%	63.9%
	Precision	60.0%	74.8%	76.6%	79.2%
Discharge (75.3%)	Sensitivity	93.4%	93.8%	94.2%	94.6%
	Precision	80.4%	86.8%	87.5%	89.1%

3.3 Results

3.3.1 C1 Classification Group

As shown in Table 14, at door (T1), without any clinical information, we can predict admission decision of ED patients with 77.9% accuracy. While the sensitivity of the admission class is around 30% at T1, the false positive rate (1-sensitivity of the discharge class) of the model is less than 7%. While incorporating more information allows the model to enhance its performance (from T1 through T4), the biggest improvement is made at triage (T2) where 56.3% of admitted patients are correctly predicted with less than 7% of false positives. With the first set of lab test orders (at T3), the improvement of prediction performance is not noticeable. However, when the test results return, there is additional information that can lead to another rise in prediction quality. Unlike the admission prediction work done at the ED in an Israeli hospital [14], the results of lab tests (from T2 to T4) seem more useful than the decisions to order specific lab tests (from T2 to T3) to enhance the prediction in our experiment hospital. This may be because of the inclusion of results of extensive lab tests.

Table 15: Comparing Prediction Model Performance at Different Care Points for C2

C1 class		T1	T2	T3	T4
Accuracy		76.6%	82.3%	83.2%	84.9%
IU (18.7%)	Sensitivity	23.5%	56.4%	58.5%	64.6%
	Precision	49.1%	66.8%	69.9%	73.1%
OU (6.0%)	Sensitivity	0.0%	6.6%	9.4%	15.3%
	Precision	-	44.3%	44.6%	55.3%
Discharge (75.3%)	Sensitivity	96.0%	95.3%	95.7%	95.8%
	Precision	79.1%	85.6%	86.2%	87.8%

3.3.2 C2 Classification Group

In the C2 class group, just like the C1 group case, the steepest increase in prediction ability is gained at triage (as shown in Table 15). It is noticeable that the OU class, being an intermediate class between the IU class and the discharge class, does not seem to have clear clinical distinction, compared to the IU and discharge classes. The lab test result items (T4) prove their predictive power at the C2 level as well. In particular, the precision of classification for the OU class exceeds 50% with doubled the sensitivity from T2, by incorporating lab test results.

3.3.3 C3 Classification Group

It is plausible that C3 class group classification would provide the most informative results for proactive resource coordination. Therefore, we present the prediction results from various angles so that detailed analysis can become possible. First, Figure 15 reports the sensitivity (Figure 15(a)) and prediction (Figure 15(b)) values of each class resulted from each prediction model. For the C3 group, we choose to report prediction results with graphics to clearly represent how prediction quality evolves as more information is gathered for each class. The C3 group classification is an imbalanced multiclass classification problem with fairly small portions of the minor class patients (especially ICU and TU with 3∼4% respectively), compared to the major class (i.e., discharge) that consists of 75.3% of the entire patients. Figure 15 shows that each class reveals different behaviours. At T1, the prediction model fails to detect those minor classes. In fact, most of the patient visits are predicted as the discharge class at T1, which proves the insufficient information to detect the minor classes. At T2 (triage), prediction quality varies depending on classes. The highest level of both sensitivity and prediction is obtained for the ICU class. It indicates that ICU patients possess most distinct features at the triage stage, and even though the number of ICU patients is small (only 3.8%), around 40% of ICU patients can be detected with about 50% of prediction precision. Triage information also carries a considerable amount of information for GPU class prediction. However, even though the number of GPU patients is comparatively large among the minor classes, which usually makes prediction easier, the prediction performance levels are inferior to those of the ICU class, emphasizing the clear clinical distinction of ICU patients at triage. We also observe that the prediction of the TU and OU classes is markedly less accurate, compared to the ICU and GPU classes at T2. It would mean that while ICU and GPU patients possess clinically distinguishable information at triage, TU and OU patients do not have clear clinical trajectory at triage, remaining as in-between states, i.e., the TU between the ICU and GPU, and the OU between the GPU and discharge. However, at T3, we can recognize that the physicians' decisions to order certain lab/imaging tests prove useful for predicting the TU class. It seems that the clinical care intensity and needs of TU patients are difficult to be estimated at triage. However,

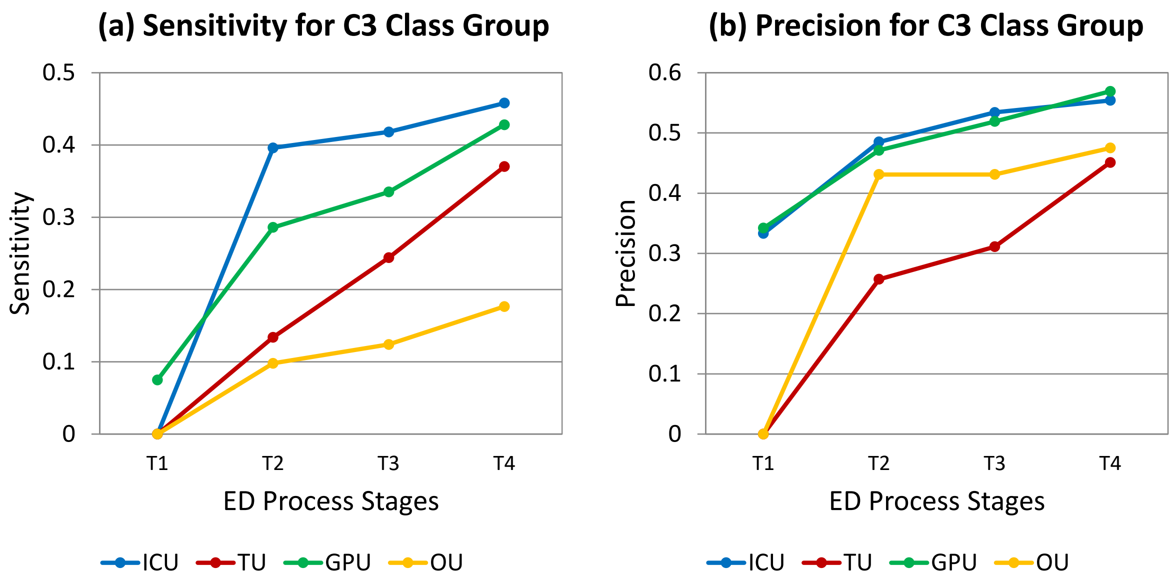


Figure 15: Comparing Prediction Model Performance at Different Care Points for C3

doctors would start to judge the clinical severity and required care services (especially constant cardiac monitoring) for patients based on triage information and try to distinguish TU patients by ordering additional lab/imaging tests.

The results in Figure 15 also show that, generally, information gain from lab test order results is larger than that from lab test order items. Especially, the precision levels of prediction are greater than 45% in all the five classes at T4, while sensitivity levels vary. The greatest sensitivity level is gained for ICU and GPU patients with more than 42% (apart from discharged patients with 96.4%), and 36.4% of TU patients can also be detected. With the comparatively high-performance levels for the ICU class at T2, the additional gain of prediction ability for the ICU class at T4 is not drastic, compared to other minor units. The distinct progression behaviours clearly indicate that any proactive task initiation strategy that utilizes ED disposition decision prediction should consider the different levels of prediction quality obtained at different care stages for each different unit. Setting the discharge class as the negative class, the false positive rate at the C3 class group is only 3.8% at T4.

3.3.4 Prediction Threshold Analysis at C3

It is important to understand how the prediction results can be further exploited to enable the operationalization of prediction information. Especially, being a probabilistic classifier, multinomial logistic regression outputs not only a predicted class for a patient but also an estimated probability value for each of the classes for a patient. The estimated probability values tell us about the confidence of membership at each disposition class. By imposing a probability threshold in making prediction, we can set up the level of confidence with which a disposition prediction is made for a patient. In other words, by applying a probability threshold, the model does not make any prediction unless any one of the classes has a higher probability value than the threshold. Therefore, as we set a higher threshold, prediction would become more reliable with increasing precision values. Figure 16 shows how the sensitivity and precision of prediction behaves for each class among the patients who have higher probability values than the different levels of the threshold (the solid lines with the left y-axis). It also displays how the fraction of patients included in the analysis decreases as the threshold value increases (the dashed line with the right y-axis). For instance, if we impose 60% threshold to the ICU class, the sensitivity of ICU class prediction among the fraction of the patients increases from 45.8% to 62.1%, and the precision of the prediction would increase from 55.4% to 68.8% while 82.6% of the patients are included in the analysis. It is noticeable that although the higher threshold values tend to bring higher sensitivity values for the ICU and TU class patients, it does not affect the GPU and OU class patients in the same way. When the model focuses on the patients who have high probability values for the selected classes, there are more predictions made for ICU and TU classes than the GPU and OU classes, compared to their actual distri-

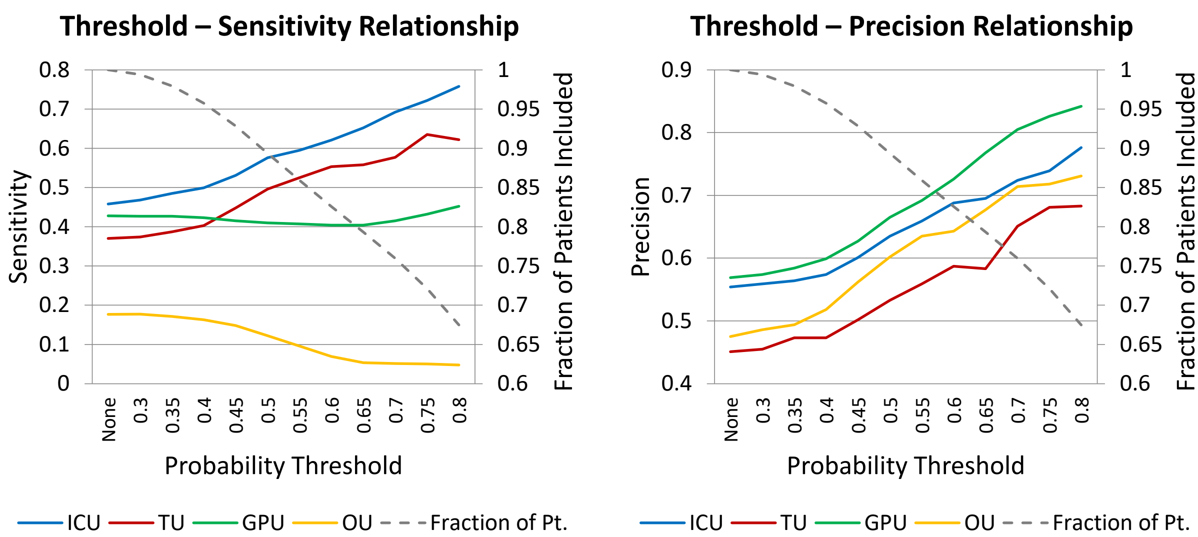


Figure 16: Performance Analysis According to Different Threshold Probability for Making Prediction

bution. This implies that the predictions of the ICU and TU classes have higher confidence (with higher probability values) than those of the GPU and OU classes because of their clinical distinctiveness.

3.3.5 Feature Analysis at T4

Besides the various prediction performance behaviours shown in the C3 classes, the importance of each feature also varies depending on the classification group. Table 16 reports the statistically most significant features at T4, calculated by the chi-square test at different classification groups. The study hospital has six compartmental primary care areas. Triage nurses determine the most appropriate primary care area for an ED patient based on the clinical information and acuity level of the patient. This feature presents high importance for disposition decision prediction in all of the three classification groups. As the classification group becomes more granular (i.e., from C1 through C3), more lab test result features become important to predict the classes. One observation from the feature analysis is that the significance of a feature is monotonically changed from C1 through C3 for most of the features. In other words, as we increase the granularity of classes, the importance of a feature keeps either increasing or decreasing. It happens to 34 among the 36 features selected within the top 25 important features at T4 in Table 16. In particular, electrocardiogram (ECG), chest x-ray, blood culture, computed tomography (CT) head without contrast, and CT angiography head neck with contrast are lab/imaging test results/items that have ever increasing importance as prediction classes are subdivided from C1 to C2 and C3. In addition, vital signs measured at triage become less important as the modelling proceeds from C1 through C3. Interestingly, as the prediction classification structure becomes more detailed (from C1 through C3), all non-lab/imaging test results/items become lower in the importance order except chief complaints (notice that the order of importance of the chief complaint items in Table 16 keeps increasing as the disposition classes become more granular).

Table 19 compares the distribution of some informative feature values over five classes in the C3 class group. The distributions are derived from the entire dataset (combined training and testing sets). The probability value is highlighted in bold when the probability at a minor class exceeds 30%, and the probability of the discharge class surpasses 95%. It is noticeable that patients with “high panic” results in the Troponin I test are likely to be admitted to the TU with 49% probability. Also, 43% of ESI level-1 patients have been admitted to the ICU.

3.4 Discussion

We attempted to frame ED disposition prediction as an analytics problem, seeking proactive task initiation for ED admissions. To the best of our knowledge, this work is the first attempt to define the ED disposition decision prediction as a hierarchical multi-class classification prob-

Table 16: Important Features to Predict Disposition Decision Classes at T4

Order of importance	C1 classification structure	C2 classification structure	C3 classification structure
1	Primary care area	Primary care area	IMG: Chest x-ray
2	LAB: Troponin I	IMG: Chest x-ray	ECG: 12-lead
3	Acuity level	ECG: 12-lead	Primary care area
4	LAB: PT/INR/PTT	LAB: Troponin I	LAB: Troponin I
5	LAB: CBC w/ differential	Acuity level	LAB: BNP
6	Age	LAB: PT/INR/PTT	LAB: Blood culture
7	LAB: Basic metabolic	LAB: CBC w/ differential	Acuity level
8	LAB: Lactate whole blood	LAB: Basic metabolic	LAB: PT/INR/PTT
9	LAB: BNP	LAB: Blood culture	LAB: Lactate whole blood
10	LAB: ECG 12-lead	Age	IMG: CT head
11	IMG: Chest x-ray	LAB: Lactate whole blood	CC: Chest pain
12	Insurance plan	LAB: BNP	LAB: CBC w/ differential
13	LAB: Blood gas venous	Insurance plan	LAB: Basic metabolic
14	LAB: Magnesium	LAB: Blood gas venous	CC: Shortness of breath
15	Arrival method	LAB: Magnesium	Age
16	LAB: Blood culture	CC: Shortness of breath	LAB: Blood gas venous
17	LAB: Liver profile	CC: Chest pain	Insurance plan
18	LAB: Phosphorus	Arrival method	LAB: Magnesium
19	LAB: POC glucose	IMG: CT head	CC: Stroke rule out
20	VT: Pulse oximetry	LAB: Liver profile	CC: Altered mental status
21	CC: Shortness of breath	LAB: Phosphorus	LAB: Liver profile
22	VT: Temperature	VT: Temperature	IMG: CT angiography head neck
23	LAB: Blood gas arterial	VT: Pulse oximetry	VT: Pulse oximetry
24	IMG: CT head	CC: Altered mental status	VT: Temperature
25	VT: Respirations	LAB: POC glucose	LAB: Urine culture

CC, chief complaint at triage; IMG, radiology (imaging) tests;LAB, lab tests; VT, vital signs at triage; POC, point-of-care. Only lab tests have test results.

Table 17: Distribution of Each Feature Value over C3 Classes (1)

Feature values	Number of cases (172,809 in total)	ICU (0.04)	TU (0.03)	GPU (0.12)	OU (0.06)	Discharge (0.75)
			(a) Chest x-ray			
Yes	20,293	0.21	0.13	0.23	0.14	0.28
			(b) Troponin I			
High panic	2,009	**0.34**	**0.49**	0.13	0.01	0.03
High	5,862	0.20	0.25	0.30	0.11	0.14
Normal	31,471	0.10	0.07	0.21	0.18	0.44
			(c) Basic metabolic			
High panic	131	**0.57**	0.06	**0.34**	0.01	0.02
High	1,034	0.20	0.08	0.27	0.07	0.37
Normal	82,771	0.06	0.06	0.17	0.10	0.61
Low	12,980	0.11	0.07	**0.31**	0.09	0.43
Low panic	400	**0.41**	0.09	**0.43**	0.03	0.04
			(d) BNP			
High	9,190	0.15	0.24	0.28	0.12	0.19
Normal	6,089	0.08	0.08	0.19	0.24	0.41
			(e) Insurance plan (four distinct examples)			
Plan 1	19,675	0.08	0.07	0.22	0.10	0.54
Plan 2	3,577	0.08	0.09	0.26	0.13	0.44
Plan 3	5,314	< 0.01	< 0.01	0.03	0.02	0.94
Plan 4	4,067	0.03	< 0.01	0.03	0.04	0.90

Table 18: Distribution of Each Feature Value over C3 Classes (2)

Feature values	Number of cases (172,809 in total)	ICU (0.04)	TU (0.03)	GPU (0.12)	OU (0.06)	Discharge (0.75)
(a) Primary care area						
Area 1	28,898	0.21	0.15	0.21	0.14	0.29
Area 2	47,008	< 0.01	0.01	0.14	0.07	0.77
Area 3	42,453	<0.01	0.01	0.13	0.06	0.79
Area 4	53,086	<0.01	< 0.01	0.01	0.01	**0.98**
(b) Acuity level						
ESI level-1	4,443	**0.43**	0.07	0.17	0.06	0.28
ESI level-2	82,674	0.07	0.07	0.20	0.11	0.55
ESI level-3	124,011	0.01	0.01	0.07	0.04	0.88
ESI level-4	14,456	< 0.01	< 0.01	< 0.01	< 0.01	**0.99**
ESI level-5	1,036	0	< 0.01	<0.01	0	**0.99**
(c) Age						
~ 2	1,999	0.03	0	< 0.01	0	**0.97**
$3 \sim 5$	2,892	0	0	0	0	**1**
$6 \sim 18$	6,679	<0.01	< 0.01	0.01	< 0.01	**0.98**
$19 \sim 44$	64,780	0.01	< 0.01	0.07	0.03	0.88
$45 \sim 64$	38,504	0.05	0.04	0.13	0.08	0.70
$65 \sim 84$	12,580	0.09	0.08	0.23	0.11	0.50
$86 \sim$	2,746	0.10	< 0.11	0.27	0.13	0.39
(d) Arrival method (four distinct examples)						
EMS 1	16,309	0.08	0.04	0.15	0.08	0.65
EMS 2	10,433	0.09	0.05	0.21	0.10	0.56
Public transportation	3,127	< 0.01	< 0.01	0.03	0.03	0.94
Walk in	5,671	0.01	< 0.02	0.07	0.86	**0.98**

EMS, emergency medical services

Table 19: Distribution of Each Feature Value over C3 Classes (3)

Feature values	Number of cases (172,809 in total)	ICU (0.04)	TU (0.03)	GPU (0.12)	OU (0.06)	Discharge (0.75)
(a) Temperature at triage (°F)						
$\sim$ 95.0	339	0.17	0.07	0.20	0.05	0.50
95.1 $\sim$ 97.7	26,457	0.04	0.04	0.13	0.07	0.73
97.8 $\sim$ 99.1	125,866	0.02	0.03	0.10	0.06	0.79
99.2 $\sim$ 100.0	6,909	0.03	0.03	0.16	0.05	0.72
100.1 $\sim$	3,538	0.06	0.03	0.24	0.03	0.64
(b) Pulse oximetry at triage (%)						
$\sim$ 94.0	7,476	0.14	0.09	0.29	0.08	0.41
94.1 $\sim$	160,677	0.03	0.03	0.11	0.06	0.77
(c) Respirations at triage (rpm)						
$\sim$ 11	215	0.33	0.04	0.15	0.04	0.44
12 $\sim$ 18	115,557	0.03	0.02	0.10	0.06	0.79
19 $\sim$ 25	44,242	0.05	0.04	0.15	0.07	0.70
26 $\sim$	7,558	0.14	0.07	0.19	0.04	0.56
(d) Pulse at triage (rpm)						
$\sim$ 59	6,263	0.05	0.06	0.10	0.10	0.69
60 $\sim$ 100	128,533	0.03	0.03	0.10	0.06	0.78
100 $\sim$	35,429	0.07	0.05	0.17	0.05	0.67

lem, categorizing the admission patients into more detailed classes so that the outcomes of the prediction can become practically useful for proactive coordination across the ED inpatient unit network. For the operationalization of the prediction information, more detailed investigation of prediction performance throughout ED care processes should be conducted. This study shows that in the study hospital, triage presents a significant amount of information to predict the disposition decision, especially for the ED patients headed to ICU. As patient information is accumulated through lab test orders and their results, the prediction power gradually increases. However, this can vary depending on the classes. It implies that a proper proactive task initiation strategy could vary in different classes depending on their own uncertainty reduction trends. This study also provides insights into physicians' disposition decisions and lab test result values through large scale data analysis.

3.5 Limitations

One of the limitations of this work is that we cannot guarantee whether the models have exploited the collected information to the greatest extent. For instance, we discretized the numerical variables such as lab test results into finite levels to transform them into categorical variables. However, more advanced data-driven methodologies, such as deep learning approaches, would be able to better extract features by thoroughly searching any interrelationship between the numerical variables. These approaches would compromise the repeatability and consistency of prediction, but the methods are rapidly becoming mature with technical advances. Another limitation of our work is that there is additional information in the EHR that could have been included in the models to precisely measure predictive power at each care step. One of these is notes entered by care providers. It is expected that a lot of clinically significant information is electronically recorded in physician notes in various formats including text. Since the note data items contain refined information that comes from the interpretation of clinical examination for patients, the inclusion of physician notes can significantly improve the prediction ability for disposition decisions.

3.6 Conclusion

This work builds on a growing academic and industrial attention to the usefulness and feasibility of proactive task initiation in health care that can be empowered by predictive analytics. Multi-class prediction that considers the actual patient flow and resource management schemes around the ED can help with effective early task initiation for admitted ED patients. The prediction system in this study reveals how prediction performances for each class evolves as ED patients go through ED care steps. Future studies will incorporate more advanced modeling techniques to exploit useful data items that are readily available in HIT systems in the ED.

CHAPTER 4: CONCLUSIONS AND FUTURE RESEARCH

This dissertation contributes to and advances the literature in exploring proactive coordination in healthcare service systems. Coupled with advanced analytics and EHR data from healthcare facilities, a promising coordination paradigm of early task initiation is proposed. In the proposed proactive coordination scheme, demand for certain care/operations services in a healthcare facility is predicted early and communicated to different department/support systems for them to take necessary steps "proactively" to enhance care quality and operational efficiency. Operations research enhanced by predictive analytics that utilizes modern machine learning methods can facilitate proactive coordination for dramatically improved patient flow and health outcomes.

Unlike the existing coordination settings that only involve patients, clinical/operational service providers, and resources, the proactive coordination should take different aspects of predicted future-state information into consideration. To enable proactive coordination through predictive analytics, it is critical to understand how the reliability of future-state information evolves as more information is accumulated for patients over time and design feasible and effective coordination mechanisms accordingly to take optimal actions at the most opportune time. To this end, this dissertation discusses the design of effective service coordination mechanisms that govern prediction updates and operational actions to reduce service delays in healthcare facilities, especially at the interface between the ED and IUs.

We now present a summary of our research along with contributions and discuss future directions of research that may stem from our work.

4.1 Summary

4.1.1 Proactive Coordination between ED and IUs to Reduce Patient Boarding

We explored the proactive management of inpatient unit bed preparation processes for ED patients in an academic urban level 1 trauma center. Unlike the general belief that the prolonged boarding delay is caused by the lack of inpatient beds, there are open yet unprepared beds most of the time when EDs suffer from severe boarding. We incorporated two pieces of future-state information of patients, i.e., remaining ED LoS and target disposition decisions of patients that can trigger proactive inpatient bed requests, into stochastic modeling processes and designed a fork-join queueing network to model the ED-IU network. The proposed methods enabled us to analytically solve the designed queueing network that shows state-dependent transition behaviors with distinct types of patients, beds, and prediction errors. The key assumption is that once a patient enters an ED and starts undergoing triage, tests (laboratory work and imaging), and treatment, there is growing amount of information for the patient within the EHR to allow reliable prediction of ED disposition decision well ahead of the final disposition decision. Instead of assuming that earlier proactive bed request signals will always lead to a greater reduction in waiting time, we investigated the operational impacts of the timing of proactive bed requests and the errors of the signals. If proactive signals are sent earlier, the operational benefits will increase, but the reliability of prediction information might be compromised due to insufficient patient information. While this trade-off relationship cannot be grasped by classification performance measures, our approach can evaluate the expected operational impact of prediction models by characterizing the bed allocation delay reduction behaviors in different operational settings.

In the ideal case where an IU is dedicated to ED patients and disposition decision predictions are perfect, the proposed bed allocation strategy can lead to about 50% reduction in bed allocation delay for ED patients in a realistic setting mimicking the actual operational parameters of the study hospital. In the more realistic case where an IU is shared by ED and non-ED patients and disposition decision predictions are imperfect, as disposition prediction quality improves, ED patients experience more reduction in bed allocation delay. On the contrary, as the prediction quality degrades, reduction in bed allocation delay for ED patients becomes smaller. However, beds that are prepared due to false positive predictions can be occupied by non-ED patients, and bed allocation delay for non-ED patients can be reduced. Moreover, we showed that as bed preparation servers operate with higher utilization, the impact of prediction quality on boarding delay reduction becomes greater. Therefore, improvement in prediction quality can become more important when the proactive bed allocation scheme is applied to bed preparation servers that are busy.

Finally, we demonstrate that one should account for the progressive nature of the ED care process in optimizing the impact of proactive coordination. If relevant information gain is drastic throughout the ED processes, postponing early bed requests for such patients can actually help improve operational efficiency. However, for patients likely to be admitted to certain inpatient units, the disposition decision prediction quality is rather good even at the earliest stages of the ED care process (e.g., even as early as triage). For such patients, it would be best to send early bed requests without wasting time to wait for more information.

4.1.2 Predicting ED Disposition Decision for Early Resource Allocation

To operationalize prediction outcomes, the form of future information should suit operational needs. We identified that while ED disposition decision prediction has received a significant attention by data scientists and practitioners, a simple binary classification compromises the actual utility of prediction results. We framed the ED disposition decision prediction problem as a multiclass hierarchical classification that could have different utility at each hierarchical level. By modeling machine-based prediction models, trade-off relationship between the reliability and timeliness of disposition decision prediction was unraveled. We demonstrated that for different IU groups, the prediction quality evolution behaves in its own unique way according to the gain of relevant information. In particular, there is significant amount of information collected during triage, where patient initial clinical information is gathered. Then, when lab test results return, there is further progress in the prediction performance. We characterized the patterns in which clinical and demographic information of patients turns into operational implications for patients, by analyzing large-scale patient information, focusing on temporal history of data resulting from ED caregiving stages and laboratory/imaging tests.

Among the different class groups, the most granular class group that includes ICU, TU, GPU, OU, and discharge classes, provides the most useful information. At the same time, however, this class group forms the most difficult classification problem due to the imbalanced distribution of true disposition decisions over the five classes. Despite this challenge, the detailed EHR data items contain useful information that can enhance prediction quality. For instance, the information gathered until the return of the first set of lab tests ordered for a patient enhances the prediction performance that the precision levels in all the five classes reach 45%. In particular, the precision levels for both the ICU and GPU classes exceed 50%, which detect more than 42% of patients in the two classes. Even though the utility of the binary classification problem (i.e., admission vs. discharge) is limited, the prediction performance of the binary classification is appealing with the sensitivity and precision levels of the admission class that are greater than 60% and 75%, respectively. While the five-class classification helps facilitate resource coordination at each IU, the outcome of the binary classification can contribute to improved care processes and expedited discharges in the ED.

Moreover, by imposing a probability threshold in making disposition decision predictions, a hospital can choose to send early bed requests only when a certain level of confidence (i.e., probability of admission to a certain IU) is obtained. This strategy will reduce the total number of patients who will subject to proactive coordination but will enhance the effectiveness of coordination for the selected subset of patients. While the strategy will improve the precision levels for all of the minor classes, the sensitivity levels increased only for the ICU and TU classes in our study hospital. This can be attributed to the fact that the information gathered at T4 has distinct features for these two classes than other minor classes.

Finally, our study also presents insights into the relationship between individual features and physicians' disposition decisions through large scale data analysis. The total number of features we incorporated into the prediction models is over a thousand. While prediction models exploit the features to the greatest extent possible for making predictions, each single feature, especially from vital signs and lab test results, provides interesting insights into how clinical values are related to physicians' disposition decisions.

4.2 Future Research

4.2.1 Sequential Decision Making with Prediction Information

Despite a number of research works in clinical and operational prediction in healthcare settings, study on decision making based upon predictive analytics is limited. Sequential decision making is imperative in healthcare operations since services performed by different providers and departments are interdependent, and actions taken by one service provider or department affects the outcome of actions taken by another. Currently, most sequential decision making under uncertainty, such as a partially observable Markov decision process, only assumes probability distributions without consideration for the forms of prediction outcomes. If predictive analytics is to be a useful tool in real-world healthcare systems and not just a collection of advanced techniques, science must progress in developing decision making algorithms that incorporate predictions. Even though sequential decision making based on prediction information should be problem-specific because the forms and types of future-state information to be created depend on specific clinical and operational settings, the types of future-state information can be categorized into two main types: outcome (e.g., relapse of an illness and intensive care unit admission) and timing (e.g., chronic disease onset timing and admission timing). Operationalization of future-state information is not very effective unless forecasting risks, including prediction errors, can be properly considered and handled through decision making processes. Unfortunately, effectiveness of prediction information in actual clinical and operational settings is very rarely studied. Improvement of understanding of relationship between information quality and decision making is highly demanded by a growing number of real world healthcare systems.

4.2.2 Distributed Multi-Agent Coordination Modeling

Coordination mechanisms are an integral topic in distributed multi-agent systems that cooperate based on information exchange. How to define the interactions between patients, clinical/operational service providers, and information is important to effectively and efficiently characterize systems for presenting actionable insights in distributed service environments. Most existing approaches assume that multiple agents (i.e., service providers) are aware of the state of other agents through communication, which implies that the implementation of coordination of each agent's action requires the agent to communicate its current state to other agents at each step. However, considering the complexity and urgency of services in hospitals, despite the existence of HIT systems, it is difficult to fully update the information at the exact time that actions are performed, thus making existing modeling approaches invalid. Therefore, modeling the interactions between the agents by identifying most informative communication between them through empirical data analysis is necessary. This modeling method will in turn provide essential coordination mechanisms while dramatically reducing the degree of communication in decision modeling with only a slight compromise on optimality.

REFERENCES

[1] Abelson R (2013) E.R.'s account for half of hospital admissions, study says, The New York Times, New York, NY. http://www.nytimes.com/2013/05/21/business/half-of-hospital-admissions-from-emergency-rooms.html, Accessed: 2017-06-10.

[2] Abraham J, Reddy MC (2010) Challenges to inter-departmental coordination of patient transfers: A workflow perspective. *International Journal of Medical Informatics* 79(2):112–122.

[3] Ahmed MA, Alkhamis TM (2009) Simulation optimization for an emergency department healthcare unit in Kuwait. *European Journal of Operational Research* 198(3):936–942.

[4] Al-Araidah O, Momani A, Khasawneh M, Momani M (2010) Lead–Time Reduction Utilizing Lean Tools Applied to Healthcare: The Inpatient Pharmacy at a Local Hospital. *Journal for Healthcare Quality* 32(1):59–66.

[5] Amarasingham R, Patzer RE, Huesch M, Nguyen NQ, Xie B (2014) Implementing electronic health care predictive analytics: considerations and challenges. *Health Affairs* 33(7):1148–1154.

[6] Armony M, Israelit S, Mandelbaum A, Marmor YN, Tseytlin Y, Yom-Tov GB (2015) On patient flow in hospitals: A data-based queueing-science perspective. *Stochastic Systems* 5(1):146–194.

[7] Ashour OM, Kremer GE (2016) Dynamic patient grouping and prioritization: a new approach to emergency department flow improvement. *Health Care Management Science* 19(2):192–205.

[8] Asplin BR, Magid DJ, Rhodes KV, Solberg LI, Lurie N, Camargo CA (2003) A conceptual model of emergency department crowding. *Annals of Emergency Medicine* 42(2):173–180.

[9] Augustine JJ (2016a) 2015 emergency department survey shows spike in volume, structural changes, patient boarding concerns. http://www.acepnow.com/article/2015-emergency-department-survey-shows-spike-volume-structural-changes-patient-boarding-concerns/, Accessed: 2017-06-10.

[10] Augustine JJ (2016b) Long emergency department boarding times drive walkaways, revenue losses. http://www.acepnow.com/article/now-boarding/, Accessed: 2017-06-10.

[11] Bahensky JA, Roe J, Bolton R (2005) Lean sigma–will it work for healthcare. *Journal of Healthcare Information Management* 19(1): 39–44.

[12] Bair AE, Song WT, Chen YC, Morris BA (2010) The impact of inpatient boarding on ED efficiency: a discrete-event simulation study. *Journal of Medical Systems* 34(5):919–929.

[13] Ballé M, Régnier A (2007) Lean as a learning system in a hospital ward. *Leadership in Health Services* 20(1):33–41.

[14] Barak-Corren Y, Israelit SH, Reis BY (2017) Progressive prediction of hospitalisation in the emergency department: uncovering hidden patterns to improve patient flow. *Emergency Medicine Journal* 34(5):308-âĂŞ314.

[15] Batt RJ and Terwiesch C (2016) Early task initiation and other load-adaptive mechanisms in the emergency department. *Management Science* 63(11):3531–3551.

[16] Bayley MD, Schwartz JS, Shofer FS, Weiner M, Sites FD, Traber KB, Hollander JE (2005) The financial burden of emergency department congestion and hospital crowding for chest pain patients awaiting admission. *Annals of Emergency Medicine* 45(2):110–117.

[17] Begley CE, Chang Y, Wood RC, Weltge A (2004) Emergency department diversion and trauma mortality: evidence from houston, Texas. *Journal of Trauma* 57(6):1260–1265

[18] Ben-Tovim DI, Bassham JE, Bolch D, Martin MA, Dougherty M, Szwarcbord M (2007)

Lean thinking across a hospital: redesigning care at the Flinders Medical Centre. *Australian Health Review* 31(1):10–15.

[19] Ben-Tovim DI, Bassham JE, Bennett DM, Dougherty ML, Martin MA, O'Neill SJ, Sincock JL, Szwarcbord MG (2008) Redesigning care at the Flinders Medical Centre: clinical process redesign using "lean thinking" *Medical Journal of Australia* 188(6):S27.

[20] Boyle A, Beniuk K, Higginson I, Atkinson P (2012) Emergency department crowding: Time for interventions and policy evaluations. *Emergency Medicine International* 2012.

[21] Boyle J, Jessup M, Crilly J, Green D, Lind J, Wallis M, Miller P, Fitzgerald G (2012) Predicting emergency department admissions. *Emergency Medicine Journal* 29(5):358–365.

[22] Burke PJ (1956) The output of a queuing system. *Operations Research* 4(6):699–704.

[23] Burt CW, McCaig LF, Valverde RH (2006) Analysis of ambulance transports and diversions among US emergency departments. *Annals of Emergency Medicine* 47(4):317–326.

[24] Cameron A, Rodgers K, Ireland A, Jamdar R, McKay GA (2015) A simple tool to predict admission at the time of triage. *Emergency Medicine Journal* 32(3):174–179.

[25] Carboneau C, Benge E, Jaco MT, Robinson M (2010) A lean six sigma team increases hand hygiene compliance and reduces hospital-acquired MRSA infections by 51%. *Journal for Healthcare Quality* 32(4):61–70.

[26] Carter EJ, Pouch SM, Larson EL (2014) The relationship between emergency department crowding and patient outcomes: A systematic review. *Journal of Nursing Scholarship* 46(2):106–115.

[27] Casalino E, Wargon M, Peroziello A, Choquet C, Leroy C, Beaune S, Pereira L, Bernard J, Buzzi JC (2014) Predictive factors for longer length of stay in an emergency department:

a prospective multicentre study evaluating the impact of age, patient's clinical acuity and complexity, and care pathways. *Emergency Medicine Journal*. 31(5):361-368.

[28] CDC (2011) National Hospital Ambulatory Medical Care Survey: 2011 ED Summary Table 27. https://www.cdc.gov/nchs/data/ahcd/nhamcs_emergency/2011_ed_web_tables.pdf, Accessed: 2017-06-10.

[29] Chalfin DB, Trzeciak S, Likourezos A, Baumann BM, Dellinger RP (2007) Impact of delayed transfer of critically ill patients from the emergency department to the intensive care unit. *Critical Care Medicine* 35(6):1477–1483.

[30] Chaou CH, Chen HH, Chang SH, Tang P, Pan SL, Yen AMF, Chiu TF (2017) Predicting length of stay among patients discharged from the ED using an accelerated failure time model. PLoS ONE, 12(1): p.e0165756.

[31] CMS (2018) Hospital Compare datasets. https://data.medicare.gov/data/hospital-compare/, Accessed: 2018-06-11.

[32] CMS (2018) National Health Expenditure Data. Available at: https://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/NationalHealthExpendData/index.html. Accessed: 2018-05-01

[33] CMS and Joint Commission (2017) Specifications Manual for National Hospital Inpatient Quality Measures. https://www.jointcommission.org/specifications_manual_for_national_hospital_inpatient_quality_measures.aspx, Accessed: 2017-08-17.

[34] Considine J, Kropman M, Kelly E, Winter C (2008) Effect of emergency department fast track on emergency department length of stay: a caseâĂŞcontrol study. *Emergency Medicine Journal* 25(12):815–819.

[35] Davis K, Stremikis K, Squires D, Schoen C (2014) Mirror, mirror on the wall. How the

performance of the US Health care system compares internationally. New York: CommonWealth Fund.

[36] Dickson EW, Singh S, Cheung DS, Wyatt CC, Nugent AS (2009) Application of lean manufacturing techniques in the emergency department. *The Journal of Emergency Medicine* 37(2):177–182.

[37] Dinh MM, Russell SB, Bein KJ, Rogers K, Muscatello D, Paoloni R, Hayman J, Chalkley DR, Ivers R (2016) The Sydney triage to admission risk tool (START) to predict emergency department disposition: A derivation and internal validation study using retrospective state-wide data from New South Wales, Australia. *BMC Emergency Medicine* 16(1):46.

[38] Disney RL, Farrell RL, De Morais PR (1973) A characterization of M/G/1 queues with renewal departure processes. *Management Science*, 19(11):1222–1228.

[39] Disney RL, König D (1985) Queueing networks: A survey of their random processes. *SIAM review*, 27(3):335–403.

[40] Eckstein M, Chan LS (2004) The effect of emergency department crowding on paramedic ambulance availability. *Annals of Emergency Medicine* 43(1):100–105.

[41] Eitel DR, Rudkin SE, Malvehy MA, Killeen JP, Pines JM (2010) Improving service quality by understanding emergency department flow: A white paper and position statement prepared for the American Academy of Emergency Medicine. *Journal of Emergency Medicine* 38(1):70–79.

[42] El-Eid GR, Kaddoum R, Tamim H, Hitti EA (2015) Improving hospital discharge time: a successful implementation of six sigma methodology. *Medicine* 94(12):e633.

[43] Fatovich DM, Nagree Y, Sprivulis P (2005) Access block causes emergency department

overcrowding and ambulance diversion in Perth, Western Australia. *Emergency Medicine Journal* 22(5):351–354.

[44] Fleischman RJ, Kaji AH, Diaz VM, McKenzie K, Soltero P, Van Natta TL, Spellberg B (2015) A simple intervention to improve hospital flow from emergency department to inpatient units. *JAMA internal medicine* 175(2):289–290.

[45] Gordon PC, Wanklin JM, Harvey NH, Hatcher GH (1966) An approach to patient care classification. *Canadian Medical Association Journal* 95(24):1228.

[46] Green L (1984) A queueing system with auxiliary servers. *Management sci.*, 30(10):1207–1216.

[47] Green LV, Soares J, Giglio JF, Green RA. Using queueing theory to increase the effectiveness of emergency department provider staffing. *Academic Emergency Medicine* 13(1):61–68.

[48] Han JH, Zhou C, France DJ, Zhong S, Jones I, Storrow AB, Aronsky D (2007) The effect of emergency department expansion on emergency department overcrowding. *Academic Emergency Medicine* 14(4):338–343.

[49] Handly N, Thompson DA, Venkat A (2013) Derivation and validation of a hospital admission prediction model adding coded chief complaint to demographic, emergency department operational and patient acuity data available at emergency department triage using neural net methodology. *Annals of Emergency Medicine* 62(4):S138.

[50] Henry J, Pylypchuk Y, Searcy T, Patel V (2016) Adoption of EHR systems among U.S. non-federal acute care hospitals: 2008-2015. The Office of National Coordination for Health Information Technology (ONC/HIT Data Brief) `https://www.healthit.gov/`

`sites/default/files/briefs/2015_hospital_adoption_db_v17.pdf`, Accessed: 2018-06-10.

[51] Holden RJ (2011) Lean thinking in emergency departments: a critical review. *Annals of Emergency Medicine* 57(3):265–278.

[52] Hoot NR, Aronsky D (2008) Systematic review of emergency department crowding: Causes, effects, and solutions. *Annals of Emergency Medicine* 52(2):126–136.

[53] Hopp WJ, Lovejoy WS (2012) Hospital operations: Principles of high efficiency health care: FT Press.

[54] Jelinek GA, Weiland TJ, Mackinlay C (2010) Supervision and feedback for junior medical staff in Australian emergency departments: Findings from the emergency medicine capacity assessment study. *BMC Medical Education* 10(1):74.

[55] Jones D, Mitchell A (2006) Lean thinking for the NHS. London: NHS confederation.

[56] Kellermann AL, Jones SS (2013) What it will take to achieve the as-yet-unfulfilled promises of health information technology. *Health Affairs* 32(1):63–68.

[57] Kelly AM, Bryant M, Cox L, Jolley D (2007) Improving emergency department efficiency by patient streaming to outcomes-based teams. *Australian Health Review* 31(1):16–21.

[58] Kerber R. (1992) Chimerge: Discretization of numeric attributes. In Proceedings of the tenth national conference on Artificial intelligence; 123–128. Aaai Press.

[59] Kim SW, Li JY, Hakendorf P, Teubner DJ, Ben-Tovim DI, Thompson CH (2014) Predicting admission of patients by their presentation to the emergency department. *Emergency Medicine Australasia* 26(4):361–367.

[60] King DL, Ben-Tovim DI, Bassham J (2006) Redesigning emergency department pa-

tient flows: application of lean thinking to health care. *Emergency Medicine Australasia* 18(4):391–397.

[61] Kocher KE, Meurer WJ, Desmond JS, Nallamothu BK (2012) Effect of testing and treatment on emergency department length of stay using a national database. *Academic Emergency Medicine* 19(5):525–534.

[62] Koning H, Verver JP, Heuvel J, Bisgaard S, Does RJ (2006) Lean six sigma in healthcare. *Journal for Healthcare Quality* 28(2):4–11.

[63] Kulstad EB, Hart KM, Waghchoure S (2010) Occupancy rates and emergency department work index scores correlate with leaving without being seen. *Western Journal of Emergency Medicine* 11(4):324.

[64] Kulstad EB, Kelly KM (2009) Overcrowding is associated with delays in percutaneous coronary intervention for acute myocardial infarction. *International Journal of Emergency Medicine* 2(3): 149–154.

[65] Lee EK, Atallah HY, Wright MD, Post ET, Thomas IV C, Wu DT, Haley Jr LL (2015) Transforming hospital emergency department workflow and patient care. *Interfaces* 45(1):58–82.

[66] Liu S, Hobgood C, Brice JH (2003) Impact of critical bed status on emergency department patient flow and overcrowding. *Academic Emergency Medicine* 10(4):382–385.

[67] Mandahawi N, Al-Shihabi S, Abdallah AA, Alfarah YM (2010) Reducing waiting time at an emergency department using design for Six Sigma and discrete event simulation. *International Journal of Six Sigma and Competitive Advantage* 6(1-2):91–104.

[68] Martin AB, Hartman M, Washington B, Catlin A, National Health Expenditure Accounts

Team (2016) National health spending: faster growth in 2015 as coverage expands and utilization increases. *Health Affairs* 36(1):166–176.

[69] McCarthy ML, Aronsky D, Jones ID, Miner JR, Band RA, Baren JM, Desmond JS, Baumlin KM, Ding R, Shesser R (2008) The emergency department occupancy rate: A simple measure of emergency department crowding? *Annals of Emergency Medicine* 51(1):15–24.

[70] McCarthy ML, Zeger SL, Ding R, Levin SR, Desmond JS, Lee J, Aronsky D (2009) Crowding delays treatment and lengthens emergency department length of stay, even among high-acuity patients. *Annals of emergency medicine* 54(4):492–503.

[71] McConnell KJ, Richards CF, Daya M, Weathers CC, Lowe RA (2006) Ambulance diversion and lost hospital revenues. *Annals of Emergency Medicine* 48(6):702–710.

[72] Moskop JC, Sklar DP, Geiderman JM, Schears RM, Bookman KJ (2009) Emergency department crowding, Part 1: Concept, causes, moral consequences. *Annals of emergency medicine* 53(5): 605–611.

[73] Mumma BE, McCue JY, Li CS, Holmes JF (2014) Effects of emergency department expansion on emergency department patient flow. *Academic Emergency Medicine* 21(5):504–509.

[74] NIH, National Library of Midicine. Available at: https://medlineplus.gov/vitalsigns.html. Accessed:2018-02-01.

[75] Niska R, Bhuiya F, Xu J (2010) National hospital ambulatory medical care survey: 2007 emergency department summary. Natl Health Stat Report; 26(26):1–31.

[76] Oh C, Novotny AM, Carter PL, Ready RK, Campbell DD, Leckie MC (2016) Use of a simulation-based decision support tool to improve emergency department throughput. *Operations Research for Health Care* 9:29–39.

[77] Olshaker JS, Rathlev NK (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.

[78] Peck JS, Benneyan JC, Nightingale DJ, Gaehde SA (2012) Predicting emergency department inpatient admissions to improve same-day patient flow, *Academic Emergency Medicine* 19(9):1045–1054.

[79] Peck JS, Benneyan JC, Nightingale DJ, Gaehde SA (2014) Characterizing the value of predictive analytics in facilitating hospital patient flow. *IIE Transactions on Healthcare Systems Engineering* 4(3):135–143.

[80] Pines JM, Batt RJ, Hilton JA, Terwiesch C (2011) The financial consequences of lost demand and reducing boarding in hospital emergency departments. *Annals of Emergency Medicine* 58(4):331–340.

[81] Pines JM, Hollander JE (2008) Emergency department crowding is associated with poor care for patients with severe pain. *Annals of Emergency Medicine* 51(1):1–5.

[82] Polevoi SK, Quinn JV, Kramer NR (2005) Factors associated with patients who leave without being seen. *Academic Emergency Medicine* 12(3): 232–236.

[83] Powell ES, Khare RK, Venkatesh AK, Van Roo BD, Adams JG, Reinhardt G (2012) The relationship between inpatient discharge timing and emergency department boarding. *The Journal of Emergency Medicine* 42(2):186–196.

[84] Qiu S, Chinnam RB, Murat A, Batarse B, Neemuchwala H, Jordan W (2015) A cost sensitive inpatient bed reservation approach to reduce ED boarding times. *Health Care Management Science* 18(1):67–85.

[85] Rodi SW, Grau MV, Orsini CM (2006) Evaluation of a fast track unit: alignment of re-

sources and demand results in improved satisfaction and decreased length of stay for emergency department patients. *Quality Management in Healthcare* 15(3):163–170.

[86] Rondeau KV, Francescutti LH, Zanardelli JJ (2005) Emergency department overcrowding: the impact of resource scarcity on physician job satisfaction. *Journal of Healthcare Management* 50(5):327–342.

[87] Rowe BH, Channan P, Bullard M, Blitz S, Saunders LD, Rosychuk RJ, Lari H, Craig WR, Holroyd BR (2006) Characteristics of patients who leave emergency departments without being seen. *Academic Emergency Medicine* 13(8):848–852.

[88] Saghafian S, Austin G, Traub SJ (2015) Operations research/management contributions to ED patient flow optimization: Review and research prospects. *IIE Trans. Healthcare Systems Engineering* 5(2):101–123.

[89] Saghafian S, Hopp WJ, Van Oyen MP, Desmond JS, Kronick SL (2012) Patient streaming as a mechanism for improving responsiveness in emergency departments. *Operations Research* 60(5):1080–1097.

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

[91] Shen YC, Hsia RY (2015) Ambulance diversion associated with reduced access to cardiac technology and increased one-year mortality. *Health Affairs* 34(8):1273–1280.

[92] Shi P, Chou MC, Dai JG, Ding D, Sim J (2015) Models and insights for hospital inpatient operations: Time-dependent ED boarding time. *Management Science* 62(1):1–28.

[93] Silla CN, Freitas AA (2011) A survey of hierarchical classification across different application domains. *Data Mining and Knowledge Discovery* 22(1-2):31–72.

[94] Sprivulis PC, Da Silva J, Jacobs IG, Frazer AR, Jelinek GA (2006) The association between hospital overcrowding and mortality among patients admitted via Western Australian emergency departments. *Medical Journal of Australia* 184(5):208–212.

[95] Sullivan C, Staib A, Eley R, Scanlon A, Flores J, Scott I (2015) National Emergency Access Targets metrics of the emergency departmentâĂŞinpatient interface: measures of patient flow and mortality for emergency admissions to hospital. *Australian Health Review* 39(5):533–538.

[96] Sun Y, Heng BH, Tay SY, Seow E (2011) Predicting hospital admissions at emergency department triage using routine administrative data. *Academic Emergency Medicine* 18(8):844–850.

[97] U.S. GAO (2003) Hospital emergency crowded conditions vary among hospitals and communities. U.S. General Accounting Office. `http://www.gao.gov/new.items/d03460.pdf`, Accessed: 2017-06-10.

[98] Van Den Heuvel J, Does RJ, Verver JP (2005) Six Sigma in healthcare: lessons learned from a hospital. *International Journal of Six Sigma and Competitive Advantage* 1(4):380–388.

[99] Wertheimer B, Jacobs RE, Bailey M, Holstein S, Chatfield S, Ohta B, Horrocks A, Hochman K (2014) Discharge before noon: An achievable hospital goal. *Journal of Hospital Medicine* 9(4):210–214.

[100] WHO Global health expenditure database. Available at: http://apps.who.int/nha/database. Accessed: 2018-05-01

[101] Wiler JL, Gentle C, Halfpenny JM, Heins A, Mehrotra A, Mikhail MG, Fite D (2010)

Optimizing emergency department front-end operations. *Annals of Emergency Medicine* 55(2):142–160.

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

[103] Zhou JC, Pan KH, Zhou DY, Zheng SW, Zhu JQ, Xu QP, Wang CL (2012) High hospital occupancy is associated with increased risk for patients boarding in the emergency department. *American Journal of Medicine* 125(4):416.e1–416.e7.

ABSTRACT

PROACTIVE COORDINATION IN HEALTHCARE SERVICE SYSTEMS THROUGH NEAR REAL–TIME ANALYTICS

by

SEUNG YUP LEE

August 2018

Advisors: Drs. Ratna Babu Chinnam & Evrim Dalkiran

Major: Industrial Engineering

Degree: Doctor of Philosophy

The United States (U.S.) healthcare system is the most expensive in the world. To improve the quality and safety of care, health information technology (HIT) is broadly adopted in hospitals. While electronic health record (EHR) systems form a critical data backbone for the facility, we need improved 'work-flow' coordination tools and platforms that can enhance real-time situational awareness and facilitate effective management of resources for enhanced and efficient care. Especially, these IT systems are mostly applied for reactive management of care services and are lacking when they come to improving the real-time "operational intelligence" of service networks that promote efficiency and quality of operations in a proactive manner. In particular, we leverage operations research and predictive analytics techniques to develop proactive coordination mechanisms and decision methods to improve the operational efficiency of bed management service in the network spanning the emergency department (ED) to inpatient units (IUs) in a hospital, a key component of healthcare in most hospitals. The purpose of this study is to deepen our knowledge on proactive coordination empowered by predictive analytics in dynamic healthcare environments populated by clinically heterogeneous patients with individual information changing throughout ED caregiving processes. To enable proactive

coordination for improved resource allocation and patient flow in the ED-IU network, we address two components of modeling/analysis tasks, i.e., the design of coordination mechanisms and the generation of future state information for ED patients.

First, we explore the benefits of early task initiation for the service network spanning the ED and IUs within a hospital. In particular, we investigate the value of proactive inpatient bed request signals from the ED to reduce ED patient boarding. Using data from a major healthcare system, we show that the EDs suffer from severe crowding and boarding not necessarily due to high IU bed occupancy but due to poor coordination of IU bed management activity. The proposed proactive IU bed allocation scheme addresses this coordination requirement without requiring additional staff resources. While the modeling framework is designed based on the inclusion of two analytical requirements, i.e., ED disposition decision prediction and remaining ED length of stay (LoS) estimation, the framework also accounts for imperfect patient disposition predictions and multiple patient sources (besides ED) to IUs. The ED-IU network setting is modeled as a fork-join queueing system. Unlike typical fork-join queue structures that respond identically to a transition, the proposed system exhibits state-dependent transition behaviors as a function of the types of entities being processed in servers. We characterize the state sets and sequences to facilitate analytical tractability. The proposed proactive bed allocation strategy can lead to significant reductions in bed allocation delay for ED patients (up to $\sim 50\%$), while not increasing delays for other IU admission sources. We also demonstrate that benefits of proactive coordination can be attained even in the absence of highly accurate models for predicting ED patient dispositions. The insights from our models should give confidence to hospital managers in embracing proactive coordination and adaptive work flow technologies enabled by modern health IT systems.

Second, we investigate the patterns of decreasing uncertainty in ED patient disposition decisions throughout the course of ED caregiving processes. The classification task of ED disposition decision prediction can be evaluated as a hierarchical classification problem, while dealing with temporal evolution and buildup of clinical information throughout the ED caregiving processes. Four different time stages within the ED course (registration, triage, first lab/imaging orders, and first lab/imaging results) are identified as the main milestone care stages. The study took place at an academic urban level 1 trauma center with an annual census of 100,000. Data for the modeling was extracted from all ED visits between May 2014 and April 2016. Both a hierarchical disposition class structure and a progressive prediction modeling approach are introduced and combined to fully facilitate the operationalization of prediction results. Multinomial logistic regression models are built for carrying out the predictions under three different classification group structures: (1) discharge vs. admission, (2) discharge vs. observation unit vs. inpatient unit, and (3) discharge vs. observation unit vs. general practice unit vs. telemetry unit vs. intensive care unit. We characterize how the accumulation of clinical information for ED patients throughout the ED caregiving processes can help improve prediction results for the three-different class groups. Each class group can enable and contribute to unique proactive coordination strategies according to the obtained future state information and prediction quality, to enhance the quality of care and operational efficiency around the ED. We also reveal that for different disposition classes, the prediction quality evolution behaves in its own unique way according to the gain of relevant information. Therefore, prediction and resource allocation and task assignment strategies can be tailored to suit the unique behavior of the progressive information accumulation for the different classes of patients as a function of their destination beyond the ED.

AUTOBIOGRAPHICAL STATEMENT

Seung Yup Lee received his B.S. degree in Materials Science and Engineering from Sungkyunkwan University, Seoul, Korea, in 2008. He worked in Semiconductor Research & Development Center at Samsung Electronics as an assistant researcher, where he investigated the integration of complex manufacturing processes through production monitoring and communication. He received his M.S. degree in Industrial Engineering from Wayne State University, Detroit, Michigan, in 2014. During his master's work, he focused on designing resource allocation strategies for enhancing situational awareness in smart grid systems through real-time monitoring and communication. During his doctoral study in Industrial Engineering at Wayne State University, he has gained a broad range of interests spanning from stochastic processes, data-driven decision making, machine learning, and artificial intelligence, for designing real-time work management and coordination strategies and mechanisms for promoting operational intelligence in healthcare operations management.