Stochastic Models of Patient Access Management in Healthcare

by

Derya Kilinc

A Dissertation Presented in Partial Fulfillment
of the Requirement for the Degree
Doctor of Philosophy

Approved July 2019 by the
Graduate Supervisory Committee:

Esma Gel, Chair
Kalyan Pasupathy
Jorge Sefair
Mustafa Sir
Hao Yan

ARIZONA STATE UNIVERSITY

August 2019

ABSTRACT

This dissertation addresses access management problems that occur in both emergency and outpatient clinics with the objective of allocating the available resources to improve performance measures by considering the trade-offs. Two main settings are considered for estimating patient willingness-to-wait (WtW) behavior for outpatient appointments with statistical analyses of data: allocation of the limited booking horizon to patients of different priorities by using time windows in an outpatient setting considering patient behavior, and allocation of hospital beds to admitted Emergency Department (ED) patients. For each chapter, a different approach based on the problem context is developed and the performance is analyzed by implementing analytical and simulation models. Real hospital data is used in the analyses to provide evidence that the methodologies introduced are beneficial in addressing real life problems, and real improvements can be achievable by using the policies that are suggested.

This dissertation starts with studying an outpatient clinic context to develop an effective resource allocation mechanism that can improve patient access to clinic appointments. I first start with identifying patient behavior in terms of willingness-to-wait to an outpatient appointment. Two statistical models are developed to estimate patient WtW distribution by using data on booked appointments and appointment requests. Several analyses are conducted on simulated data to observe effectiveness and accuracy of the estimations. Then, this dissertation introduces a time windows based policy that utilizes patient behavior to improve access by using appointment delay as a lever. The policy improves patient access by allocating the available capacity to the patients from different priorities by dividing the booking horizon into time intervals that can be used by each priority group which strategically delay lower priority patients. Finally, the patient routing between ED and inpatient units to improve the patient access to hospital beds is studied. The strategy that captures the

trade-off between patient safety and quality of care is characterized as a threshold type. Through the simulation experiments developed by real data collected from a hospital, the achievable improvement of implementing such a strategy that considers the safety-quality of care trade-off is illustrated.

To Kerime Atilla, for being my first teacher.

ACKNOWLEDGMENTS

Last six years have been quite an interesting and challenging journey for me. Throughout my Ph.D. studies I have received tremendous support and guidance from many people that I feel grateful to have them in my life.

I would like to give a special thanks to my advisor, Dr. Esma Gel. She has been a great mentor to me both academically and personally. Her guidance, expertise, vision, and encouragement taught me how to conduct an academic research and helped me greatly to improve myself as a researcher. Her sense of ethics and unique way of conducting research set me a great example. I would like to thank Dr. Mustafa Sir and Dr. Kalyan Pasupathy for serving in my committee as well as being incredible mentors for me. Their help and support allowed me to direct my career and make my research meaningful in real life. I enjoyed every moment working with them. I would also like to thank Dr. Jorge Sefair and Dr. Hao Yan for taking part in my dissertation committee and providing comments that improved quality of my research.

Lastly, I want to thank my roommate Petek Yontay for helping me greatly to have a nice life in Arizona, and being my ride or die in all of the new experiences. I would like to personally thank my friends Yigit Yildirim, Mert Ozer, Sultan Kilinc, and Berkin Arici for their friendship and being there for me whenever I need their company and help. I would like to also thank my lab-mate Aysegul Demirtas who understood my research struggles whenever I shared them with her.

I would like to express my gratitude to my parents, Sevil and Aydin Kilinc, my sister Deniz Kilicoglu ,and my brother-in-law Baris Kilicoglu. This degree could not have been completed without their support and unconditional love that I can feel even across the ocean.

TABLE OF CONTENTS

CHAPTER Page

CHAPTER Page

LIST OF TABLES

Table Page

LIST OF FIGURES

Figure Page

Figure Page

Chapter 1

INTRODUCTION

Interest in healthcare resources has grown significantly throughout the years globally. Due to this growing demand, the hospitals are facing the problem of prolonged waiting times experienced by the patients while seeking access to care. Timeliness is considered as one of the key indicators of the quality of healthcare delivery as well as patient safety. Many studies in the literature emphasize lengthy waiting times can have a negative effect on patients condition and put patients' safety at risk. Considering this vitality of timeliness of care on healthcare outcomes, it is essential for the hospital administrators to manage the available capacity effectively to satisfy patients' needs and obtain better healthcare outcomes without incurring additional costs.

Hospitals receive patient demand from various resources which can be divided into three main encounter types as emergency, inpatient, and outpatient care. The patients who require access to care are different in terms of level of care needed, urgency and stochasticity of the need, service time expectations, and resource usage. Additionally, the wait experienced by each one of these patient types differs since emergency patients physically experience the delay while they are actually in a healthcare facility while outpatients usually request an appointment on a future date. Therefore, in emergency cases, we can observe a queue accumulated in a waiting area, while in outpatient case, patients experience *indirect* waiting, mostly in days, in a virtual queue.

The differences in healthcare experience of each one of the encounter types call for unique approaches in improving healthcare delivery in terms of effectiveness, ef-

ficiency, timeliness, and safety which are directly related to prolonged waiting times experienced by patients. While there is a consensus on importance of reducing the waiting times in patient service satisfaction and safety, there are no standardized guidelines to reduce those waiting times under different settings. Inefficiencies and prolonged waiting times experienced by patients beg for decision support tools that can help to improve utilization of resources and patient outcomes. We utilize operations research, statistical estimation, and stochastic control methods to provide strategies to satisfy the goals of healthcare delivery listed above.

In this dissertation, we introduce the concept of access management, which is a new concept in healthcare systems. Our main goal in access management is developing policies that consider differences in patients' needs and differences in service level expectations rather than providing equal access to each patient. This approach can be summarized as designing an access protocol that assigns "the right capacity to the right patients with the right access delay." Access is a different concept from the scheduling where in the scheduling problems the concept includes the sequencing of the job while in the access problem, we go beyond the scheduling and focus on high-level rules that allocates capacity based on the patient characteristics and priorities.

Based on our collaborations with a healthcare institution, we identify the challenges within the system based on the data available and we focus on the allocation of healthcare resources in two different settings considering the setting specific trade-offs and objectives.

In the first setting, we examine an outpatient access problem where patients request an appointment for a future date. Unlike ED, in outpatient systems patients experience a virtual (indirect) waiting. In this setting, one of the critical components of the system is patient behavior which is not relevant in our first setting. In outpatient care, determining "right" patients are directly associated with identifying

patient priorities. While there is no general tool to determine patient priorities, we consider a setting where the priorities can be determined by institution based rules. After determining who the "right" patients are the next step of access management is allocation of the capacity with "right" delay which can be achieved through prioritization. While prioritization is the key to improve access, it is not the only criteria that needs to be considered since determining "right" delay also depends on patients' waiting time expectations and their sensitivity to experienced delay.

We then consider an Emergency Department (ED) in the second setting where patients experience *direct* waiting times while accessing the hospital beds in inpatient wards (IW). In this setting, allocating the "right" capacity to the "right" patient is directly associated with appropriateness of patient's condition to the IW that the bed belongs to and the "right" access delay can be determined based on patient's urgency.

Considering these differences in the settings and the concept of access in them, we develop two different approaches in this dissertation to respond setting specific goals. In the first setting, we focus on prioritizing patients while allocating the clinical capacity where patients exhibit reaction to access delay. Therefore, we analyze the second setting in two steps. First, in Chapter 2, we focus on identifying patients' waiting time expectations. Then, in Chapter 3, we develop an access protocol to prioritize patients considering the behavior studied in Chapter 2. In our work on the second setting which is presented in Chapter 4, we specifically focus on appropriateness and timeliness of the assignment to improve access for patients are admitted to different IWs through ED.

1.1 Contributions of this Dissertation

The first contribution is developing statistical models to understand the patient willingness-to-wait (WtW) behavior and comprehensive analysis of this behavior on

simulated and real life data. Our observations from real life systems suggest that offered appointment delay has an effect on patients' appointment booking and fulfilling behavior and patients show aversion to prolonged appointment delays. This observation suggests that it is crucial to understand WtW behavior before designing an effective access policy and consider this behavior in making access decisions. In order to develop a policy that considers patient behavior, one needs to fully characterize WtW from available data. To this end, we develop two novel statistical methods to obtain patient willingness-to-wait from available data on booked appointments. Through an extensive analysis on simulated datasets, we show that the statistical methods that we develop are effective in parameterizing WtW as a function of the offered appointment delay.

We then focus on improving patient access by developing an access protocol that utilizes WtW behavior and prioritize by using offered appointment delay as a lever. We propose a framework that schedules patients from different priorities on certain time intervals of the booking calendar which are called time windows. In this part, we focus on managing the available capacity in terms of the calendar days that can be used by each priority group and by strategically delaying the patients from lower priority groups. Our objective is to segmentize patients from different priority groups considering patient WtW and prioritize patients by not only serving higher priorities earlier in the booking horizon but also serving a higher proportion of the arriving demand from higher priorities.

The final contribution in this dissertation is developing an assignment policy that utilizes overflows to improve the access of ED to IW where the patients that require beds from different IWs experience prolonged waiting (boarding) times before they are admitted to an IW while occupying an ED bed. We employ a Markov decision process (MDP) and model the patient flow as a multi-class queueing network problem with

flexible servers where the servers are inpatient beds with the objective of minimizing the total cost consists of the cost associated with the risk of adverse events can be developed while waiting in ED and the cost of assigning patient to a secondary unit where patients can get less than ideal treatment. By analyzing the structural properties of our MDP, we identify the optimal policy as a state-dependent threshold-type policy where keeping patients in ED for a primary assignment is beneficial up until a certain number of outstanding ED bed requests. We then develop a heuristic policy which is effective and easy to use where we dynamically balance the cost associated with patient safety and quality of care. Finally, we use a simulation model which is calibrated with the real-life data to assess the performance of our proposed patient routing policy.

1.2 Dissertation Organization

The rest of this proposal is organized as follows. In Chapter 2, we focus on understanding patient willingness-to-wait to outpatient appointment by conducting empirical analyses on available appointment data. We then present our time-window based framework to improve patient access by considering patient behavior in Chapter 3. Chapter 4 focuses on our research in improving patient flow between ED to inpatient units considering patient safety and quality of care. Finally, we present our concluding remarks with research plan in Chapter 5.

Chapter 2

STATISTICAL CHARACTERIZATION OF PATIENT RESPONSE TO ACCESS DELAY USING HEALTHCARE TRANSACTIONAL DATA

2.1 Introduction

Lengthy waiting times for medical care is a growing problem that the US healthcare system faces. Appointment scheduling literature defines two main types of waiting, where *direct wait* refers to the amount of time patient waits at the care facility on the day of the appointment, whereas *indirect wait* refers to the number of days (typically) between the appointment request and the actual appointment. While direct wait is an important metric since it impacts patients' perception of quality of care, excessive indirect wait may have a larger impact on the patients' health outcomes, and may even put patients' safety and positive health outcomes in jeopardy (Murray and Berwick, 2003). Additionally, long indirect waiting times are often associated with higher cancellation and no-show rates, reducing clinic efficiency and increasing healthcare costs (LaGanga and Lawrence, 2007). Unfortunately, indirect waiting has been an growing problem in many healthcare settings. A recent survey conducted in 15 metropolitan areas shows that the average wait time for a new appointment has increased by 30% to 24.1 days since 2014 (Merrit-Hawkins, 2017). The same survey also indicates that the average new patient wait time for a physician appointment can be as high as 52.4 days in some major metropolitan areas such as Boston. Considering the growing healthcare needs of an aging population, increasing rates of chronic diseases, and significant expansion of health insurance coverage, it is essential for healthcare systems to use available capacity effectively to provide timely access

to healthcare services.

The ability to use the available care capacity effectively depends strongly on accurate understanding of patient needs, expectations and behavior. Ideally, access to healthcare resources should be provided in such a way that "the Right patient sees the Right care provider, at the Right time," which we refer to as *3R Healthcare Access*, or 3R-HA for short. Unfortunately, appointments are currently provided to patients in first-come-first-served manner by an appointment office (AO) agent, who typically lacks the ability and information to assess the potential negative consequences of offering a particular slot with a long appointment delay to a patient. For example, if the patient's needs are indeed more urgent than the offered appointment delay, the patient may "leave," (i.e., hang-up the phone) without booking an appointment (incidence referred to as PLWBA hereon), or book the offered appointment but cancel, reschedule or "no-show" (incidence referred to as C/RS/NS hereon), subsequent to booking the appointment. In the worst case, patient waits for the offered appointment and then suffers negative health outcomes due to delayed treatment and interventions. All of these undesirable outcomes mainly result from a mismatch between the patient's needs/expectations and the offered appointment delay, and the operational and financial impact of these incidences can be significant, especially in severely capacity-constrained specialty clinics. For example, PLWBA means that the patient has to seek care elsewhere (which may not be in his best interest), and the clinic foregoes an opportunity to provide care. The term PLWBA is inspired from left without being seen (LWBS) term which is a well-known concept in emergency departments (see, e.g., Baker *et al.*, 1991; Batt and Terwiesch, 2015; Lucas *et al.*, 2014). On the other hand, C/RS/NS may possibly mean that an appointment slot which could have been used to serve a patient in need of care goes unused.

While the operational and financial impacts of PLWBA and C/RS/NS have been

assessed by several studies (see, e.g., Rust *et al.*, 1995; Murray and Berwick, 2003; LaGanga and Lawrence, 2007; Dreiher *et al.*, 2008; Defife *et al.*, 2010; Zacharias and Pinedo, 2014), attempts to understand and quantitatively characterize patient behavior in terms of how they respond to offered appointment delay have been relatively scarce, even though a clear understanding of patient response may help us avoid PLWBA and C/RS/NS incidences and provide 3R-HA. Note that we are pointing to a gap in the understanding of how patients respond to "offered" appointment delays rather than a characterization of events that result in abandonment of "already booked" appointments since we are interested in the phase at which slots are offered to patients by appointment office agents.

A fundamental reason for the scarcity of work in this area is the fact that data that one can use to characterize patient response have been difficult to obtain, since healthcare systems, while they have very comprehensive IT infrastructure to collect and maintain data on appointments that have taken place for billing and insurance purposes, have not paid real attention to collecting other transactional data that are useful to characterize patient response as a function of appointment delay. For example, most systems have data on appointments that were booked by patients (even though they may subsequently result in C/RS/NS), but do not keep information on PLWBA incidences. As we will discuss below, this creates a missing data problem, where "lost" appointment requests are not observable (similar to lost sales in retail systems).

A highly useful characterization of patient response to indirect wait is the probability that an offered appointment with a k-day delay will be booked and ultimately "fulfilled" by the patient, since this characterization can then be used in several different ways to improve the degree to which the offered appointment delay matches the needs of a given patient, or make scheduling or resource allocation decisions. We

refer to this probability as the *realization probability* in this chapter. Overbooking, for example, means scheduling multiple patients in one slot, and is similar to overselling of seats on a commercial flight. It is a generally "unpublicized but relatively ubiquitous" practice to provide care to patients that are deemed-urgent by staff and to combat the loss of use of care capacity due to NS situations. Information on the fulfillment probability, for example, can be used to decide which of the already-booked appointment slots can be more reliably used for overbooking. We note that this characterization can be done either for an arbitrary patient, or for a patient from a particular subpopulation, e.g., patients that need surgical intervention among patients with low back pain. In this example, it is easy to appreciate that these patients will be more "impatient" than other groups of patients that need other, more conservative interventions.

In this chapter, our objective is develop methodologies to characterize realization probability, for a given patient subpopulation, as a function of appointment delay using existing patient transactional data. To do this, we take an approach that is inspired by the willingness-to-pay or reservation price (i.e., maximum price that a customer will pay for a specific product) concepts from the economics literature, and assume that each patient in a given population of interest has an inherent "willingness-to-wait" (WtW for short hereon) for an appointment of the same type. A characterization of the distribution of WtW for a patient population allows us to calculate the probability that a randomly chosen patient from this population will fulfill an appointment with k-days' delay (we use business days as the time unit for appointment delays) since the appointment will only be booked and subsequently fulfilled if and only if the patient's WtW is **at least** k-days. Then, from this argument, it is trivial to see that the appointment realization probability (denoted by p_k hereforth) is equal to the probability that WtW of an arbitrary patient in this population

is greater than or equal to k, i.e., $p_k = P(\text{WtW} \geq k)$. Unfortunately, direct data on WtW of patients is almost never available, and even if patients were asked questions on how long they would be willing to wait, it would probably be a moot exercise, since patients would most likely indicate a low WtW (e.g., two days) to game the system for quick access. Hence, it is necessary to develop statistical methods to make inferences on the WtW distribution using the available historical transactional data on appointment requests, bookings and fulfillment.

We build two non-parametric models to characterize WtW and then use it to estimate p_k , $k \in \{0, 1, \ldots, T\}$ where T denotes maximum possible access delay for an offered appointment. The performance of the models are assessed using the errors between the estimated p_k and what is "observed" either from real-life data or "assumed" our simulated datasets, which realistically generate with relevant variates such as appointment requests, offered appointment delays, booked or PLWBA, and final status of the appointment by assuming a WtW distribution.

The first method we present involves non-parametric **survival analysis** of the "lifetime of an offered appointment." Note that if the patient does not book an appointment and leaves the system (i.e., PLWBA) the lifetime of the offered appointment is zero. On the other hand, if the patient books the appointment and C/RS at a time τ days after the current time (but before the offered appointment date), then the lifetime of the appointment is τ days. For the case of no-show (i.e., NS), the lifetime can be considered to be equal to the delay of the booked appointment. Finally, if the appointment is fulfilled, then the lifetime of the appointment is at least as large as the delay of the booked appointment. Note that this last type of observations are right-censored since we do not get to observe an event that marks the end of the appointment life. The literature on non-parametric estimation from censored data is well established (see, e.g., Kaplan and Meier, 1958; Turnbull, 1976; Gentle-

man and Geyer, 1994; Anderson-Bergman, 2017a). Hence, survival analysis can be directly used to characterize the lifetime of an appointment with delay k, except that the exact timing of C/RS events are almost never recorded, in addition to the data being right-censored.

The second estimation method uses a **rank-based choice model** commonly used in the retail and revenue management literature to estimate customer preferences for a set of products assuming that customers have a rank-based preference list. We use the non-parametric maximum likelihood estimator, which was introduced in Farias *et al.* (2013) and van Ryzin and Vulcano (2014). Osadchiy and Kc (2017) was the first study to use a rank-based choice model to analyze patient's reactions to appointment delay in a healthcare setting; their focus was on estimating the probability of no-shows and late cancellations. There are, also, other differences in our assumptions on the nature of patient response and decisions; for example, they consider appointment booking and fulfilling decisions separately. In our study, we explore the assumption that patients have a pretty concrete idea of whether or not they will be fulfilling an appointment at the time of appointment request, guided by their inherent WtW. Additionally, we consider all possible ways that a booked appointment may not be fulfilled, such as cancellation, rescheduling, and no-shows while Osadchiy and Kc (2017) focuses only on late cancellations and no-shows. To develop the rank-based choice model, we follow a similar rationale as we did for survival model in using the transactional appointment data. We assume that each patient has a "preference list" of acceptable appointment delays, bounded by his WtW. As stated before, we assume that the patient is always offered the earliest appointment available. Hence, the patient only needs to evaluate whether the offered appointment delay is in the preference list or not. For instance, if the patient has a WtW of w days, his preference list includes all waiting times from 0 to w, in integer increments and a lower waiting

time always preferred to a longer waiting time. If a patient is observed to have fulfilled the offered appointment, we can infer that the offered delay is in his preference list.

Our extensive experimentation and rigorous statistical testing on simulated data show that the proposed methods are effective in estimating WtW. We further scrutinize the performance of the proposed methods by comparing them with a baseline model that directly estimates p_k from data, by calculating the fraction of realized appointments among all offered appointments with delay of k days. We observe that estimating p_k through the use of predicted WtW distributions by either the survival model and the rank-based choice model outperforms the baseline method in terms of various error metrics and robustness with respect to the availability of data.

An important contribution of the chapter is to provide a comprehensive methodology to use patient transactional data, including imputation techniques to complement the available data. Our testing demonstrates that imputation significantly improves the estimation performance, and furthermore, the estimates obtained with imputed datasets look almost as good as the ones that would be obtained by much more comprehensive datasets that include all details on all patient "encounters." The study also makes a case for collecting data over time, and more information on different aspects of patient response (such as details on PLWBA), since we observe that the estimates also improve as the number of data points collected increases. Finally, we use the proposed methods on real patient data from a highly specialized clinic that offers destination medicine to obtain insights into the differences in sensitivity to delay of different patient populations (e.g., new and established patients) and make suggestions on possible ways of employing the estimated probabilities for developing policies to improve patient access.

The rest of the chapter is organized as follows. Section 2.2 describes the available data used to build the models. Section 2.3 and Section 2.4 present the two proposed

methods for modeling patient WtW behavior. In Section 2.5, we describe a typical patient flow process, which we use to obtain simulated data and conduct numerical experiments. We report detailed analyses of the performance of the models in estimating the parameters of the underlying WtW distribution. In Section 2.6, we present our results on real patient data. We give concluding remarks in Section 2.7.

2.2 Problem Description and Available Data

We assume that at the time of his appointment request, the patient (to whom we will referring with the "he" pronoun hereon) knows his WtW; for example, he knows that he would not be willing to wait more than two weeks for this appointment. Hence, if the appointment delay for the appointment offered by the agent is less than two weeks, the patient will book this appointment but may later C/RS/NS due to reasons unrelated to the length of indirect wait, such as insurance coverage or an unexpected improvement in medical condition. If the offered appointment delay is more than two weeks, we assume that the patient knows this is longer than the time that he is prepared to wait, but with a certain probability, he may still decide to book the appointment to "keep his options open," and have a booked appointment while he seeks care elsewhere (i.e., misclassification at the time of appointment). We assume that this patient will, subsequently, "abandon" this appointment through C/RS/NS (not particularly important which one).

We are well aware of the fact that there are numerous other situations we are unable to consider with this basic model, but we have found that it sufficiently models patient behavior at a population level. Figure 2.1 provides an outcome tree describing all of the above defined possible outcomes of an appointment request by a patient. In addition, the red check and cross marks in the figure refers to the outcomes for which data are typically recorded and maintained by healthcare systems. In general,

several different pieces of information on booked appointment instances are available, but we do not have (at least at the time of writing of this chapter) access to detailed data on appointment requests resulting in PLWBA (although some partial data, as discussed below, is available on those instances).

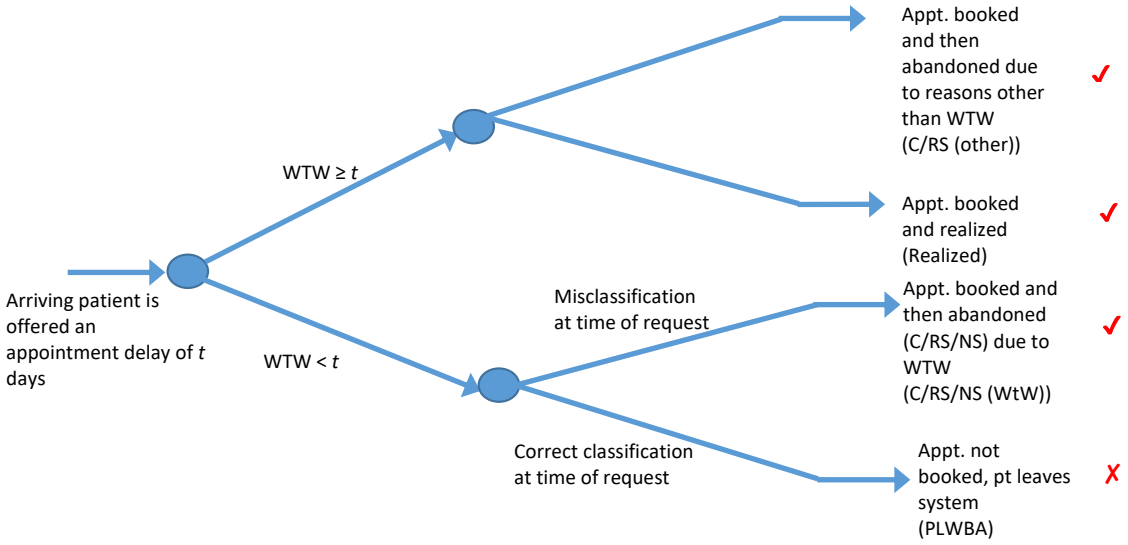


Figure 2.1: All Possible Outcomes of an Appointment Request.

Since the availability of data on appointment request and scheduling transactions may change in different contexts, we define different "data settings." For example, in some healthcare systems, patient requests for an appointment may be recorded and maintained along with information on various traits of the patient, as well as the offered appointment delay, etc. even when the patient decides not to book an appointment and leaves (i.e., PLWBA). We refer to such a set of data as the **Full Encounter Dataset** (FED for short) since an instance in that dataset represents a patient request for an appointment, and hence includes all of the "encounters" that AO agents have with patients, along with the relevant data on the encounter such as delay(s) of the offered appointment(s), final status of the request (i.e., the patient booked an appointment or left without booking), and the final status of the booked

appointments (i.e., fulfilled, cancelled, rescheduled, no-show).

The available data at the institution that we conducted this study did not include information on appointment requests that did not result in a booked appointment. We only had access to information on encounters that resulted in booked appointments. We refer to such datasets as the **Booked Appointment Dataset** (BAD for short), which is a subset of FED. The BAD we had access to included information on the usual suspects like the delay of the booked appointment, but did not, for example, include any information on whether there were other appointment slots with possibly different delays offered to the patient that the patient did not accept. A depiction of FED, as well as BAD as a subset of FED are given in Tables 2.1 and 2.2. In this example, the appointment records 1001 thru 1006 show all of the appointment requests received in this FED. The BAD that corresponds to this FED is given in Table 2.2; only records 1001, 1003, 1004 and 1006 resulted in a booked appointment and were recorded as part of BAD.

Table 2.1: A Depiction of Data Contained in FED, for Each Appointment Request Record

Record ID	Appt ID	Appt Delay	Patient Type	Status
1001	123	3	New	Seen
1002	-	5	Established	Not Booked
1003	124	2	Established	Cancel
1004	125	6	New	No-show
1005	-	30	New	Not Booked
1006	126	10	New	Seen

In healthcare settings there is usually a set booking horizon (say, 6 months) that

Table 2.2: A Depiction of Data Contained in BAD, for Each Booked Appointment

Record ID	Appt ID	Appt Delay	Patient Type	Status
1001	123	3	New	Seen
1003	124	2	Established	Cancel
1004	125	6	New	No-show
1006	126	10	New	Seen

the appointment scheduling agents have access to book appointments. We use T to denote this booking horizon length, which is also the highest delay observable from the transactional appointment data. Note that, although not likely, the maximum possible WtW for the patient population may be higher than T. Therefore, we denote the maximum WtW with Z and we set it to a large value to allow for this behavior.

In our study, we were able to alleviate the unavailability of data on PLWBA instances by using a separate source of data, called the **Demand Universe Dataset** (DUD for short), which includes information about calls patients make to the AO to request an appointment and the final status of these requests (i.e., booked an appointment or left without booking). The DUD we had access to also included a free-text column containing the reason for PLWBA outcome, for example, "patient refused the offered appointment and is no longer interested in booking an appointment." A representation of DUD is available in Table 2.3. Using DUD, we determined that approximately 6.4% of the patients who called to request an appointment during the study period left the system without booking. This rate represents the probability of PLWBA given that the patient is not denied due to any other reasons. Unfortunately, DUD does not include further details on the patient, or the offered appointment delays that were offered to but not accepted by the patient.

Table 2.3: A Depiction of Data Contained in DUD

Record ID	Final Status	Reason
1001	Booked	-
1002	Not Booked	Patient rejected due to delay
1003	Booked	-
1004	Booked	-
1005	Not Booked	Denied by the department
1006	Booked	-

A simplifying assumption worth noting is that we assume that patients will prefer an earlier appointment to a later appointment, even though this assumption may not always hold due to the patient's personal scheduling conflicts. Such conflicts are almost never properly reflected in the data so there was no way that we could accurately consider such effects. Furthermore, this simplifying assumption allowed us to identify an interval within which the patient's WtW lies. For example, it allowed us to infer that the patient's WtW should be at least k days if this particular patient fulfilled an appointment with k days' delay. Similarly, a C/RS/NS record indicates a WtW that is at most k days.

Another assumption we made is that patients are provided appointments in the order of their call; that is, a patient calling before another one will be offered an appointment with lower appointment delay. Due to queuing, multiple agents serving patients on the phone, and other effects due to patient's scheduling conflicts, we understand this may not be the case, but again, this assumption allowed us to develop imputation methods to append BAD with probable PLWBA instances and their respective offered delays. Our method obtains probable PLWBA instances by randomly

sampling instances in BAD and inserting "likely" PLWBA instances. The resulting dataset is one that includes imputed PLWBA instances that would be statistically equivalent to those that would be found in a FED. This imputed dataset, which we call this an Imputed BAD (I-BAD), actually comes quite close to having FED in terms of the performance of the estimation procedures.

The idea behind the imputation via sampling can be considered as following. Consider a point in time that a patient arrives and an appointment with certain delay is offered to the patient during the call. The patient evaluates that wait based on his WtW and leaves without booking an appointment. Following this PLWBA, the same appointment will most likely be offered to the next arriving patient that requests an appointment since the slot is still not utilized. If the new patient's WtW is more than the offered delay, the current patient books the appointment and the appointment slot that is offered more than once will appear in BAD only once. While we consider PLWBA as a behavior that is similar to C/RS/NS, we do not necessarily sample from the appointments that result in C/RS/NS to avoid introducing bias into our analyses. In Section 2.3 and Section 2.4, we further discuss the details of the imputation methods for each of our proposed models separately.

Our real-life dataset is a BAD, and consists of almost three years of information on all of the booked appointments for multiple outpatient clinics in a specialty unit. A nice feature of our dataset is that it includes data on two different patient types: (i) new patients, who are calling the clinic for the first time to request an initial appointment, and (ii) established patients, who have been previously seen by a provider in the clinic. In our study, we analyze the WtW behavior of these two patient types separately to explore the ability of our methods to capture the different reactions of new and existing patients to different levels of offered appointment delays.

Unlike typical appointment scheduling data that only include timestamp data

with basic appointment type information, our dataset captures the reason of cancellations (i.e., C) and rescheduled appointments (i.e., RS). Using this information, we can clearly distinguish whether an appointment is canceled or rescheduled due to WtW related reasons or not. The reason for no shows (NS) cannot be obtained from the patient, therefore, we assume that all NS occurs due to WtW. Note that other approaches are possible to incorporate to our methodology as well; we have found this assumption to work sufficiently well for our estimation problem. If a patient's WtW is higher than the offered appointment delay, the patient would book the appointment but later may cancel or reschedule it due to reasons unrelated to waiting time (denoted as C/RS (other) below) such as insurance coverage or unexpected improvement in medical condition. On the other hand, if the patient's WtW is less than the offered delay, the patient, even though he/she is expected to not book the offered appointment, may still book the appointment, but would abandon the appointment subsequently through cancellation, rescheduling to an earlier appointment, which can be identified since the reasons of C/RS are recorded or not showing up at the time of the appointment (denoted as C/RS/NS (WtW) below). Accordingly, all of the three top outcomes (marked with a check mark in Figure 2.1), can be observed in our dataset, which covers all of the booked appointments during the study period. The unobserved instances of patient appointment requests are due to patients leaving the system without making an appointment.

Table 2.4 lists descriptive statistics on appointment delays for new and established patients using our real-life data. We see that the new patients consistently show higher average delays for all three appointment status types (significant at $p = 0.005$), which may be due to various ways that established patients are "prioritized." For example, in many cases nurses and physician assistants are able to "work in" (using overbooking) established patients who want to be seen due to a patient-reported

urgency. Additionally, we observe that fulfilled appointments have a significantly lower average appointment delay (significant at $p = 0.005$).

Table 2.4: Descriptive Statistics from Our Real-life Dataset

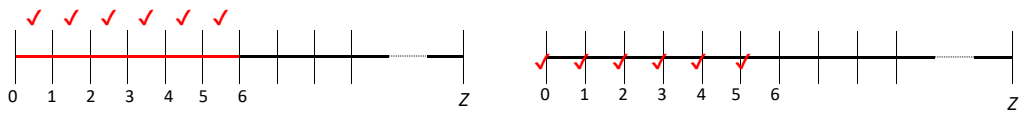
	Established Patients			New Patients		
Final Status	n	Avg. Delay (days)	Std. Delay (days)	n	Avg. Delay (days)	Std. Delay (days)
Fulfilled	5350	21.5	20.4	5757	30.0	22.8
C/RS/NS (WtW)	2441	26.9	20.2	2851	36.1	22.6
C/RS (other)	639	23.5	19.4	586	34.0	22.6

We propose two non-parametric models, which we refer to as **survival model** and **rank-based choice model**, to empirically characterize patient WtW. We consider non-parametric models since our goal is to develop good models that are generalizable to a broad range of situations and at different clinical settings. The main difference between two models is the way they treat patient WtW. Survival analysis model treats WtW as a continuous variable, and therefore, the model generates and estimate of the probability that patient WtW belongs to a certain interval.

In particular, the survival model starts by empirically generating the set of intervals that patient WtW can belong to and provides the probability that an arbitrary patient's WtW is contained in the interval $[s_{j-1}, s_j) \subseteq \mathcal{T}$, where $\mathcal{T} = \{[s_0, s_1), [s_1, s_2), \ldots, [s_y, s_{y+1})\}$ and s_j where $j \in \{1, \ldots, y+1\}$ are determined empirically, $s_0 = 0$ and $s_{y+1} = \infty$. We discuss how these intervals are determined in Section 2.1 in more details. In comparison, the rank-based choice model estimates the probability that a given patient belongs to "patient group k," which is defined as the group of

patients with WtW equal to k days, where $k \in \mathcal{D} = \{0, 1, \ldots, Z\}$. In rank-based model, we use Z as the maximum value that a patient's WtW can take and set it to a large number to avoid having infinitely many discrete values for patient WtW. Suppose that we observe from data that a booked appointment with k day's delay is not fulfilled. Upon observing this instance, rank-based choice model infers that the patient's WtW is one of the values in the set $\{0, 1, \ldots, k-1\}$. The same instance for the survival model is an observation that the WtW belongs to one of intervals in the set $\{[s_0, s_1), [s_1, s_2), \ldots, [s_d, s_{d+1})\} \subseteq [0, k) \subseteq \mathcal{T}$.

Suppose we obtain unit intervals $\{[0, 1), [1, 2), \ldots, [d, d+1)\}$ from a dataset empirically (which may not be the case for the FED that the data in Table 2.1 sampled from), an illustration of how each model considers the observation with record ID 1004 from Table 2.1 is shown in Figure 2.2.



(a) Survival Model (b) Rank-based Choice Model

Figure 2.2: Illustration of Observation with Record ID 1004 Based on Each Model

Before we start providing further details on the models, we find it useful to summarize the different types of datasets described above, in Table 2.5.

2.3 WtW Estimation Using Survival Analysis

The literature on survival analysis is vast. Successful applications of survival analysis are observed in clinical trials, health-related systems, and reliability engineering (see, e.g., Wei, 1992; Kessler *et al.*, 1995; Fleming and Lin, 2000). We employ an approach that is similar to traditional survival analysis to characterize the "lifetime of an offered appointment" and use it to estimate the WtW distribution.

Table 2.5: Different Dataset Types Defined in the Study

Available Data	Abbrv.	Description
Full Encounter Dataset	FED	All appt. requests including PLWBA, with dates and final status
Booked Appt. Dataset	BAD	Only booked appts. with dates and final status
Demand Universe Dataset	DUD	All appt requests and their outcomes (booked or PLWBA)
Imputed Booked Appt. Dataset	I-BAD	Dataset appended with PLWBA instances through imputation methods

The following analysis uses a FED, which includes all appointment requests, regardless of the final status (i.e., booked or PLWBA). We evaluate each offered appointment of delay k separately to classify them into two categories (0 or 1). This classification is done based on the relationship between the observed lifetime of the appointment and the delay of that particular appointment. An offered appointment is considered from category 1 if it is fulfilled (i.e., observed lifetime of the appointment is more than the appointment's delay), and 0 if it is either PLWBA or C/RS/NS. Additionally, we transform the data by defining observation intervals $[L, R)$ for every observation. Each observation interval is determined based on that specific appointment's delay and is considered as the interval that contains the appointment's actual lifetime, which is, actually, never observed in the data. For each appointment record i from category 1 in FED, we say that the observation is right-censored; the appointment delay k is the left side, L, of the interval that contains the unobserved appoint-

ment lifetime, and $[L_i, R_i) = [k, \infty)$ for observation i. Similarly, k would be the right side of the interval, R, for the appointments from category 0, and hence, and the observation interval for the appointment record i in this case is $[L_i, R_i) = [0, k)$. Note that this means, in traditional survival analysis terms, that in FED, all of the observations (i.e., appointment records in FED) are either left-censored or right-censored. As a result of this processing, the FED is transformed, as shown in Table 2.6 (for the original FED sample given in Table 2.1).

Table 2.6: Transformed Version of the FED Data Sample in Table 2.1

Record ID	Appt. ID	Patient Type	Category	Appointment Lifetime	
				L	R
1001	123	New	1	3	∞
1002	-	Established	0	0	5
1003	124	Established	0	0	2
1004	125	Established	0	0	6
1005	-	New	0	0	30
1006	126	New	1	10	∞

We use a non-parametric maximum likelihood estimator (NPMLE) developed as a generalized version of the Kaplan-Meier estimator to allow interval censoring (Turnbull, 1976), and we use the notation denoted in (Zhang and Sun, 2010). Let M be the total number of observations (i.e., the total number of appointment requests records in FED), and L_i and R_i be the left and right side of the interval that contains lifetime of record i, respectively. Our case is similar to the clinical studies where there is only one observation time for each subject to observe whether or not the subject "has died" before the observation time (see, e.g., Sun and Kalbfleisch, 1993; Andersen and

Ronn, 1995).

Let s_j, $j \in \{1, \ldots, y+1\}$ be the ordered elements of set $\{0, \cup_i L_i, \cup_i R_i, \infty\}$ that is obtained empirically from the data and let $\mathcal{T} = \{[s_0, s_1), \ldots, [s_y, s_{y+1})\}$. Suppose ρ_{ij} is an indicator variable that shows whether interval j is contained in observation i, $[s_{j-1}, s_j) \subseteq [L_i, R_i)$. As an example, consider the transformed FED in Table 2.6. From this table, we can generate s_j values from set $\{0, 2, 3, 5, 6, 10, 30, \infty\}$ with set $\mathcal{T} = \{[0, 2), [2, 3), [3, 5), [5, 6), [6, 10), [10, 30), [30, \infty)\}$ (where $y = 6$). Then, for record 1001, the indicator variables are $\rho_{11} = \rho_{12} = 0$ and $\rho_{13} = \rho_{14} = \cdots = \rho_{17} = 1$. These intervals are referred as Turnbull intervals, which are defined as the union of disjoint intervals $[l, r]$ that are determined empirically, from available data on lifetime observations.

The data that we have used in this small example results in relatively longer intervals (for instance, $[10, 30)$) since there are only six observations. However, the real-life dataset from the specialty unit that we study consists of almost three years of information on all of the booked appointments. Therefore, the set of intervals generated from our data is almost equal to the set of unit length intervals.

The probability p_j for each interval j is defined as $F(s_j) - F(s_{j-1})$ where F is a non-decreasing function representing the cumulative distribution function of lifetime of an appointment. Then, the log-likelihood function for set of p_j of in vector form $\mathbf{p}$ is proportional to:

$$\mathcal{L}(\mathbf{p}) = \sum_{i=1}^{M} \left(\sum_{j=1}^{y+1} \log \left(\rho_{ij} \left(F(s_j) - F(s_{j-1}) \right) \right) \right) = \sum_{i=1}^{M} \left(\sum_{j=1}^{y+1} \log \left(p_j \rho_{ij} \right) \right). \qquad (2.1)$$

By employing the notation and approach for Turnbull intervals, we estimate

NPMLE, by using our set of observations with the following model:

$$\max \quad \mathcal{L}(\mathbf{p}) \tag{2.2}$$

$$\text{s.t.} \quad \sum_{j=1}^{y+1} p_j = 1, \tag{2.3}$$

$$p_j \geq 0 \ , \qquad \forall\, j \in \{0, \ldots, y+1\}. \tag{2.4}$$

where the objective function maximizes the log-likelihood function based on each observation. Considering each request observed, the model assigns the probabilities of intervals that are contained by the observation to maximize the log-likelihood function. Turnbull (1976) shows that the objective function can be reduced to include only intervals that are contained within the observations, therefore, the model assigns zero probability to intervals that are not observed from the dataset and it is shown that the maximum likelihood estimate cannot be out of the set characterized as the union of Turnbull intervals and likelihood is determined by interval boundaries $[l, r]$.

We use R package icenReg to solve the model and obtain probabilities for each interval (see Anderson-Bergman, 2017a,b, for details) by sequentially updating them based on our data. icenReg package uses an algorithm called EMICM, which was developed to reduce the computation time of the Expectation Maximization (EM) algorithm (Wellner and Zhan, 1997). The algorithm uses EM algorithm in a way that instead of direct maximization of log-likelihood function, the function is reparametrized and approximated by using second-order Taylor expansion to empirically estimate the probabilities (Anderson-Bergman, 2017a).

Recall that in our case FED is not available since no data is collected for the PLWBA instances. Estimating the probabilities p_j by solely using the booked appointments in BAD would result in overestimation of p_j. Therefore, to avoid this overestimation, we impute a set of statistically equivalent PLWBA instances and append them into BAD via random sampling to obtain I-BAD. Ideally, number of

observations sampled should be equal to the number of PLWBA instances. In our case, we only observe the proportion of patients that PLWBA from DUD. Therefore, we sample a certain fraction from BAD to obtain I-BAD where the sampling fraction is determined based on DUD (6.4% in the real data).

For instance, for the depicted BAD in Table 2.2, suppose that two data points are randomly sampled from BAD to insert PLWBA instances (observations with record ID 1002 and 1005) to obtain an Imputed BAD (I-BAD), and further suppose that the sampled observations are the ones with record ID 1001 and 1003. The sampled observations are imputed to BAD (with record IDs 1001-I and 1003-I) to obtain I-BAD that is given in Table 2.7.

Table 2.7: Depiction of Data Contained in I-BAD from BAD in Table 2.2

Record ID	Appt ID	Appt Delay	Patient Type	Status
1001	123	3	New	Seen
1003	124	2	Established	Cancel
1003-I	-	2	Established	Not Booked
1004	125	6	New	No-show
1006	126	10	New	Seen
1001-I	-	3	New	Not Booked

After obtaining I-BAD via random sampling, we treat I-BAD as FED in our survival model. We first transform I-BAD similar to how we transform FED in Table 2.6. Then we use this transformed I-BAD in our model given by Equations (2.2) thru (2.4).

2.4 WtW Estimation Using a Rank-Based Choice Model

As a second approach to understanding patients' appointment fulfilling behavior, we employ a rank-based choice model that is commonly used in retail operations and revenue management. In a retail management setting, upon arrival, each customer either chooses a product among the set of available products at the time of their arrival or leaves without any purchase. The product selection is done based on each customer's rank-based preference list. Therefore, if none of the products in customer's preference list is available at the time of the arrival, the customer leaves without any purchase. Each customer group is assumed to be characterized based on their preference list. In this setting, customers are assumed to always make rank-based decisions based on their preference list. Therefore, based on the type of the purchased product or lack thereof, arriving customer's group can be determined with a certain probability. Notice that total number of customer groups can be as large as the total number of all possible combinations of products of interest.

In our setting, we assume that each patient's preference list is fundamentally driven by his inherent WtW. As we indicate in Section 2.1, we make the simplifying (and generally valid) assumption that patients prefer an earlier appointment to a later one. Therefore, each patient's rank-ordered preference list is assumed to be all of the possible discrete appointment delay values up to and including the patient's WtW. Note that this is an ordered list, from smallest to largest delay values, indicating patient's preference, that is, for a patient with a WtW of k days, the preference list is $(0, 1, \ldots, k)$. We divide the patient population under consideration into a number of "patient groups" that are characterized by distinct preference lists resulting from the different WtW values that patients can have. Specifically, if a patient is from Group k, then the patient is willing to wait at most k days for an appointment. This

indicates that we allow $Z+1$ possible patient groups due to the upper limit Z that we assume on the WtW of a patient. An appointment with a delay of zero days is a same-day appointment; therefore, WtW group 0 includes the patients who are only willing to take a same-day appointment.

Transactional data on appointments provide information on a patients' WtW group in the following way. When a patient books an appointment with a delay of w days and later abandons this appointment via canceling or rescheduling indicating a WtW reason, or the appointment results in a no-show, we infer that the patient is from any one of the WtW groups $k \in \{0, 1, \ldots, w-1\}$. If the patient decides not to book at the time of request, the patient is again inferred to be from any one of WtW groups $k \in \{0, 1, \ldots, w-1\}$. On the other hand, a patient is from any one of WtW groups $k \in \{w, w+1, \ldots, Z\}$ if the patient fulfills a booked appointment with a w-day delay, or cancels/reschedules it for a non-WtW related reason.

We assume that each business day is divided into H time buckets that are short enough such that appointment requests arrive according to a discrete-time homogeneous Bernoulli process with probability $0 < \lambda < 1$. The parameter λ is set to a value obtained from FED (or if FED is not available, jointly from BAD and DUD), representing the proportion of time buckets that an appointment request arrives. We show how we calculate λ in Equation (2.5). The length of time buckets can be set according to the observed daily patient demand. We assume that appointment request arrivals are stationary throughout the day. Our objective is to estimate the probability that an arriving appointment request is from a patient of Group k, denoted as p_k. At each time bucket t, either an appointment request occurs with probability λ or no patient calls. Notice that for the survival model presented in Section 2.3, we denote the total number of observations with M where $H \geq M$ since M can be considered as the total number of time buckets that an appointment request arrives. Also note that,

in a traditional implementation of the rank-based choice model, keeping track of the available products at each time bucket is important since the type of the customer can be estimated by considering the choice that the customer makes between the available product offerings. In our case, we only need to keep track of the earliest available appointment delay at the time of each request.

If we have access to FED, we can identify the time buckets in which an arrival occurs, the offered appointment delay at each time bucket, whether the patient decided to book the appointment, and whether the appointment booked at that time bucket is ultimately fulfilled or the reasons for the abandonment. This allows us to divide the study period, consisting of H time buckets, into four disjoint sets. The sets $\mathcal{S}$, $\mathcal{A}$ and $\mathcal{W}$ represent the set of time buckets that an appointment request results in: either a fulfilled appointment or a booked appointment that is later cancelled or rescheduled due to a non-WtW reason, a booked appointment that is subsequently abandoned through either a no-show or cancelled or rescheduled due to a stated reason that is WtW-related, and no booking (i.e., PLWBA) due to a WtW reason, respectively. The set $\mathcal{N}$, on the other hand, represents the set of time buckets with no appointment requests. We denote the set of all time buckets with $\mathcal{H} = \mathcal{S} \cup \mathcal{A} \cup \mathcal{W} \cup \mathcal{N}$.

As a first step for our analysis, similar to what we did in Section 2.3, we transform FED. The main difference here is the existence of time buckets in the rank-based choice model. Therefore, we first assign each observation into an associated time bucket, t, of the day based on time of the request (which is included in the dataset) and include the time buckets with no appointment request. Note that we have two different patient types (i.e., new and established) in our real-life dataset. Since different patient types may have different arrival patterns, we assign time buckets separately for each type and analyze them separately in our models. We then assign each time bucket to their associated set (i.e., $\mathcal{S}$, $\mathcal{A}$, $\mathcal{W}$, or $\mathcal{N}$).

Table 2.8: Updated FED After Including Time Buckets

Day	t	Record ID	Appt ID	Appt Delay (w_t)	Pt. Type	Status	Set
5	1	1001	123	3	New	Seen	$\mathcal{S}$
5	2	-	-	-	-	No Arrival	$\mathcal{N}$
5	3	-	-	-	-	No Arrival	$\mathcal{N}$
5	4	1004	125	6	New	No-show	$\mathcal{A}$
5	5	1005	-	30	New	Not Booked	$\mathcal{W}$
5	6	1006	126	10	New	Seen	$\mathcal{S}$

Table 2.9: Transformed Version of the Data in Table 2.8

Day	t	Record ID	Appt ID	$\Theta_t(w_t)$	Patient Type	Status	Set
5	1	1001	123	$\{3, \ldots, Z\}$	New	Seen	$\mathcal{S}$
5	2	-	-	-	-	No Arrival	$\mathcal{N}$
5	3	-	-	-	-	No Arrival	$\mathcal{N}$
5	4	1004	125	$\{0, \ldots, 5\}$	New	No-show	$\mathcal{A}$
5	5	1005	-	$\{0, \ldots, 29\}$	New	Not Booked	$\mathcal{W}$
5	6	1006	126	$\{10, \ldots, Z\}$	New	Seen	$\mathcal{S}$

Table 2.8 shows the transformed version of the FED we have presented above. This dataset is then transformed again to indicate the discrete set that the WtW of the patient in each record can belong to, denoted as $\Theta_t(w_t)$, which is a function of the appointment delay as well as the set membership of the time bucket t. The transformed-again dataset is shown in Table 2.9, which we use for the analysis.

We use a maximum likelihood estimator (MLE) for estimating the probabilities p_k, where $k \in \{0, 1, \ldots, Z\}$, or in vector notation, $\mathbf{p}$. If we have access to a FED, we can

obtain the set membership of each time bucket, denoted as q_t for all $t \in \{0, \ldots, H\}$, or $\mathbf{q}$ in vector notation, as well as the offered appointment delays, w_t. In addition, from FED, we can obtain:

(i) an estimate for rate λ as the fraction of time buckets that an appointment is requested, i.e.,

$$\lambda = \frac{|\mathcal{S}| + |\mathcal{A}| + |\mathcal{W}|}{H} = 1 - \frac{|\mathcal{N}|}{H} \ , \text{ and} \tag{2.5}$$

(ii) an estimate for the probability that a patient with WtW less than the offered delay will still go ahead and book an appointment, which we denote as $1 - \alpha$, and will subsequently abandon the appointment, that is,

$$\alpha = \frac{|\mathcal{W}|}{|\mathcal{W}| + |\mathcal{A}|} \tag{2.6}$$

If one only has access to BAD, it is not possible to differentiate the time buckets in set $\mathcal{W}$ (i.e., PLWBA instances) and the time buckets in $\mathcal{N}$ (i.e., no appointment request arrivals). In this case, we treat the set of time buckets with no recorded events in BAD as $\mathcal{B} = \mathcal{W} \cup \mathcal{N}$. To be able to conduct the analysis on BAD, we first update BAD by including the time buckets into BAD, as given in Table 2.10.

After that, for any time bucket $t \in \mathcal{B}$, we impute an appointment delay value that might have been offered to an arriving appointment request in that time bucket using the following argument. In any time bucket with no registered event in the BAD, if a patient arrived and left without booking (i.e., PLWBA), the offered appointment was then offered to and booked by another subsequent patient, which is included in BAD. We impute "likely" appointment delays for the time buckets with no events and obtain I-BAD shown in Table 2.11.

Note that the difference between the data shown in Table 2.11 (I-BAD) and Table 2.9 (FED) is the fact that in Table 2.11 we do not know which of the time buckets

Table 2.10: Updated BAD After Including Time Buckets

Day	t	Record ID	Appt ID	Appt Delay (w_t)	Pt. Type	Status	Set
5	1	1001	123	3	New	Seen	$\mathcal{S}$
5	2	-	-	-	-	No Event	$\mathcal{B}$
5	3	-	-	-	-	No Event	$\mathcal{B}$
5	4	1004	125	6	New	No-show	$\mathcal{A}$
5	5	-	-	-	-	No Event	$\mathcal{B}$
5	6	1006	126	10	New	Seen	$\mathcal{S}$

Table 2.11: Depiction of I-BAD Obtained by Imputation for the Rank-based Choice Model, from the Updated BAD Given in Table 2.10

Day	t	Record ID	Appt ID	Appt Delay (w_t)	Pt. Type	Status	Set
5	1	1001	123	3	New	Seen	$\mathcal{S}$
5	2	-	-	6	New	No Event	$\mathcal{B}$
5	3	-	-	6	New	No Event	$\mathcal{B}$
5	4	1004	125	6	New	No-show	$\mathcal{A}$
5	5	-	-	10	New	No Event	$\mathcal{B}$
5	6	1006	126	10	New	Seen	$\mathcal{S}$

in $\mathcal{B}$ are actually in $\mathcal{W}$ and which are in $\mathcal{N}$. Recall that even in the case of BAD, α is still parameterizable due to availability of DUD. DUD allows us observe the fraction of appointment requests that result in PLWBA over all appointment requests (which is 6.4% for the dataset that we use). Denoting this fraction obtained from DUD as β, we can write

$$\alpha = \frac{|\mathcal{W}|}{|\mathcal{W}| + |\mathcal{A}|} = \frac{|\mathcal{W}|}{|\mathcal{S}| + |\mathcal{W}| + |\mathcal{A}|} \cdot \frac{|\mathcal{S}| + |\mathcal{W}| + |\mathcal{A}|}{|\mathcal{W}| + |\mathcal{A}|} = \beta \, \frac{\lambda \, H}{\lambda \, H - |\mathcal{S}|} \ , \quad (2.7)$$

where the terms in the RHS are all available. For the instances that are observed to be in $\mathcal{B}$ from BAD, we can write

$$P(t \in \mathcal{B}) = P(t \in \mathcal{W}) + P(t \in \mathcal{N}) = \lambda\alpha \sum_{k=0}^{w_t-1} p_k \; + \; (1-\lambda) \; . \tag{2.8}$$

In order to address the incompleteness due to not observing the exact set membership in $\mathcal{B}$, we employ the approach that is introduced in van Ryzin and Vulcano (2017). In the study, van Ryzin and Vulcano (2017) suggest using a simplified approach that considers the log-likelihood function that one would obtain for a dataset that also includes information on the WtW group of the patients in each time bucket, t. Let g_t denote the observed WtW group of the patient in time bucket t. Then, given the vector of set memberships for each time bucket observed in the dataset, $\mathbf{q}$, w_t for all t and the estimated parameters λ and α, we can write the log-likelihood of a given probability vector, $\mathbf{p}$, as follows.

$$\begin{aligned} \mathcal{L}_C(\mathbf{q}, \mathbf{g}|\mathbf{p}) \;\; = \;\; & \sum_{t\in\mathcal{S}} \log\left(\lambda \, p_{g_t}\right) + \sum_{t\in\mathcal{A}} \log\left(\lambda(1-\alpha) p_{g_t}\right) + \sum_{t\in\mathcal{W}} \log\left(\lambda\alpha p_{g_t}\right) \\ & + \sum_{t\in\mathcal{N}} \log(1-\lambda) = \sum_{t\in\mathcal{H}} a_t \log\lambda + (1-a_t)\log(1-\lambda) \\ & + \sum_{t\in\mathcal{H}} a_t \log p_{g_t} + \sum_{t\in\mathcal{A}} \log(1-\alpha) + \sum_{t\in\mathcal{W}} \log\alpha \; , \end{aligned} \tag{2.9}$$

where $a_t = \mathbb{1}\{q_t \in \mathcal{S} \cup \mathcal{A} \cup \mathcal{W}\}$. To maximize this log-likelihood function, it is sufficient to maximize the term

$$\widetilde{\mathcal{L}}_C(\mathbf{q}, \mathbf{g}|\mathbf{p}) = \sum_{t\in\mathcal{H}} a_t \log p_{g_t} = \sum_{k=0}^{Z} \sum_{t\in\mathcal{H}} a_t \mathbb{1}\{k = g_t\} \log p_k = \sum_{k=0}^{Z} m_k \log p_k \tag{2.10}$$

with respect to the values p_{g_t} where m_k indicates number of patients arriving from each group k. Then, the model becomes

$$\max \qquad \sum_{k=0}^{Z} m_k \log p_k \tag{2.11}$$

$$\text{s.t.} \qquad \sum_{k=0}^{Z} p_k = 1, \tag{2.12}$$

$$p_k \geq 0 \; , \qquad \forall \; k \in \{0, \ldots, Z\}. \tag{2.13}$$

We need to use KKT conditions to identify optimal $\mathbf{p}^*$ where we can write the Lagrangian as

$$L(\mathbf{p}, \mu) = \sum_{k=0}^{Z} m_k \log p_k + \mu \left(1 - \sum_{k=0}^{Z} p_k\right) \ , \tag{2.14}$$

which results in $p_k^* = \frac{m_k}{\mu}$ and $\mu = \sum_{k=0}^{Z} m_k$ which satisfies $p_k \geq 0$. Therefore, for the complete log-likelihood case p_k^* is simply the fraction of patients arriving from group k.

In reality, even when one has access to the FED, the observations only include a_t values. In that case, the log-likelihood function that one needs to optimize is obtained by taking an expectation over the WtW group of the patients, as follows.

$$\begin{aligned} \widetilde{\mathcal{L}}_{\text{FED}}(\mathbf{q}|\mathbf{p}) &= E[E[\widetilde{\mathcal{L}}_{\text{C}}(\mathbf{q}|\mathbf{p})|\mathbf{G}]] = \sum_{t \in \mathcal{H}} a_t \sum_{k=0}^{Z} \log p_k \ P(G_t = k) \\ &= \sum_{t \in \mathcal{H}} a_t \sum_{k=0}^{Z} \mathbb{1}\{k \in \Theta_t(w_t)\} \frac{1}{|\Theta_t(w_t)|} \log p_k \end{aligned} \tag{2.15}$$

where G_t denotes the random variable that represents the WtW group of the patient in time bucket t, assuming that the patient belongs to each one of the possible WtW groups in $\Theta_t(w_t)$ with equal likelihood.

In the case of availability of only I-BAD (as in our case), we additionally have to take an expectation over the arrivals, since we do not exactly know which time buckets in the set $\mathcal{B}$ has arrivals (i.e., PLWBA) and no arrivals. Let A_t denote the binary random variable that indicates arrival in a time bucket t. Then, in this case,

$$\begin{aligned} \widetilde{\mathcal{L}}_{\text{I-BAD}}(\mathbf{q}|\mathbf{p}) &= E[E[\widetilde{\mathcal{L}}_{\text{C}}(\mathbf{q}|\mathbf{p})|\mathbf{G}, \mathbf{A}]] \\ &= \sum_{t \in \mathcal{S} \cup \mathcal{A}} \sum_{k=0}^{Z} \log p_k \ P(G_t = k) + \sum_{t \in \mathcal{B}} P(A_t = 1 | t \in \mathcal{B}) \sum_{k=0}^{Z} \log p_k \ P(G_t = k) \\ &= \sum_{t \in \mathcal{S} \cup \mathcal{A}} \sum_{k=0}^{Z} \mathbb{1}\{k \in \Theta_t(w_t)\} \frac{1}{|\Theta_t(w_t)|} \log p_k \\ &\quad + \sum_{t \in \mathcal{B}} \frac{\lambda \alpha \sum_{k=0}^{w_t - 1} p_k}{\lambda \alpha \sum_{k=0}^{w_t - 1} p_k + (1 - \lambda)} \sum_{k=0}^{Z} \mathbb{1}\{k \in \Theta_t(w_t)\} \frac{1}{|\Theta_t(w_t)|} \log p_k \end{aligned} \tag{2.16}$$

Solving the model for $\widetilde{\mathcal{L}}_{\text{I-BAD}}(\mathbf{q}|\mathbf{p})$ to obtain $\mathbf{p}^*$ is not straightforward, therefore, we utilize an Expectation-Maximization (EM) algorithm (Dempster *et al.* (1977)). The details of EM algorithm are presented in Appendix A.1.

2.5 Testing Performance of Estimations on Simulated Data

To test the effectiveness of our methodology, we generate realistic datasets carefully simulated to resemble real-life transactional data. This testing methodology gives us the ability to assess the accuracy of the predictions with respect to the assumed characterizations of patient WtW in the simulation model. In particular, we focus on a single patient type and build a discrete event model that simulates patient arrivals, appointment delays offered to them and the decisions that they make, which are randomly generated from the assumed WtW distribution. We rigorously test the performance of the estimation methods given in Sections 3 and 4 by comparing the appointment realization probabilities indicated by our estimations with the ones implied by the WtW distribution assumed in our simulation model. Using a simulated dataset also allows us to test the behavior of our estimation methodologies under different cases of available data; essentialy, we generate a complete FED that includes all patient responses including PLWBA, and then remove them to assess the degree to which the developed imputation methods can improve estimations from a BAD.

2.5.1 Test Data Generation

The real data is collected from a system with highly complex dynamics that are not easy to replicate. Therefore, we generate the simulated data with the fundamental characteristics that we can observe from the data. We simulate patient arrivals as a Poisson process where, upon each arrival, the first available appointment slot is offered to the patient from a calendar system with limited capacity and limited booking

horizon. For each delay value k, there is an assumed "true" realization probability, which is non-increasing in k, in accordance with our assumption that earlier appointments are always preferred. As in Section 4, let p_k denote this "true" probability that an offered appointment with delay k will be realized (i.e., booked and subsequently fulfilled). Note that $p_k = P(\text{WtW} \geq k)$, so the patient WtW distribution is represented as a discrete cdf by the p_k values assumed in the simulation model.

At the time of each appointment request, we generate a discrete WtW value for the patient requesting the appointment. If the patient's WtW is more than the offered delay, the appointment is booked and realized. If the patient's WtW is less than the offered delay, the patient can do one of the three possible things; (i) PLWBA, (ii) books the appointment and then cancels, or (iii) decides not to show up at the appointment. We do not include rescheduled appointments into the simulation model; instead, we consider them as canceled appointments and consider the appointment that is scheduled as a result of reschedule request from the patient as a new appointment request since we only consider each encounter in the real data. We simulate this process by distributing patients according to a set of delay dependent probabilities, which are directly obtained from the available data. To represent the real system, we fix PLWBA rate at 6.4% to match the observed lost patient rate from our DUD. Then from the data, we calculate the delay dependent no-show probabilities as the fraction of no-shows among the appointments that are not fulfilled for each delay value. We calculate the delay dependent cancellation probabilities in a similar manner. The delay dependent no-show and cancellation probabilities are shown in Figure 2.3. For instance, for a delay value of 15 days, 62.5% of the patients who have WtW less than 15 days end up canceling their booked appointments while 12.5% of them do not show up to their appointments and the remaining ones PLWBA. Notice that the probability of PLWBA among the ones with WtW less than the appointment delay

is equal to the value α that we described in Equation (2.6) and represents the 6.4% of the total encounters.

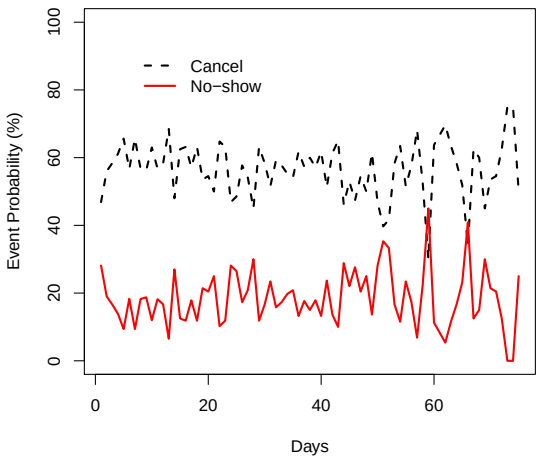


Figure 2.3: Delay Dependent Cancellation and Rescheduling Probabilities

Additionally, if the appointment is booked and then cancelled, we generate the day of cancellation. For each delay value, a cancellation date before the day of appointment is assigned to the patient with a certain probability. These probabilities are determined from the C/RS cases in the data as the fractions of appointments that were cancelled or rescheduled on days $\{1, 2, \ldots, k-1\}$ among the appointments with offered with k days delay. In the case of cancellations, after a certain amount of time (but before the day of appointment), the patient cancels the appointment and the appointment slot becomes available, while, in the case of no-shows, the appointment slot is never released once it is booked since we only observe no-shows at the time of the appointment.

We run the simulation model to obtain a simulated FED (referred to as sFED). Each row of sFED represents a generated appointment request, and columns show the day and time bucket of the day, the offered appointment delay as well as the

final status of the encounter (fulfilled, canceled, no-showed, and not booked). We consider 48 time buckets per day in our simulation analyses. In addition to testing the performance of our models with respect to "true" realization probabilities, simulated datasets allow us to consider the effect of the number of data points on our estimations. We run our simulation model for 100 and 1000 days to observe whether probability estimates can be improved over longer periods of data collection. We generate two separate sFEDs for 100 day model and 1000 day model separately and generate sFED(100) and sFED(1000). After obtaining sFED(100) and sFED(1000), we generate a total of 10 random folds from the each of these datasets. We obtain 10 different training sets each of which consists of 9 of the 10 random folds; we refer to each of them as Tr-1 to Tr-10. Each of the training sets allow us to observe whether a time bucket results in PLWBA or not. From each training set, Tr-1 to Tr-10, we removed the observations with PLWBA and obtained 10 different sBADs. We refer these sBADs as sBAD-Tr-1 to sBAD-Tr-10.

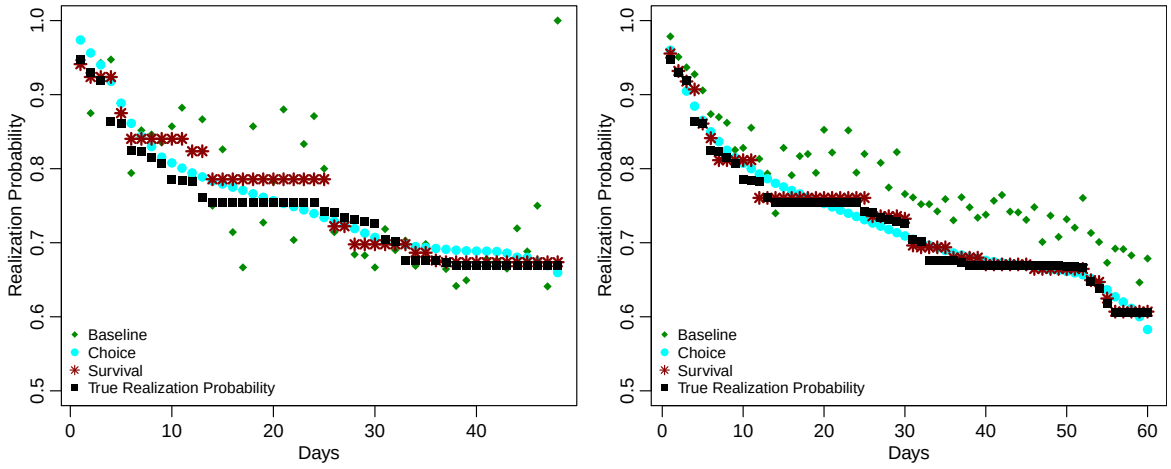
2.5.2 Observed Errors of Estimations Obtained by Each Model

The two models that we propose in Sections 2.3 and 2.4 aim to characterize an underlying WtW distribution for the patient population under study. This characterization is expressed and tested with respect to the accuracy of estimates for p_k, the probability that an offered appointment with delay k will be realized (i.e., booked and eventually fulfilled), which is equal to $P(\text{WtW} \leq k)$. In particular, we compare the estimated p_k values by the two methods with the "true" realization probabilities assumed in the simulation model and calculate statistics on absolute errors.

To provide a benchmark, we consider the method of estimating the p_k directly from the data by calculating the fraction of realized appointments among all appointments with delay k. We refer the p_k estimates obtained in this fashion as "baseline." Note

that a major issue with the baseline method is that the data may not be homogeneous with respect to the observations with different levels of appointment delay. Hence, some of the p_k estimates with the baseline method are obtained based on a lot fewer observations, making our confidence in some of these estimations less than other ones. Furthermore, in some cases, there may be no observations with delay k, making it impossible to obtain an estimate for p_k. In contrast, both survival and rank-based choice models use the available data in a more "holistic" manner, to make inferences on the patients' WtW.

We employ 10-fold cross validation to obtain absolute errors. We use 150 days as an upper bound for patients' WtW throughout the study (i.e., Z value for the rank-based choice model, determined via experimentation with different values), and using each model, we obtain 10 different sets of realization probabilities estimated from each fold of sFED(100) and from each fold of sFED(1000). The reported statistics on mean absolute errors (average, minimum and maximum MAD) are obtained over the 10-folds for sFED(100) and 10-folds for sFED(1000). Figures 2.4(a) and 2.4(b) show the realization probabilities estimated with the three methods along with the "true" realization probabilities for one fold of sFED(100) and sFED(1000, respectively. The figures suggest that both models obtain probability estimates that are close to "true" probabilities and show significantly better estimation performance compared to baseline model, while representing reasonable patient responses to delays (e.g., nonincreasing in k), in accordance with our assumptions. Additionally, our results show that probability estimates obtained with the survival and rank-based choice models drastically improve through the use of more data points since we observe sFED(1000) estimates look closer to "true" probabilities compared to sFED(100) in the figures. This observation is not trivial since we see below that the error metrics for the baseline method does not necessarily improve with more data. We observe similar results



(a) Single Fold Results from sFED(100)

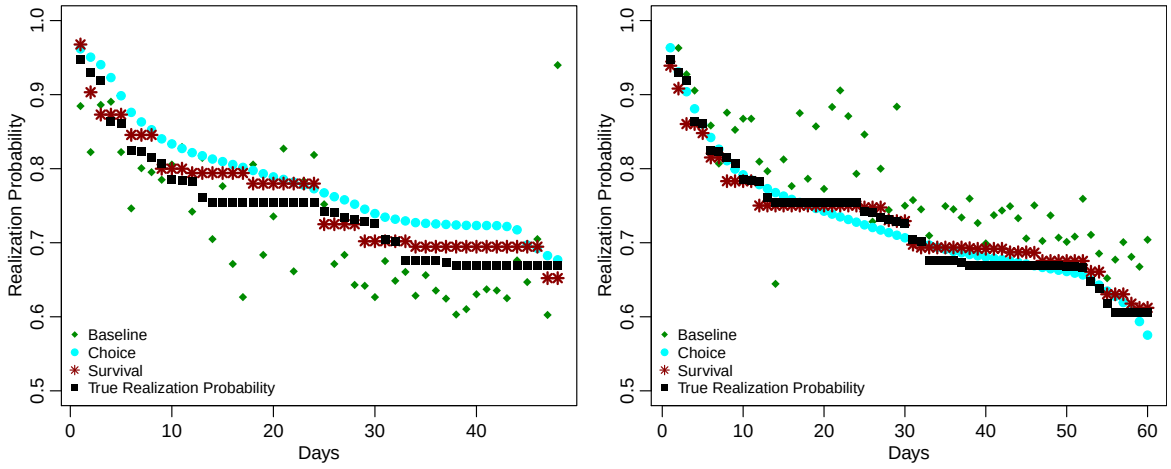
(b) Single Fold Results from sFED(1000)

Figure 2.4: Estimates Obtained by the Three Different Models for One Fold of sFED(100) and sFED(1000)

from our estimates on each training set that we generate.

We next test the performance of the three estimation methods with and without imputation. To this end, we obtain imputed versions of each of the 10 folds (denoted as I-sBAD-Tr-1 to I-sBAD-Tr-10) for the survival model and the rank-based choice model; recall that the imputation method is slightly different for the two algorithms. For the baseline model, we use the imputation strategy presented for the survival model and calculate the fraction of realized appointments from this I-sBAD. Figures 2.5(a) and 2.5(b) show the estimations obtained from a single fold of I-sBAD(100) and I-sBAD(1000), respectively. We make similar observations to the ones made for the estimations obtained from sFED folds, which indicates that it is possible to obtain similar characterizations with FED and BAD, through imputation and estimation methods we have outlined.

To emphasize the above visual observations through a quantification of the errors, we calculate the MAD of each model for all training sets; the error statistics are



(a) Single Fold Results from I-sBAD(100) (b) Single Fold Results from I-sBAD(1000)

Figure 2.5: Estimates Obtained by the Three Different Models for One Fold of I-sBAD(100) and I-sBAD(1000)

provided in Table 2.12. We denote survival model as S, rank-based choice model as C, and baseline model as B in the table.As indicated above, we first note that the errors observed for both survival model and the rank-based choice model improve drastically (from about 2% to less than 1%) as the data increases from 100 days to 1000 days' worth of patient transactional data. The baseline model estimations actually get slightly worse, as more data is used, raising questions about the robustness and intuitiveness of the method since one generally expects the estimations to get better as more data is used in the estimations. An additional fundamental issue with using baseline model is that since it depends on observations for each individual delay value, it might result in no estimation for a delay value if the delay is not observed in the data, or it might estimate the probabilities for a delay value as 0 or 1, if there is a single observation for that delay. We observe MAD range for baseline being wider than the other models since baseline model can arbitrarily estimate the realization probabilities since it does not assume an underlying distribution or a functional form

to estimate the probabilities.

Table 2.12: Statistics on MAD for Simulated Data on 100 Days Versus 1000 Days

Results from 100-days simulated data						
	MAD Range (min, max) over 10 training data folds			Average MAD		
	sFED	I-sBAD	sBAD	sFED	I-sBAD	sBAD
S	(0.0000, 0.0945)	(0.0000, 0.9473)	(0.0000, 0.1998)	0.020	0.024	0.060
C	(0.0002, 0.0787)	(0.0063, 0.1352)	(0.0003, 0.3876)	0.026	0.045	0.065
B	(0.0002, 0.2712)	(0.0004, 0.3313)	(0.0004, 0.7427)	0.044	0.050	0.115
Results from 1000-days simulated data						
	MAD Range (min, max) over 10 training data folds			Average MAD		
	sFED	I-sBAD	sBAD	sFED	I-sBAD	sBAD
S	(0.0000, 0.0433)	(0.0000, 0.0624)	(0.0000, 0.1726)	0.009	0.014	0.0624
C	(0.0000, 0.0327)	(0.0000, 0.0400)	(0.0000, 0.3158)	0.010	0.011	0.063
B	(0.0000, 0.1295)	(0.0000, 0.1571)	(0.0001, 0.4231)	0.055	0.061	0.107

Our suggested models are using the available data in a more comprehensive way since the p_k estimations are made by using all observations that contain that delay value. For instance, while baseline model estimates the probability for 10 days' delay only using the observations with 10 days' offered delay, survival and rank-based choice models use the data from appointments being realized with delays 10 days or more, and appointments that are not fulfilled with delay less than 10 days.

For all three methods, imputation improves the quality of estimations drastically; for the survival model and the rank-based choice model the average MAD values reduce to about one fifth of those observed without imputation. Furthermore, the average MAD values obtained with imputed BAD are almost as low as those obtained

with the FED, clearly demonstrating the feasibility of using our methods on less-than-complete datasets that only include information on booked appointments.

2.5.3 Testing Goodness-of-Fit of Obtained WtW Distributions

In this section, we present a statistical test to evaluate the goodness-of-fit of an obtained WtW distribution to be used in real-life data analysis, when "true" realization probabilities are not available. In particular, we employ a hypothesis testing approach that checks the statistical significance of an estimate for the WtW distribution.

Consider 95% confidence intervals (CI) on the fraction of appointments realized for each delay value, calculated from the dataset, FED or I-BAD, which includes PLWBA instances. Since the number of realized appointments are binomially distributed with p_k, the so-called Wilson confidence interval (see, e.g., Brown *et al.* (2001)) provides the range of values that includes the true probability that an offered appointment with k-days' delay will be realized with, say, 95% probability. Wilson CIs on fraction of appointments realized for each appointment delay is obtained by

$$\frac{n_s + z^2/2}{n + z^2} \pm \frac{z}{n + z^2}\sqrt{\frac{n_s n_f}{n} + \frac{z^2}{4}}\ , \tag{2.17}$$

where n_s refers to the number of appointments realized out of n appointments offered with a certain delay, n_f is the number of appointments not realized, and z denotes the z-value from standard normal distribution where z is equal to 1.96 in our case (for 95% CI). Hence, under the hypothesis that the WtW characterization is a good one, each p_k estimate will lie in the associated Wilson interval with 95% probability, and assuming independence, the number of p_k estimates that fall outside the Wilson intervals should be binomially distributed with T trials (number of possible appointment delay values) and 0.05 "success" probability. Hence, we conduct a one-tailed Binomial

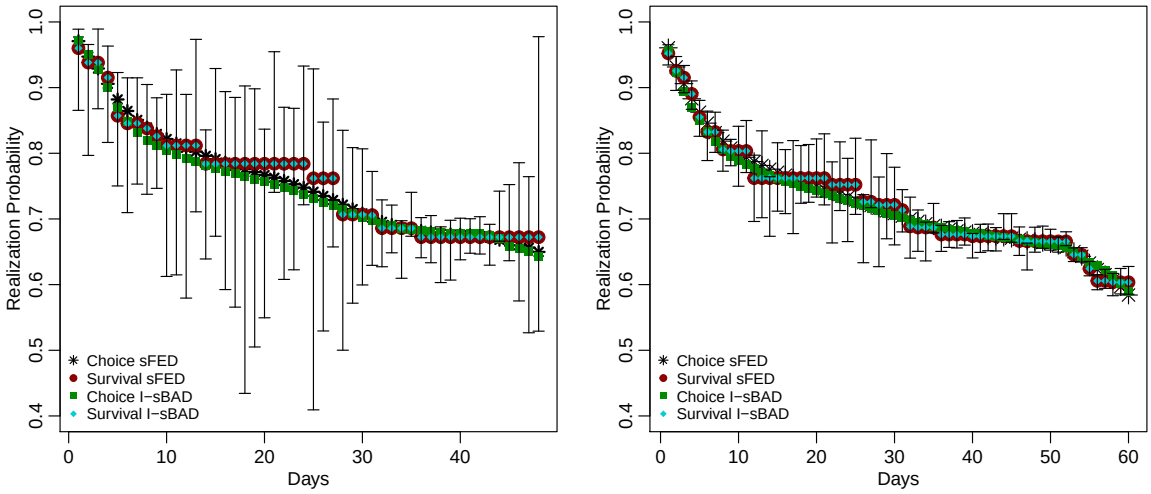
test (with H_0 : "success" probability = 0.05 and H_a : "success" probability > 0.05) and calculate a p value using the observed number of estimates that fall outside the Wilson CIs.

We follow the above-mentioned procedure for the realization probabilities obtained through the two models on sFED(100), sFED(1000), I-sBAD(100), and I-sBAD(1000) by generating the 95% Wilson CIs for each delay value k. We report the p-values in Table 2.13, and plot the probability estimates and Wilson CIs as a function of appointment delay in Figure 2.6. We observe that the data depicts behavior that would be strongly aligned with the WtW characterizations obtained with either of the two models (i.e., fail to reject the null hypothesis at $\alpha = 0.05$) and hence, can be considered to be effective mechanisms to estimate WtW distribution with respect to the data at hand.

Table 2.13: p-values from the Goodness-of-fit Tests

	sFED(100)	sFED(1000)	I-sBAD(100)	I-sBAD(1000)
Survival Model	1.000	0.954	1.000	0.954
Choice Model	1.000	0.180	1.000	0.079

Note that since we have fewer number of observations in sFED(100) compared to sFED(1000), and hence, the confidence intervals that are generated from sFED(100) are wider. This again shows us the importance of collecting more data on offered (and/or booked) appointments to obtain more accurate estimations for realization probabilities.



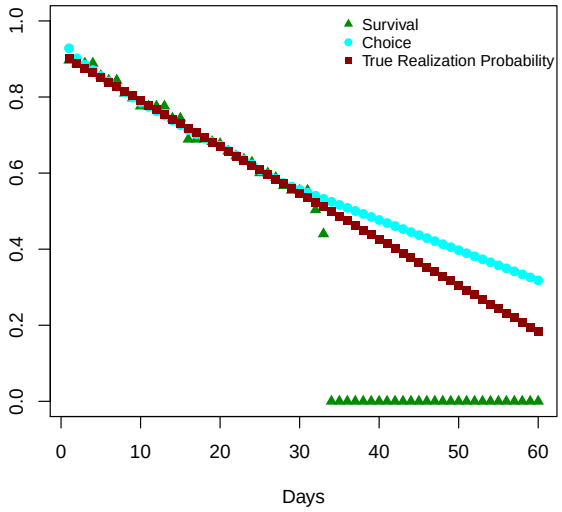
(a) Results from 100 Days Simulated Data (b) Results from 1000 Days Simulated Data

Figure 2.6: Comparison of Estimates from Different Models with 95% CIs from sFED

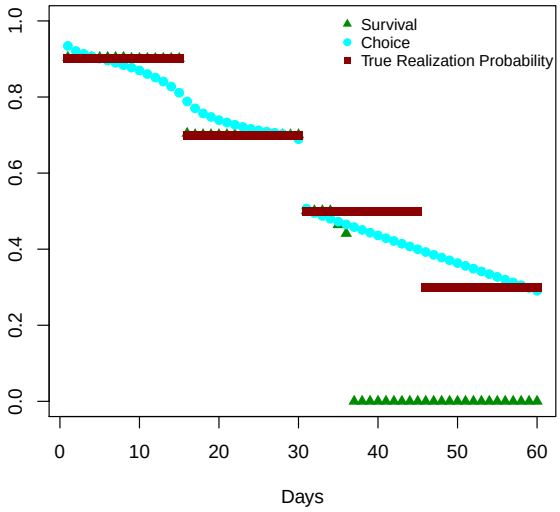
2.5.4 Further Analysis with Multiple Simulated FEDs

In the above analyses, we used a single set of "true" realization probabilities that reflected the WtW behavior observed from our real-life dataset (details of which are provided in the next section) to generate sFED(100) and sFED(1000). To gain further insights into how our models perform under different settings of patient WtW behavior, we extend our numerical analysis by considering alternative sets of "true" realization probabilities (Cases 1 thru 4), and sFEDs generated for each case (referred to as FED#1 thru FED#4). The four cases of true realization probabilities depict somewhat extreme types of responses; ranging from linearly/nonlinearly decreasing to step-wise constant. In our simulation model, we keep the other parameters the same while generating new sFEDs. We employ both estimation models on each sFED we generate, and report on performance of the models.

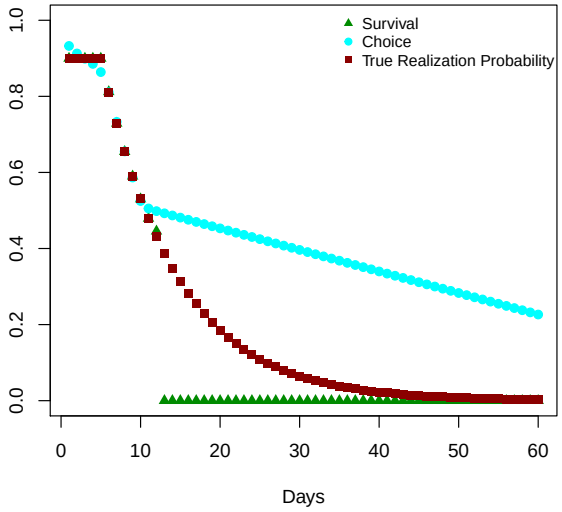
Figure 2.7 shows the estimations obtained from each sFED along with the "true" realization probabilities used to generate the four sFEDs. We observe that the models



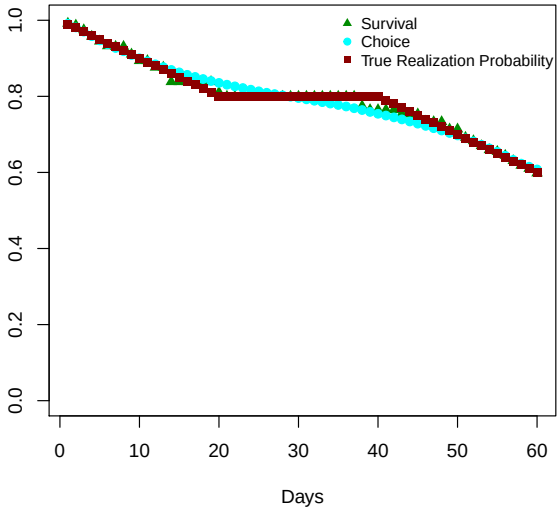
(a) Realization Probabilities for FED#1



(b) Realization Probabilities for FED#2



(c) Realization Probabilities for FED#3



(d) Realization Probabilities for FED#4

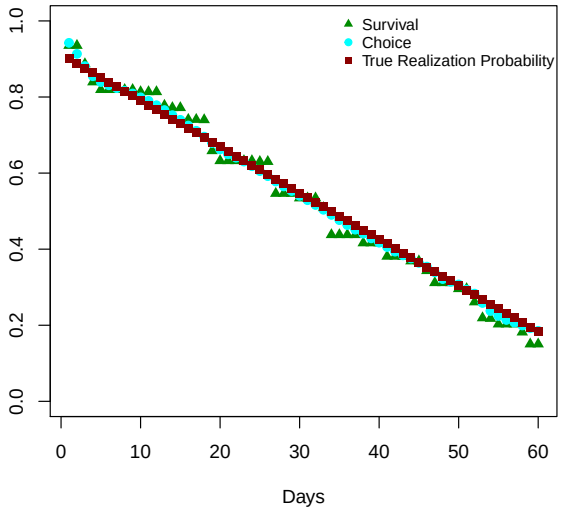
Figure 2.7: Performance of Proposed Models on Alternative sFEDs

perform poorly after a certain appointment delay except for FED#4. The reason is that our estimates are limited by the available data, hence we expect higher errors on the p_k values with very rarely observed offered delays. Not observing certain delay values might be due to the length of the booking horizon or the congestion in the system. For instance, we expect to observe high delay values rarely in less congested systems.

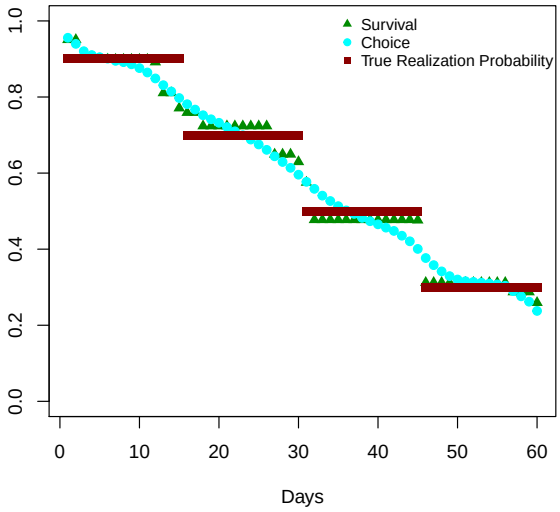
Therefore, it is important to have homogeneous data that contains transactions with different offered appointment delays. For instance, in FED#1, the maximum offered appointment delay is observed as 33 days. Therefore, our models start to perform poorly after day 33 where survival analysis assigns zero to p_k for $k > 33$ since there are no records collected for those values while the rank-based choice model assigns the same probability to each individual value. Notice that the realization probability that the models start to perform poorly (maximum observed delay in each FED) is almost the same for all FEDs since this value represents the required dilution level in arrival rate to stabilize the system.

We expect our proposed models to perform better in a more congested system under Cases 1, 2 and 3, and perform worse in a less congested system under the realization probability Case 4. We test this claim by obtaining new sFEDs (referred to as FED#1-2 thru FED#4-2) by using the same "true" probability cases but with less service capacity for Cases 1-3, and with more service capacity for Case 4. Figure 2.8 shows our estimations with the new set of sFEDs.

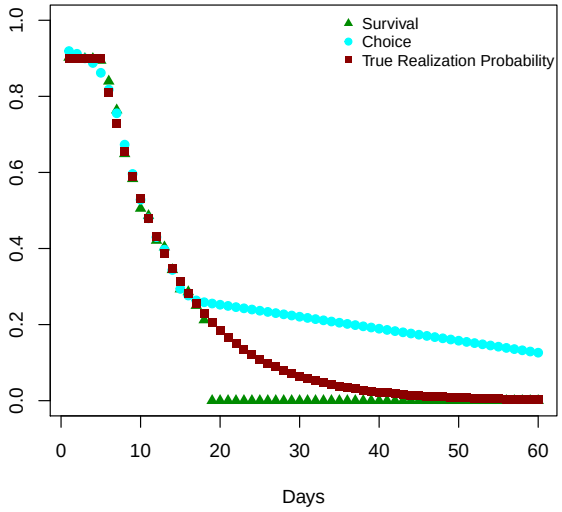
Our experiments on new sFEDs show that both models are effective in estimating the probabilities as long as we can observe the related data for the delay. One can discuss whether it is crucial to estimate the realization probabilities for the delays not offered and not observed. While these appointment delays are not offered to the patients during a certain study period, those delay values are possible due to length



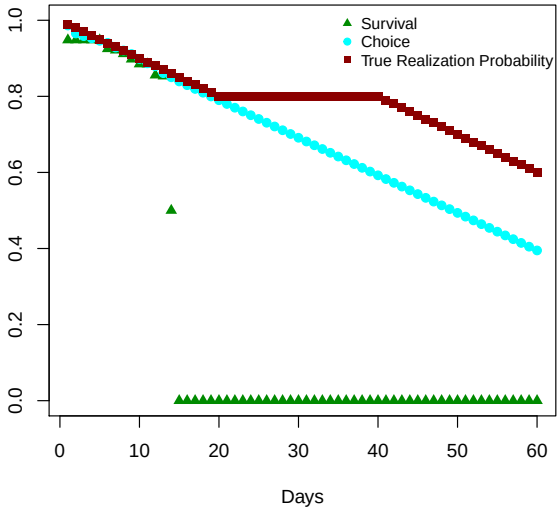
(a) Realization Probabilities for FED#1-2



(b) Realization Probabilities for FED#2-2



(c) Realization Probabilities for FED#3-2



(d) Realization Probabilities for FED#4-2

Figure 2.8: Performance of Proposed Models on Second Set of Alternative sFEDs

of booking horizon. Therefore, in the case of changing the scheduling policy and utilizing appointment slots that are further into future, it is suggested to collect FED or BAD, if possible, for a certain period of time to observe different values of offered delays and re-estimate the realization probabilities by using the proposed models.

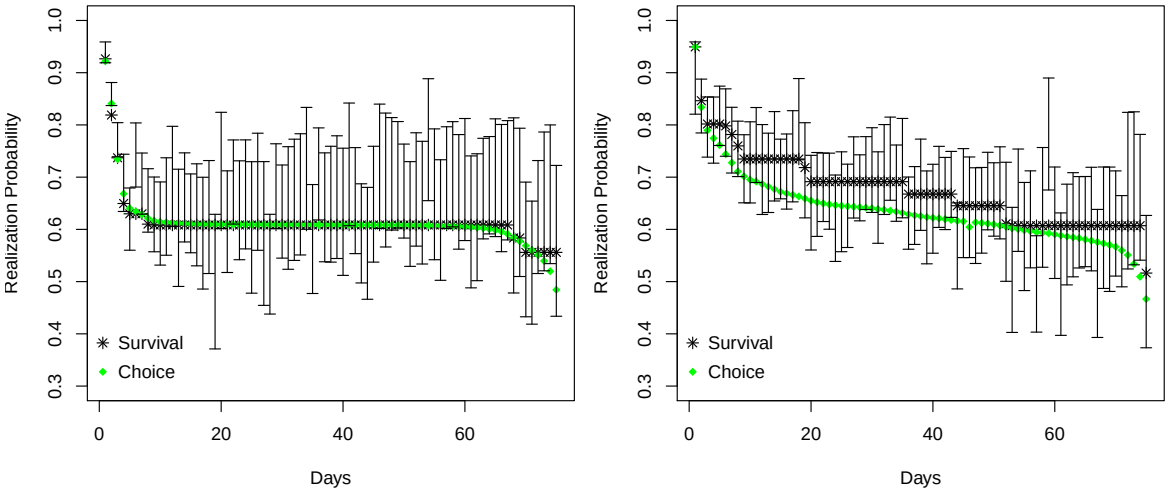
Another way to generate an alternative FED is using the same set of "true" realization probabilities but changing the distribution of the no-shows and cancellations for each delay value for the appointments that are not realized. Note that changing the distribution depicted in Figure 2.3 only affects the offered delays observed. If patients decide to no-show more compared to Figure 2.3, we expect to observe higher delays offered to patients since the system congestion will be higher due to no-show slots being occupied until the time of the appointment while cancellations result in slots being available before the time of the appointment. We use two different sets of no-show and cancellation probabilities as inputs and generate two alternative sFEDs to observe whether the choice of input changes the quality of estimates. We use the statistical testing framework that we presented in Section 2.5.3 to test the quality of our estimates and conclude that the choice of cancellation and no-show probabilities for the appointment is not realized do not affect the performance of the estimation methods significantly.

2.6 Case Study on Real-life Patient Transactional Data

In this section, we use the data that is described in Section 2.2 to gain insights into patient responses in a real-life clinical setting with two types patients (i.e., established and new patients), which we analyze separately. Since our real-life data only includes data on booked appointments (i.e., it is a BAD), we first use the imputation methods outlined in the previous sections to construct I-BAD, using 6.4% as the lost patient rate, estimated from the demand universe dataset (i.e., DUD) available in the clinical

setting we study.

Our data comes from a specialty unit that provides highly specialized, destination medicine to patients, and hence, our assumption on the PLWBA and C/RS/NS observations only occurring due to patient's WtW is highly appropriate for most patients in this context. In other settings, we acknowledge the fact that WtW may not be the only reason for PLWBA or C/RS/NS observations. We consider 48 time buckets per day that an appointment can be requested for both new and established patients.



(a) Probability Estimates for Established Patients

(b) Probability Estimates for New Patients

Figure 2.9: Probability Estimates from I-BAD

Figures 2.9(a) and 2.9(b) demonstrate the delay-based realization probabilities obtained with the survival and rank-based choice models for established and new patients, respectively. We plot the realization probabilities on a backdrop of the 95% CIs obtained from the fraction of realized appointments with each level of delay observed from I-BAD. For new patients, the number of p_k estimates outside the CIs results in p-values of 0.7300 and 0.1724 for the WtW distributions estimated from the survival model and rank-based choice models, respectively, failing to reject the null

hypothesis that the obtained WtW distributions effectively reflect patient response. Both models yield a p-value of 0.0810 for established patients, again failing to reject the null hypothesis.

Given the clinic under study is one that provides very specialized care to patients in a destination medicine setting, it is probably not very surprising that we do not observe a high sensitivity to appointment delays. For both patient types, we observe a steep decline in realization probabilities as delays increase from zero to five business days (i.e., one-week delay).

However, a side-by-side comparison of the realization probabilities show that new patients tend to be more "patient" toward experienced appointment delays (i.e., higher realization probabilities for almost all delay values, k). This can be explained by the fact that new patients are probably expecting longer delays to start treatment in a new, highly reputable healthcare facility.

A comparison of the realization probabilities for new and established patients show a notable difference in their responses; established patients show relatively stable realization probabilities for delay values of more than a week, whereas we see that the realization probabilities degrade consistently as delay values increase. This observation can probably be explained by the fundamental healthcare trade-off between speed and service quality which in our case corresponds to access delay and continuity of care (Saultz, 2003; Liu *et al.*, 2017). Considering that established patients are already in the system and have an ongoing relationship with their physician or the institution, they may have a relatively stable response to delays ranging from 10 to 60 business days due to the nature of their ongoing, long-term, periodic medical care needs provided by a provider who is familiar with their condition or stay within a healthcare system that they have prior experience with.

Due to the way that the two models work, the realization probabilities estimated

by the survival model for new patients resemble a step-wise function, resulting in similar realization values for delays in certain ranges, but decreasing as delays increase beyond a certain level. In comparison, the rank-based choice model for new patients depicts almost strictly decreasing behavior for the new patients. For established patients, the two models provide almost the same "stable" characterization. While both models provide WtW characterizations that are statistically significant, the choice of which one to use may be a function of the specific medical context, which may make one model more appropriate than the other.

We can discuss possible ways of using our results to improve patient access to care. There are many ways that such a characterization of patients WtW can be used to reduce waste (i.e., no-shows, cancellations, etc.) and improve patient outcomes and provider satisfaction. A better understanding of patient response can lead to improved strategies to offer appointments to patients, considering various system parameters such as capacity and slot availability as well as how a given patient may react to an offered appointment delay. Such efforts may effectively reduce no-shows and cancellations, and improve patient access times, rather than simply monitoring these performance measures.

One relatively direct way is identifying a good strategy to use overbooking while scheduling patients. Overbooking basically means assigning a patient to an appointment slot that is already occupied by another patient. It is commonly used in practice for accommodating patients who are in urgent need of care, as well as reducing the negative effects of patient no-shows or late cancellations and increase slot utilization. Unfortunately, overbooking generally leads to higher direct wait experienced by patients and/or overtime for providers. Our realization probability estimates (which are obtained as a function of patient type and appointment delay) can be used to determine which appointment slot should be considered as a good candidate slot to

overbook a patient if overbooking is unavoidable. Suppose two patients are scheduled to different appointment slots on the same clinic day. Further suppose that one of the patients requested the appointment 3 days prior to the appointment date while the second one experiences 10 days' appointment delay. Our analyses suggest that an appointment with 3 day delay is more likely to be realized than an appointment with 10 day delay. Therefore, the slot that the patient is scheduled with 10 day delay is a better overbook candidate compared to one with 3 day delay.

A similar decision can be made based on the patient type that is scheduled to an appointment slot. An overbooking policy can be designed by considering these different realization probabilities of each appointment slot based on the delay experienced by patients who are scheduled in those slots and types of these patients. Additionally, we can estimate the expected load in the system from those probabilities and amount of overbooking that should be allowed can be determined based on the capacity required to serve that demand load.

2.7 Conclusions

Understanding patient sensitivity to appointment delays is crucial to develop policies to effectively utilize the available care capacity. While patient response is a critical component to understand and quantify to improve access to care, the data that represent patients' sensitivity to appointment delays are not collected at the time of the appointment request and mostly not easy to collect from the patients. Therefore, direct estimation of patient WtW is not possible with the available data that we can access. However, the transactional appointment data collected let us observe each offered appointment with their delay and the patient's reaction to that particular appointment. While these data do not directly give us the probability that an appointment being booked and subsequently fulfilled by the patient, it allows us to

observe the intervals that patient's WtW belongs to.

We focus on developing methods that utilize transactional appointment data to be used under different healthcare settings and for various patient populations. Therefore, instead of developing models that are specifically designed for the available data that we have from a single institution and a single specialty clinic, we consider developing good models that can be adapted to different settings as long as transactional appointment data is available. Additionally, we assume each patient population has an underlying inherited WtW distribution and focus on characterizing this distribution via statistical methods.

Hence, our objective becomes developing statistical models to characterize patient WtW as the probability of an appointment with a certain delay being realized. While the collected data allow us gain insights into patient WtW, we do not have full access on FED since it is incomplete in the sense that the data on offered delays for PLWBA is not collected. Even though we do not have access to FED, the two datasets that we can access, BAD and DUD, let us represent PLWBA and append an estimate of PLWBA to BAD via imputation with random sampling to obtain an I-BAD.

We develop two non-parametric statistical model, namely survival model and rank-based choice model, to observe the effect of offered appointment delay on patients' decision on fulfilling a particular appointment. For each model, we first discuss the modeling approach that we take if FED is available. Then separately for each model, we discuss the possible imputation strategies to obtain I-BAD and introduce how we modify our models to utilize I-BAD instead of FED.

We start with testing our proposed models on simulated datasets, sFED and sBAD, that are generated by simulating a patient flow that resembles the process in the setting that we focus on. Using simulated data helps us observe FED and test the effect of amount of data on the estimations. Additionally, since we use a set of

probabilities to represent WtW in simulation model, it allows us to test effectiveness of proposed models by calculating the error of the estimates compared the "true" realization probabilities that we use as inputs in our simulation model.

By conducting the analyses on simulated datasets, we make numerous observations on performance of our models. By calculating error metrics compared to "true" realization probabilities, we observe that proposed models perform better than a baseline model that directly estimates probability of realizing an appointment as a function of appointment delay without considering any underlying WtW distribution. Our observations show that continuous collection of data is effectively improving estimates. Additionally, our results suggest that collection of FED is crucial to obtain probability estimates that are close to reality. However, if collection of FED is not possible, time-consuming, or costly, it is observed that using I-BAD instead of BAD leads to results close to reality and close to estimates that are obtained from using FED. We then develop a goodness-of-fit test procedure to assess how well our estimates represent WtW distribution. Our analyses show that the estimates generated from both models on sFED and I-sBAD are effectively representing WtW distribution.

From various analyses of simulated data, we conclude that both models are useful in estimating WtW distribution and generating estimates that are close to "true" realization probabilities. We then employ our models on a real BAD and DUD for multiple outpatient clinics in a specialty unit that includes data on two different patient types, new patients and established patients. We use the data on those two different patient types to observe differences in patient responses. Our observations suggest that established patients show more stable responses to delay values compared to new patients which supports the trade-off between access to care and continuity of care.

Our research only focuses on WtW aspect of the patient behavior. In real life,

patient behavior is complex and there are several patient specific aspects that can impact patients' decisions. While understanding the underlying components of the patient behavior is critical to improve patient flow, characterizing patient behavior completely requires substantial amount of data which are mostly not collected and possibly not easy to capture. Thus, we focus only on WtW behavior which is one of the most critical patient behavior that has a substantial effect on patient experience in accessing healthcare resources and this behavior can be reliably estimated from the available data.

In the present study, we present effective statistical models to estimate patient response from transactional appointment data. Our analyses show that the probability of realizing an appointment can be parametrized which let us incorporate patient WtW into developing prioritization policies while assigning patients to appointments to improve patient access. In Section 2.2, our analyses on average appointment delays show that established patients have relatively lower average delays than new patients which signals a prioritization scheme that is used to schedule patients to appointment slots. While we focus on two patient types in this study, generally hospitals use multiple patient types that might be based on urgency, referral type (internal or external), location (in-state, out-of-state, international, etc.), or based on the visit type as the dataset that we consider (new, established, etc.). One possible future direction can be extending the prioritization scheme that we observe to cover various patient types considering patient WtW and develop an access policy that assigns capacity to patients' based on their priority and WtW.

Chapter 3

TIME WINDOWS POLICY TO IMPROVE PATIENT ACCESS

3.1 Introduction

Providing timely access is one of the main indicators of quality of healthcare delivery. In the recent years, increasing gap between the need for healthcare resources and available capacity leads longer waiting times experienced by the patients ((Merrit-Hawkins, 2017)) when they request appointments from outpatient clinics. Long waiting times not only lead to lower patient satisfaction but also have potential negative impact on patient safety and healthcare outcomes.

One may think of an obvious way of managing the mismatch between the demand and supply is increasing the available capacity. However, increasing the number of resources in healthcare is costly and the supply for the resources is limited, especially if the resource is a human component. One way to improve system performance and healthcare delivery is allocating the available capacity based on patient needs and expectations. This task goes beyond simply scheduling the patients to the available appointment slots since it requires an approach that captures multiple components of the system in a more comprehensive manner and considers the unique characteristics of different healthcare setting to improve patient experience rather than simply treating patients identically. It is a challenging task to identify how to allocate the available capacity due to heterogeneity of the patient characteristics, their priorities, and service needs as well as uncertainties involved in patient arrivals and service time durations. An ideal access policy should take differences in patient characteristics into account and satisfy patient needs accordingly. Our objective is developing an ac-

cess protocol by considering patient priorities and their care needs to provide timely care and improve patient experience. While access can be defined for all sorts of patient-provider interaction, our focus is on improving patient access to outpatient appointments in this study.

Access is a complex concept and definition of it varies among institutions and specialties (see e.g.,Levesque *et al.* (2013)) based on clinic dependent characteristics and goals. Additionally, differences in patient characteristics and care needs result in different access requirements for each patient. To this end, we redefine improving patient access and determine our goal as using the existing transactional data to meet the *right* patients with *right* provider with *right* access delay. This definition contains not only prioritizing the *right* patients but also doing it under the concept of *right* access delay which is characterized based on patients' delay expectations and their urgencies.

3.1.1 Outpatient Scheduling

Outpatient appointment scheduling has been extensively studied in the literature. The appointment scheduling literature can be classified into two groups according to the wait type the patients experience; direct wait and indirect wait. In direct wait, the focus is on the waiting time spent during the day of appointment after the patient arrives while indirect wait considers the virtual waiting that patient experiences after they are scheduled to an appointment in a future day on a booking horizon. This way of scheduling is named as advanced scheduling in the literature. We focus on advanced scheduling systems and indirect waiting times in the context of improving patient access in this study. Gupta and Denton (2008) and Cayirli and Veral (2003) provide an extensive literature review on the appointment scheduling systems in healthcare. We also refer readers to Ahmadi-Javid *et al.* (2017) for detailed review on models on

outpatient scheduling systems.

We focus on a system where patients call an appointment office to schedule an appointment in a specialized clinic of a large hospital system. The appointment office agent obtains patient information at the time of the request and offers an available appointment slot with a provider based on patient's medical needs and characteristics. If the patient books the appointment, the patients' access delay (*indirect*) is calculated in business days and we refer this metric as time to appointment (TtA). In the system we consider, the appointment calendar of each provider consist of certain number of fixed-length appointment slots on each clinical day on a limited booking horizon. When a patient is scheduled to an appointment slot, we assume that both the patient and the physician are punctual so that the *direct* waiting times are negligible.

There are several studies in literature that present different approaches for reducing indirect wait. Even though it is a well studied concept in the literature, the studies on settings with patients from multiple priority classes are limited. Most of the research in outpatient scheduling either assume the patients are homogeneous in terms of the service needs or there are priority classes based on urgency (see, e.g., Klassen and Rohleder, 2004; Wang and Gupta, 2011; Ayvaz and Huh, 2010; Truong, 2015), encounter type (Patrick *et al.*, 2008), service time characteristics (Klassen and Rohleder, 1996; Cayirli *et al.*, 2008) or patient location and diagnosis characteristics Kazemian *et al.* (2017).

In multipriority setting, the literature mainly addresses allocation of available limited capacity to patients from different priority classes. The study of Patrick and Puterman (2007) proposes a simple way to allocate the available capacity to patients of higher priorities for a two day planning problem and use overtime when the patients cannot be scheduled. Patrick *et al.* (2008), is one the leading studies in the appointment scheduling area where authors consider an advanced scheduling system

with multipriority patients in a diagnostic imaging facility. They use a prioritization scheme based on patients' encounter type which are emergency, inpatient, and outpatient. They model the problem as an infinite horizon Markov Decision Process (MDP), and solve it with an LP-based Approximate Dynamic Programming (ADP) model. Similar to Patrick *et al.* (2008), in Saure *et al.* (2012) authors model the problem as an infinite horizon MDP for multi-priority patients by extending the model with including multi stage appointments. In the study, the patient priority classes refer to urgency levels, cancer site, and treatment intent with different waiting time targets. Another study that focuses on surgical scheduling is Astaraky and Patrick (2015) where authors develop a model for both operating room and recovery bed scheduling under the presence of multiple priority classes again the priority classes represent urgency of the patients.

While these studies show useful strategies to improve patient access, there is a clear distinction between our work and this body of work in outpatient scheduling. Unlike many studies in literature, we do not focus on giving slot-based decisions, instead, we focus on developing higher level prioritization policies as set of rules that can help appointment office agents to offer multiple appointment slot options to patients based on patients' needs and specific characteristics. Our encounters health systems show us that appointment office agents act as a gatekeeper to the health system since they are the ones that patients directly interact. Therefore, an effective access policy should be clear and easy to implement by the appointment office agents. Additionally, while allocating the capacity, the policy should provide options for the patients on appointment days so that patients can choose among them based on their personal schedules.

3.1.2 Patient Behavior

Mainstream outpatient appointment scheduling literature directly focuses on controlling the available capacity and allocating it to the patients from different priority classes with the assumption that the patient demand is independent of the scheduling policies that the institutions use. However, a stream of research in literature suggests that patients show aversion to prolonged waiting times (see, e.g., Gallucci *et al.*, 2005; Liu *et al.*, 2010; Norris *et al.*, 2014; Osadchiy and Kc, 2017), which indicates that offered appointment delay can affect patients' appointment booking and fulfilling behavior. We refer this behavior as willingness to wait (WtW), which is the maximum access delay that a patient can tolerate to wait. Therefore, if the offered appointment delay is more than patient's WtW, patient does not to fulfill the offered appointment and abandon the system without booking.

WtW is a similar to the willingness-to-pay concept in revenue (yield) management. Yield management is one of the most popular techniques that is successfully implemented in for profit service industry, such as airline companies and hotels, to manage the available limited capacity (see, e.g., Belobaba, 1987; Smith *et al.*, 1992) through segmenting the market based on customers' willingness-to-pay. While there are obvious differences between the systems that revenue management is successful in and healthcare systems, some of the insights from revenue management can be used to address the issues in access context. We inspire from the idea of segmenting the market in yield management and focus on developing access protocols that use WtW to divide patient population in segments by offering the patients different delay values.

3.1.3 Time Window Based Policy

This insights from literature shows us observed delay is an important factor in patient decisions and it is possible to use appointment delays to control patient demand to address the mismatch between available capacity and demand. Our goal is to bring a new perspective to studies in outpatient scheduling area by focusing on how to control the demand to improve patient access to care by impacting appointment booking decisions of non-urgent, lower priority patients. Our policy not only focuses on serving higher priority patients early but also considers serving higher proportion of them. Therefore, we can identify our approach as matching available clinical capacity and patient demand subject to patient behavior.

To this end, we focus on policies that considers patients' inherent aversion to waiting times and use this behavior to provide diverse care for patients from different priorities. In this study, we introduce a time-windows based access protocol (TWP) to improve patient access in multipriority patient environment. Instead of allocating the capacity in terms of appointment slots, the policy allocates the available capacity as time intervals on booking calendar that each priority can be scheduled. By strategically delaying the lower priority patients, goal of time window based policy is to dilute the arrivals from lower priority levels by inducing a higher rate of abandonments. Our goal is to determine the optimal time windows for each priority class to be scheduled in considering the patient abandonments due to appointment delay to improve patient access.

A successful implementation of time window based access protocol lies in clearly identifying patient priorities along with patients' sensitivity to waiting times. Notice that the reason why time window based policy is an effective prioritization scheme is due to inherent patient WtW and dilution of demand as a result of it. Therefore, it

is critical to understand patients' WtW and quantify this behavior from the available data to successfully implement a time window based policy. In Chapter 2, two alternative statistical methods to estimate patients' response to appointment delays from transactional appointment data are provided. This study shows us that the parameters associated with WtW can be effectively estimated from the available data, hence, can be used in developing an effective access protocols. We focus on a system where patient priorities are easy to identify. In particular, we assume that there is a set of guidelines that can be used to assign priority classes to patients at the time of their request for an appointment with a provider. Additionally, we assume that complete characterization of patient abandonment behavior is possible through analyzing the transactional appointment data.

Considering that the concept of access and perception of improving the access are different for each institution and each specialty, there is no single model that can cover all those cases. Therefore, we focus on a specific setting that is common in healthcare institutions where decision makers have a clear hierarchical preference for higher priority patients over patients from lower priorities. Demand from each patient class is satisfied in a hierarchical manner considering the available capacity. Since patient behavior is an undeniable component of the patient satisfaction, we take it into account while determining the time windows for each priority class. Therefore, the access policy that we develop targets prioritizing patients not only serving higher priority patients earlier but also serving higher priority patients with fewer abandonments.

In our study, in addition to regular demand, we also consider overbook decisions. In the literature, overbooking is mostly used to prevent negative effects of patient no-shows and cancellations (see, e.g., LaGanga and Lawrence, 2007; Zacharias and Pinedo, 2014; Liu and Ziya, 2014; Parizi and Ghate, 2016). However, in the case of less

than expected no-shows and cancellations case, overbooks can lead to increased direct waiting times which lead patient dissatisfaction or it can cause overtime for medical providers which is costly and leads lower provider satisfaction. Unlike these studies, we do not employ overbooks to avoid unutilized appointments due to cancellations and no-shows since time window based policy already makes the assignment decisions considering the estimated dilution of patient demand.

Our main target while making overbooking decisions is to consider the trade-off between making a patient wait longer and overbooking the patient at an earlier appointment slot. Overbooks can help us to provide a certain level of service when the regular available capacity is insufficient to provide the targeted service level. We provide a set of solutions to decision makers under alternative overbook capacities to help them decide on the level of capacity that should be used o reach certain average TtA and service level targets. Additionally, it helps us to capture the trade-off between increasing the capacity and increasing the performance measures.

Determining the optimal time windows on a calendar system is a challenging task. The literature on outpatient scheduling shows that advanced scheduling systems tend to fail to "curse of dimensionalty" due to necessity to keep track of full appointment calendar. In order to avoid this consequence, we consider the calendar as a collection of uncapacitated bins where each bin represents the time window for a certain priority class. Uncapacitated bin representation along with the assumption that each patient is equally likely scheduled to any day on their time windows help us to characterize the optimal time windows for each priority class based on demand load, available regular and overbook capacity, and patient behavior.

3.1.4 Common Policies for Patient Prioritization

One family of policies commonly used in practice to prioritize patients is "template type" policies. This family of policies uses appointment templates that are designed in a way that each appointment slot is designated for a specific class of patients. This way, the policy limits the number of slots that can be allocated to lower priority patients and prioritizes higher priorities by allowing them access to a higher portion of the capacity. Even though this family of policies is commonly used in practice, it requires an accurate estimation of the traffic to allocate the slots between multiple patient classes. Since this family of policies is strict in reserving the capacity to patients, if they are not managed correctly, they can lead to low utilization levels while result in poor access levels for lower priority patients at the same time. Additionally, this type of policies are not easy to manage and adjust to the changes in demand mix since it does not have any dynamic features.

A less restricted way of protecting appointment slots is to use protection level policies that protect some portion of capacity from lower priority. The policy mainly does not allow the capacity to be used for lower priority patients if the available capacity is below a certain protection level. In Patrick and Puterman (2007), a capacity protection policy is proposed for higher priority patients each day and certain capacity is reserved in the subsequent day for carry-over demand of lower priority patients, and overtime used when those reserved capacities do not satisfy the incoming demand. The study of Ayvaz and Huh (2010) studies two patient case where the urgent patient demand is lost when it is not satisfied instantly, the authors use a dynamic programming formulation. Simple heuristics that reserve capacity for urgent patients is studied by Patrick and Puterman (2007) for the same setting. A dynamic allocation problem in a general setting on job processing is studied by Erdelyi and

Topaloglu (2009), where a model that reserves a certain amount of capacity in each day for jobs from different priorities is presented and a sample-path based solution methodology is discussed.

In our study, unlike the above mentioned families of policies, instead of reserving capacity in terms of appointment slots, we (weakly) allocate capacity by restricting the time intervals that each priority class can be scheduled in. One nice feature of using such a policy is that, since it does not restrict the capacity for certain patient types unlike to a template style calendar, it may provide better utilization levels due its flexibility. Additionally, since it allocates a certain time range that patients can be assigned to, multiple appointment slot alternatives can be offered to the patients.

A study that uses an approach that resembles our time windows is Kazemian *et al.* (2017), where the authors consider an appointment system where clinical and surgical appointments are scheduled in coordination in a multipriority patient- multiple provider environment. The authors suggest various alternative policies that include the delay target-dependent time windows similar to the ones suggested in Patrick *et al.* (2008), and evaluate the performance of those policies by using simulation. Then, the most effective policy in terms of average operating room overtime is fine-tuned through investigation. In our study, unlike Kazemian *et al.* (2017), we provide the optimal time windows that patients can be scheduled into considering hierarchical preference. Additionally, our study generates the time windows not only considering the targets but also acknowledging patients' sensitivity to appointment delays.

We conduct various simulation experiments under different arrival and WtW cases to test the performance of the time window based policy as a comparison to the policies that we mention above. Our extensive analyses show that time window based policy is effectively prioritizing patients by assigning lower priority patients to appointment slots that are further into the future, and diluting higher proportion

of the demand from lower priority patients. We observe that both time window based policy and protection level policy are effective policies to improve patient access without resulting in high number of overbooks. We then compare these two policies from an application point of view, and discuss the advantages of using the time window based policy in terms of its flexibility and ease of implementation.

We provide a case study on our approach by using a real-life patient transactional data from a specialty clinic that provides destination medicine. We use the dataset to identify patient priority classes and estimate patient WtW from the data by using one of the statistical methods that are introduced in Chapter 2. We then calculate the optimal time windows for each priority class and utilize a simulation model to observe the amount of improvement in patient access due to implementing time windows based policy to prioritize patients. Our results suggest that time windows based policy results in lower access delay and fewer WtW-related abandonments from higher priorities while appropriately delaying the lower priority patients and diluting the demand of patients from those priority classes.

Following the real-life case study, we introduce the concept of "compromised prioritization" where the decision maker targets certain levels of performance measures for each priority class rather than strictly prioritizing higher level patients. We discuss the trade-off between not serving a lower priority class completely and compromising from the service level of higher priority class. By utilizing the real data, we generate managerial insights and provide an alternative set of solutions that can be used under different targets on performance measures and overbook capacities to help the decision maker.

The rest of the chapter is organized as follows. Section 3.2 presents the modeling framework of proposed scheduling policy that utilizes time windows to improve access in a setting with distinct priorities and hierarchical preferences of higher priority

classes over lower priorities. In Section 3.3, we describe our simulation model of the appointment system and compare the performance of time window based policy with other policies. In Section 3.4, we present our case study on a real dataset. Section 3.5 introduces compromised prioritization and the managerial insights that we generate. We conclude the chapter in Section 3.6.

3.2 Problem Description

We consider the problem of constructing time windows for each of N patient priority classes on a limited booking horizon of length T. We assume that it is possible to assign each patient to a priority class based on available patient traits, such as acuity or the urgency of the patient's medical needs where priority class 1 denotes the higher priority class. Following an appointment request from a patient from priority class n, the scheduling agent only search for an appointment within the specified time window for priority class n.

Previous studies on patient access and appointment scheduling show that keeping track of the full appointment calendar leads to "curse of dimensionality." In order to avoid "curse of dimensionality," we make a simplifying assumption that each time windows is represented as uncapacitated bins. At any day t, any arriving patient type n can be accommodated by their respective bin only. In this bin representation, we consider total used capacity within each bin rather than usage in separate days. However, this simplification is not representative enough since in the actual calendar whenever day changes, the first day of time window for patient type n becomes the last day of time window for patient type $n-1$. Therefore, we represent the process as when the day rolls (at the end of each day) each bin n(except the last one) receives a certain amount of used capacity from bin $n+1$ since the time windows are in sequential order in the actual calendar. This representation helps us to capture the

characteristics of a calendar and the ways days are rolled without keeping track of the full calendar.

In order to calculate the distribution of used capacity within each bin, we start with considering priority class N, the last priority class. Let T_N denote the length of time window for priority class N. Hence, the time windows policy assigns arrivals from priority class N into bin that represents the time period that starts with $(T+1)-T_N$th day from the current day until the end of day T.

To calculate the distribution of appointments in bin for priority class N, we keep track of the number of patients on a given day (we call this the "marked day"), from the first calendar day that the marked day enters the booking horizon (we refer to this calendar day as "day 1" without loss of generality) to the calendar day that the marked day becomes the first day in the time window. Notice that at the end of each day when the calendar is rolled, a calendar day with no used capacity appears at the end of booking horizon, we refer this day as the "marked day" in this representation. Assume that on a given day, the type N arrivals are distributed randomly (i.e., equally likely to any day within the time window and the number of type N arrivals on any day are distributed Poisson with rate λ_N. Let $W_N(1)$ denote the random number of type N patients received on day 1, and distributed equally among the days in time window N, where the last of these days is our marked day. In Figure 3.1 the marked day is shown with the diagonal hatch pattern with type N time window length $T_N = 6$. Since the $W_N(1)$ are distributed equally among the T_N days in the time window, the number of day 1 patients who are scheduled on our marked day (denoted as $Y_N(1)$) can be modeled as the sum of $W_N(1)$ independent, identically distributed Bernoulli variables, $I_1(1), I_2(1), \ldots, I_{W_N(1)}(1)$ with success probability, $p = 1/T_N$. On day 2, this marked day gets another $Y_N(2)$ patients scheduled on it that is equal to the sum of $W_N(2)$ independent, identically

distributed Bernoulli variables, $I_1(2), I_2(2), \ldots, I_{W_N(2)}(2)$, again with success probability $p = 1/T_N$ when the day is rolled. Hence, the number of patients scheduled on our marked day by calendar day T_N (i.e., $S_N(T_N) = \sum_{t=1}^{T_N} Y_N(t)$) can be modeled by the following probability generating function (since the type N arrivals in subsequent days are independent Poisson variables with rate λ_N:

$$\begin{aligned}
\pi_{S_N(T_N)}(u) = \prod_{t=1}^{T_N} \pi_{Y_N(t)}(u) &= \prod_{t=1}^{T_N} \pi_{W_N(t)}(\pi_{I_1(t)}(u)) \\
&= \prod_{t=1}^{T_N} \pi_{W_N(t)}\left(1 - p + pu\right) \\
&= \prod_{t=1}^{T_N} e^{-\lambda_N\left[1-\left(1-\frac{1}{T_N}+\frac{1}{T_N}u\right)\right]} = e^{-T_N\lambda_N\frac{1}{T_N}(1-u)} \\
&= e^{-\lambda_N(1-u)} \, . \qquad (3.1)
\end{aligned}$$

The probability generating function for $S_N(T_N)$ shows that the number of type N patients scheduled on the marked day is Poisson distributed with rate λ_N.

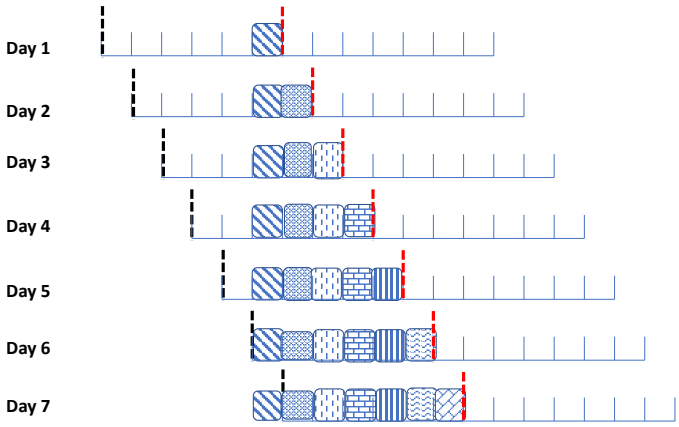


Figure 3.1: Days in the Time Windows

Using a similar argument, we can calculate the total number of patients scheduled in the time window for type N patients by calendar day T_N. The number of type N patients scheduled on the day after our marked day can be shown to be Poisson

distributed random variable with rate $\frac{T_N-1}{T_N}\lambda_N$. Furthermore, it is straightforward to show that this random variable is independent of $S_N(T_N)$, the number of patients scheduled on the marked day, since the number of type N patients arriving each day is a Poisson random variable, and the allocation of these Poisson arrivals on each day is determined randomly with probability $1/T_N$. Summing T_N independent Poisson random variables with rates $\lambda_N, \frac{T_N-1}{T_N}\lambda_N, \ldots, \frac{2}{T_N}\lambda_N, \frac{1}{T_N}\lambda_N$, we find that X_N, which represents the number of patients scheduled in T_N days to be distributed Poisson with rate $\frac{T_N+1}{2}\lambda_N$.

Since the number of patients in time window N includes the patients that are allocated in the first day of this period (i.e., marked day), the number of patients in the first day of the period is not independent of the total number of patients scheduled in the time window. As the calendar is rolled, the number of patients in this first day becomes part of the number of patients scheduled in the time window for priority class $N-1$. Therefore, it is important to analyze the conditional probability of the number of patients on the first day of the time window given the total number of patients in the time window. That is, for $x \geq s$

$$
\begin{aligned}
P(S_N(T_N) = s | X_N = x) &= \frac{P(S_N(T_N) = s, X_N = x)}{P(X_N = x)} && (3.2)\\
&= \frac{P(S_N(T_N) = s, X_N - S_N(T_N) = x - s)}{P(X_N = x)} && (3.3)\\
&= \frac{P(S_N(T_N) = s) P(X_N - S_N(T_N) = x - s)}{P(X_N = x)}\ , && (3.4)
\end{aligned}
$$

where the last equality is due to the fact that the number of patients scheduled on the first day of the time window is independent of the number of patients scheduled on the other days of the time window. Then, using the fact that the number of patients on the remaining $T_N - 1$ days of time window N on day T_N is Poisson with rate

$\frac{T_N-1}{2}\lambda_N$, we have

$$P(S_N = s|X_N = x) = \frac{\frac{e^{-\lambda_N}\lambda_N{}^s}{s!} \cdot \frac{e^{-\frac{T_N-1}{2}\lambda_N}\left(\frac{T_N-1}{2}\lambda_N\right)^{x-s}}{(x-s)!}}{\frac{e^{-\frac{T_N+1}{2}\lambda_N}\left(\frac{T_N+1}{2}\lambda_N\right)^{x}}{x!}} \tag{3.5}$$

$$= \binom{x}{s}\left(\frac{2}{T_N+1}\right)^s\left(1-\frac{2}{T_N+1}\right)^{x-s} \tag{3.6}$$

which equals to Binom(x, $2/(T_N+1)$).

To calculate the distribution of the number of patients in the time window for priority $N-1$, we note that each day in this time window has a number of priority N patients, independent and identically Poisson distributed with rate λ_N. Hence, using independence, we can show that the total number of patients in time window $N-1$ (which is of length T_{N-1}) is Poisson with rate $\left(\frac{T_{N-1}+1}{2}\right)\lambda_{N-1} + T_{N-1}\lambda_N$. Hence, as before, the rate is a function of the length of the time window used. The independence implied by random splitting of Poisson arrivals further imply that the total number of patients (of priority $N-1$ and N) in time window $N-1$ is independent of the number of patients in time window N.

It is possible to generalize this result for any time window for patient class $n \in \{1, 2, \ldots, N-1\}$. In general, the number of patients in priority time window n, X_n, is Poisson distributed with rate $\Lambda_n = \left(\frac{T_n+1}{2}\right)\lambda_n + T_n\sum_{k=n}^{N-1}\lambda_{k+1}$.

3.2.1 Modeling Abandonment

One of the main reasons that time window based policy is an effective method to improve patient access is that patients show aversion to wait. This time sensitivity leads to abandonments especially if the patients experience lengthy appointment delays. Due to this behavior, realization of an offered appointment becomes dependent on delay of the appointment and which patient class it is offered to. Chapter 2 provides two alternative methods that can be used to estimate the probability of real-

izing an appointment as a function of appointment delay. Those probabilities directly show patients' WtW for an appointment with k-day delay since an appointment is realized only when the patient's WtW $\geq k$. Therefore, we can obtain an estimate for $P(\text{WtW} \geq k)$ for each patient class by employing the models suggested in Chapter 2 on real patient data.

Under a time window based access protocol, we expect to observe more abandonment from lower priority classes compared to the higher priority classes since the lower priority patients are scheduled further into the future. This dynamic makes time windows an effective method in controlling the patient demand. Therefore, our objective is to include appointment delay dependent abandonment rates into our analyses to reflect the true effect of time windows on the system in this section.

Suppose that on a given day, the arrivals from priority N are distributed equally among the days in time window N. In addition to notation defined in sections above, we also define the beginning and end of time window N by day B_N and E_N, respectively. The number of patients on the first day of time window N (referred to as S_N) is distributed Poisson with rate $\frac{1}{E_N - B_N + 1}\lambda_N \sum_{k=B_N}^{E_N} P(\text{WtW} \geq k)$. Also suppose that WtW follows cumulative distribution function (cdf) F_n for patient priority class n. Hence, the distribution of S_N is Poisson with rate:

$$\begin{aligned} \Lambda_N &= \frac{1}{E_N - B_N + 1}\lambda_N \sum_{k=B_N}^{E_N} 1 - P(\text{WtW} \leq k-1), \\ &= \frac{1}{E_N - B_N + 1}\lambda_N \sum_{k=B_N}^{E_N} (1 - F_N(k-1)). \end{aligned}$$

Therefore, the distribution of S_N is a function of both B_N and E_N, the beginning and end dates of the time windows for priority N. For any priority level n, the calculation is the same where the number of class n patients on the first day of the time window for n will be Poisson with the same rate that depends on the beginning and end dates

of the time window for n.

Consider the number of patients on the first day of the booking horizon, which includes patients from all priority classes. Since the Poisson arrivals of patients from different priority levels are independent due to the fact that unlimited amount of overbooking is possible, the number of patients from each priority level on the first day S_1 are Poisson distributed with the above calculated rates. Assuming that for the first priority time window starts on day 1 (i.e., $B_1 = 1$), the total number of patients scheduled over the entire booking horizon are then Poisson (sum of independent Poisson variables) with rate

$$\sum_{n=1}^{N} \frac{1}{E_n - B_n + 1} \sum_{k=B_n}^{E_n} (1 - F_n(k-1))\lambda_n. \tag{3.7}$$

Note that $(1 - F_n(k))$ is equal to the probability of realizing an appointment with k day delay; this can be estimated from the available data for each priority class n and can be utilized in calculation of the realized arrival rates.

After including the abandonments due to appointment delays in our model, we now focus on obtaining the optimal time windows.

3.2.2 Strict Prioritization Model

We consider the setting that is described in Section 3.1 where there is a clear order on patient priorities to be served. Particularly, we prefer serving priority n patients over priority $l \in \{n+1, \ldots, N\}$ patients. For instance, under three priority classes case, we first focus on allocating the available capacity to priority 1 patients, the remaining capacity from priority 1 patients are allocated to priority 2 patients. Lastly, priority 3 fills the remaining slots.

In our model, we fix a minimum and a maximum time windows length, $T_{\min}$ and $T_{\max}$, respectively. These parameters help us obtain time windows that are suitable for

applying in real hospitals. Notice that putting a lower bound on time window length results in dilution from priority 1 patient demand which is not desirable. However, this minimum time window length ensures that a reasonable period on the booking calendar is allocated to patients and it reflects the nature of the system where patients abandon even though they are served within the earliest possible time windows. Additionally, setting an upper bound, $T_{\max}$, for the length of time windows helps us avoid observing high variances in appointment delays experienced by the patients from same priority class.

We also consider an upper limit on the expected number of overbooks, denoted by $\theta_{\max}$, which can be expressed as

$$\sum_{s_1=0}^{\infty} \frac{e^{-(\sum_{n=1}^{N}\Lambda_n)}(\sum_{n=1}^{N}\Lambda_n)^{s_1}}{s_1!} \max\{0, s_1 - C\} \leq \theta_{\max} \ , \tag{3.8}$$

where s_1 is the total number of patients on the first day of the booking horizon. Notice that $\theta_{\max}$ along with the total capacity C limits the total demand rate that can be served with the regular capacity and overbooks.

We define $\Lambda = (\sum_{n=1}^{N}\Lambda_n)$ and rewrite the expression in (3.8) as:

$$\Lambda + \sum_{s_1=0}^{C} \frac{e^{-\Lambda}\Lambda^{s_1}}{s_1!} (C - s_1) \leq C + \theta_{\max} \ . \tag{3.9}$$

From this, we say that there exists a Λ^* that satisfies (3.9) in equality which is the maximum rate that can be served with total regular and overbook capacity; we refer to this quantity as the effective capacity. In our model, our focus is on allocating this effective capacity among patients from different priority classes with clear hierarchical preference by setting the time windows while considering the fill rates for each patient class where fill rate is defined as the proportion of demand that is served.

For each patient class n with a time window that is defined as $[B_n, E_n]$, fill rate

of priority class n, denoted as β_n, is

$$\beta_n = \frac{\Lambda_n}{\lambda_n} = \frac{1}{E_n - B_n + 1} \sum_{k=B_n}^{E_n} (1 - F_n(k-1)) \ , \tag{3.10}$$

where we refer to Λ_n as the diluted demand of priority class n. For our case with N patient priority classes, we can write

$$\sum_{n=1}^{N} \beta_n \lambda_n \leq \Lambda^*. \tag{3.11}$$

The expression (3.11) is an inequality (rather than a strict equality) since β_n values for each patient class n belongs to a discrete set that is determined by the underlying WtW distribution with cdf F_n and possible $[B_n, E_n]$ values, which may not make use of the effective capacity, Λ^*.

Under strict prioritization, we start with determining the time window for priority 1 patients that maximizes the fill rate (minimizing the dilution) of those patients. After determining the time window and the resultant Λ_1^* for priority 1 patients, we continue with patients from the next priority level. Our model becomes setting time windows $[B_n, E_n]$ for each patient priority class $n \in \{1, \ldots, N\}$ where we solve the following model for each n separately in the order of priority

$$\max \quad \beta_n \tag{3.12}$$

s.t.

$$\beta_n = \frac{1}{E_n - B_n + 1} \sum_{k=B_n}^{E_n} (1 - F_n(k-1)) \ , \tag{3.13}$$

$$\sum_{s_1=0}^{\infty} \frac{e^{-(\sum_{i=1}^{n-1} \Lambda_i^* + \lambda_n \beta_n)} (\sum_{i=1}^{n-1} \Lambda_i^* + \lambda_n \beta_n)^{s_1}}{s_1!} \ \max\{0, s_1 - C\} \leq \theta_{\max}, \tag{3.14}$$

$$B_n \geq 1, \tag{3.15}$$

$$T \geq E_n \geq B_n + T_{\min} - 1, \tag{3.16}$$

$$T \geq B_n + T_{\max} - 1 \geq E_n \ , \tag{3.17}$$

$$B_n, E_n \in \mathbb{Z}^+ \ . \tag{3.18}$$

While solving the model for priority n patients, we fix the Λ_i^*, $i \in \{1, 2, \ldots, n-1\}$, values for patients from higher priority classes and continue in this manner until either the effective capacity is filled or a time window is set for all priority classes. Notice that since we follow a strict prioritization scheme, the capacity can be filled before a particular patient priority class and any classes below that are served. In that case, we conclude that desired service cannot be provided to those patient classes with the current available capacity.

It is trivial to see that the model assigns the time windows that result in minimum possible dilution to each patient class until inequality (3.14) is satisfied with equality or with minimum slack (since equality cannot be reached for some WtW distributions). Therefore, in any cases where we observe that the patient class n is not served with minimum possible dilution (in the earliest time window) that means the patient classes $l \in \{n+1, \ldots, N\}$ cannot be served with the available capacity.

Our solution approach starts with creating a dilution table which indicates the achievable dilution for each $[B, E]$ pair that are constrained by (3.15)-(3.17) where achievable dilution for a patient class n, for each $[B_n, E_n]$ pair can be denoted as $1 - \beta_n$. Note that we create the dilution table since we do not assume any specific functional form for $F_n(k)$, therefore, we need to enumerate the dilution table based on WtW distribution. However, it may be possible to denote a closed-form expression for the achievable dilution as a function of $[B_n, E_n]$ pairs under certain underlying WtW distribution cases. For instance, if the WtW is geometrically distributed with rate p, the cdf is $P(\text{WtW} \leq k) = 1 - (1-p)^k$. Hence, the diluted demand rate for patient class n can be written as

$$\frac{1}{E_N - B_N + 1}\lambda_N \sum_{k=B_N}^{E_N} (1-p)^{k-1}$$
$$= \frac{1}{E_n - B_n + 1}\left[(1-p_n)^{B_n-1}\right]\left(\frac{1-(1-p_n)^{E_n-B_n+1}}{p_n}\right)\lambda_n \ . \quad (3.19)$$

We present the pseudocode for the solution methodology for the strict prioritization case in Appendix B.1. In the next section, we study the performance of the obtained strict prioritization time windows on a realistic system and compare it to that of other policies that we have introduced in Section 3.1.

3.3 Numerical Experiments via Simulation

We develop a discrete event simulation model to assess the performance of the time windows approach in a realistic system. Simulation model helps us to obtain the level of multiple performance measures considering the uncertainty of appointment request arrivals from different priority classes and compare these performance measures with those obtained by alternative policies that can be used in an outpatient setting.

We analyze the performance of time window based policy in two steps. In the first step, we consider various parameter combinations that are artificially generated and observe insights into how time window based policy performs and whether it is more beneficial in improving patient access compared to other policies. In the second step, we conduct a case study on a real dataset from a real healthcare institution. In this case, we observe how time windows perform in a real life setting in comparison to current performance of the system.

3.3.1 Simulation Study on Generated Data

We model a patient flow with simulation that is similar to the appointment system that we observe. We focus on the two priority classes case to easily observe and

compare the performance measures. When a patient requests an appointment, an appointment is offered to a patient based on his priority class and the clinic's scheduling policy. We assume that the institution has an effective tool to assign each patient to a priority class at the time of the appointment request. After the appointment is offered to the patient, patient evaluates the delay against his WtW. We assume that each priority class has a separate inherent WtW distribution and lower priority patients tend to be more patient towards appointment delays. Note that this assumption reflects a worst case scenario for the performance of the TWP, the effectiveness of which is relatively limited when lower priority patients are less sensitive to delays and do not abandon the system easily. If the offered appointment delay satisfies the patient's expectations, he books the appointment and is assigned to the slot, and the TtA is recorded. Otherwise, patient abandons the system without booking and the event is recorded. In our simulation model, we assume that available number of regular appointment slots and overbooks slots are equal on each day of the booking horizon.

In the implementation of time windows based policy, we randomly assign any available regular slot to a patient within the appropriate time window to reflect our modeling assumption that the patient can be scheduled at any day during his time window. If there is no regular capacity available within his time window, patient is assigned to an overbook slot within his time window. If there is no available regular or overbook slots within the time window, the patient is assigned to first available slot on the rest of the booking calendar. Patients are only rejected when there is no regular and overbook capacity available on the booking calendar after the first day on their time windows.

We test the performance of time window based policy and three families of policies that are commonly used in the literature and in practice, template type policies (TP),

protection level policies (PLP), and first come first served (FCFS).

Under TP, certain appointment slots are strictly templated to only accommodate higher priority patients where a higher number of slots are typically made available to patients of higher priority since lower priority patients cannot use a certain portion of the slots. There is a nested allocation in TP where all appointment slots that can be used by priority class $n+1$ can also be used by class n. When an appointment request arrives, TP searches for an appointment slot that is templated for the requesting patient's priority class and schedules the patient to the first available such slot. We relax the strict rules of TP and generate another policy that allows lower priority patients to be scheduled to slots templated for higher priorities if the slot is not utilized five business days prior to the appointment date. We name this policy as template type policy utilizing unused slots (TP-U).

In designing the template for each week, we set the number of appointments to be templated to higher priority on each day as the value that sets Poisson cdf to 99% based on the arrival rate. If that value is higher than the capacity, we allow one appointment slot per day for lower priority patients. We set certain rules to avoid request rejections and utilize overbooks when needed. That is, if there is no available capacity that a patient class can use within the regular calendar considering the template, the patients are assigned to any overbook slot within the booking horizon. Request rejections only happen when there is no regular and overbook capacity on the booking horizon.

Under PLP, a certain level of weekly capacity is protected for higher priority classes by not allowing lower priority patients to be scheduled within a week if the available capacity is lower than a certain level. In PLP, we again take a similar approach as we did with template type policies to determine the number of appointment slots that need to be protected for higher priority patients on each week. For PLP,

we additionally put an upper bound on the delay that higher priority patients can experience to make it similar to time window policy for equal comparison and to trigger overbooks from this patient class. Instead of assigning the first available slot, we assign priority 1 patient to an available slot within 10 days of the appointment request. If there is no available regular slot within 10 days, patient is assigned to an overbook slot within 10 days if there is one available. Otherwise, patient is assigned to first available slot within the booking horizon. Priority 1 patients are only rejected if there is absolutely no regular and overbook capacity within the limited booking horizon. On the other hand, priority 2 patients are assigned to an appointment slot on any day of a week that has total number of available slots that are more than a determined protection level. If there is no week that satisfies this condition within the booking horizon, the patient is assigned to any available overbook slot within the booking horizon.

Lastly, we consider the FCFS policy as a benchmark to show the performance when no prioritization is used. In FCFS policy, we assign the first available slot for any patient that requests an appointment without considering their priorities. To make our comparison fair, in FCFS, we trigger overbooks for priority 1 patients with certain probabilities. These probabilities are obtained from the result of time window based policy. We calculate the proportion of priority 1 patients overbooked in our simulation on time window based policy and use these proportions in each case as the probabilities for FCFS policy to trigger the overbooks. If the priority 1 patient is decided to be overbooked, he is assigned to first available overbook slot on the booking horizon. We additionally utilize overbooks when regular calendar is completely booked for both patient classes.

In order to gain insight into the performances of alternative access policies under different system conditions, we generate a large variety of parameter combinations.

Additionally, we consider alternative overbook capacity levels as a percentage of regular capacity where fix regular capacity as 20 slots per business day and overbook slots are determined as 10%, 25% and 50% of the regular capacity (2, 5, and 10 overbook slots per day). Overbook slots can be considered as regular slots that an additional patient can be assigned to even when they are already utilized.

We only focus on overloaded systems where arriving total demand is higher than the total regular capacity and we assume that the demand is stationary. The arrival rate combinations used in the simulation model are presented in Table 3.1. We mark the arrival regimes as H, E, and L based on the combination between two patient types. H denotes the cases where priority 1 patients have higher demand compared to priority 2, L denotes the ones where priority 1 patients have lower demand compared to priority 2, and E denotes the ones where the demand from priority 1 and priority 2 is equal. We additionally the mark the regimes on a figure in Table 3.1 and show their positions with respect to $\lambda_1 + \lambda_2 = C$ line that is shown in the figure.

We consider three different patient populations in terms of WtW behavior. We can name these three populations as impatient (I), average (A), and patient (P) population. In population I, underlying WtW distribution is right-skewed where a larger portion of the population have lower WtW. Population A represents a population where more patients tend to have average values for WtW and very low and very high WtW is relatively rare, and population P represents the population where more patients have higher WtW values. We present the WtW distributions and their resulting realization probabilities in Figure 3.2.

We run the simulation for three years with one year of warm-up period, and use 100 replications. Since we are using multiple replications, we illustrate our results with boxplots.

For each case, we calculate five different performance measures; average TtA for

Regime ID	λ_1	λ_2
L1	4	20
L2	8	16
L3	5	25
L4	10	20
E1	12	12
E2	15	15
H1	20	4
H2	16	8
H3	25	5
H4	20	10

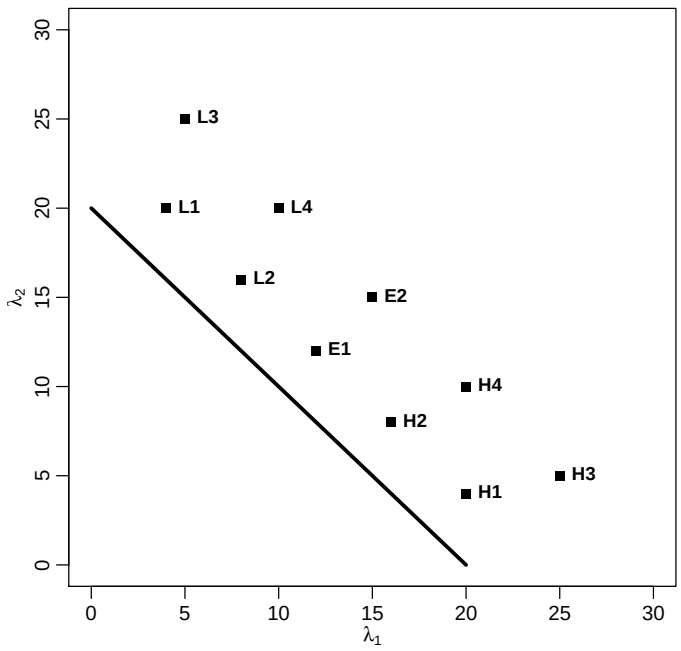
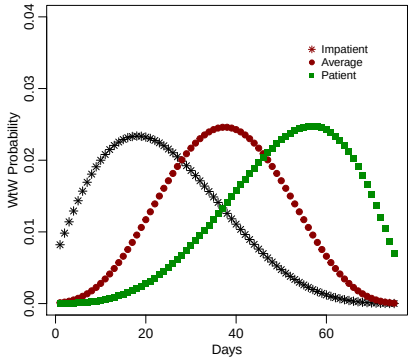


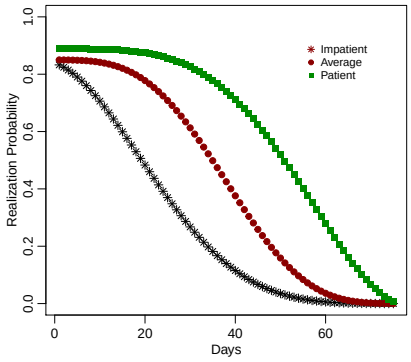
Table 3.1: Arrival Regimes

each priority class, fill rate (FR) for each priority class, percent rejected for each priority class, utilization of regular slots, and utilization of overbook slots for each policy. Since we have 90 (10 arrival regimes, 3 WtW cases, and 3 alternative overbook capacities) cases in our simulation model with five performance measures, we only report three representative cases to generate insights. In Figures 3.3 through 3.5, we illustrate the results for the patient population under WtW P with $\theta_{\max} = 5$ for arrival cases L1, E2, and H2, respectively. We do not observe any rejections in any of the cases, therefore, we do not report rejections. We also illustrate the results on the same arrival cases with I and A WtW groups in Appendix B.2.

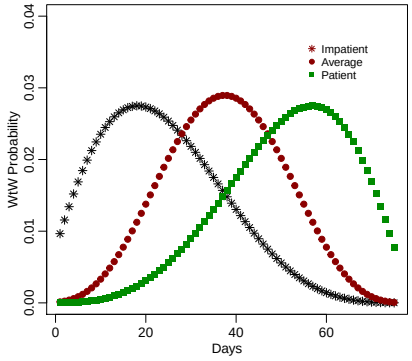
Under WtW cases P and A, we observe that TWP leads to significant improvement in performance measures compared to FCFS. However, WtW case I, almost all of the policies that we test in the simulation model, except TP, perform similarly since the



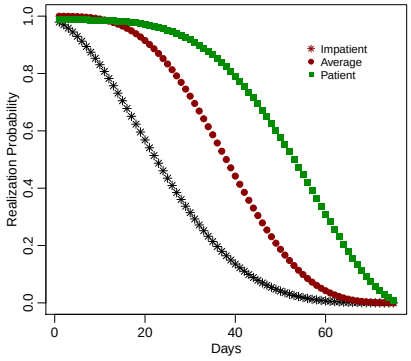
(a) WtW Cases-Priority 1



(b) Realization Prob.-Priority 1



(c) WtW Cases-Priority 2

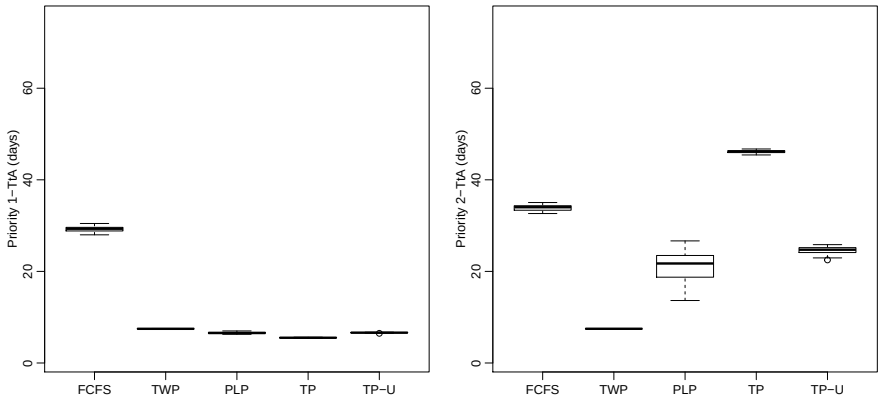


(d) Realization Prob.-Priority 2

Figure 3.2: WtW Parameters Used in Simulation Model

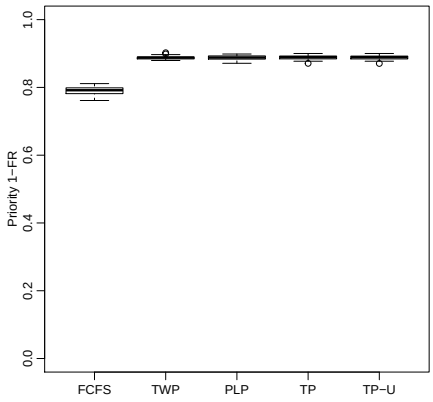
natural dilution in WtW case I leads to system not being congested. For instance, under WtW case I for the cases where $\lambda_1 + \lambda_2 = 24$, we rarely observe overbook slots being utilized, while we start to observe overbook slot utilization and differences in performance measures between the policies when $\lambda_1 + \lambda_2 = 30$.

Under WtW cases A and P, we observe that arrival regime H3 when $\theta_{\max} = 2$ is the only case where priority 1 patients are not served within 10 days under the policies that uses prioritization (the ones except FCFS). Under arrival regime H3,

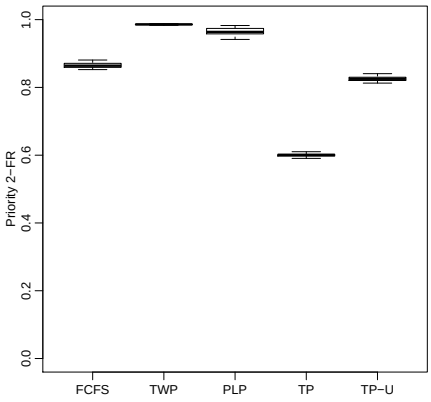


(a) Average TtA for Priority 1

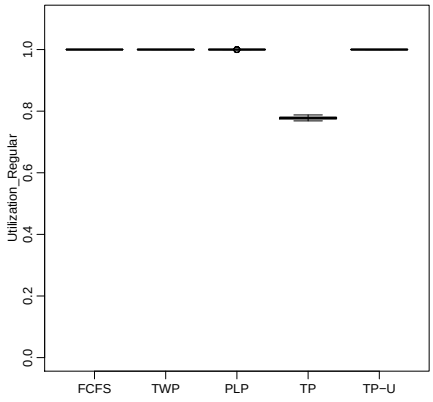
(b) Average TtA for Priority 2



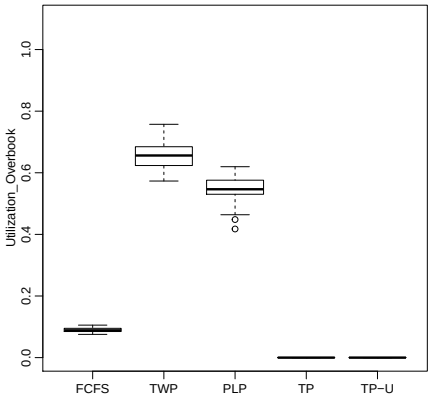
(c) Fill Rate for Priority 1



(d) Fill Rate for Priority 2

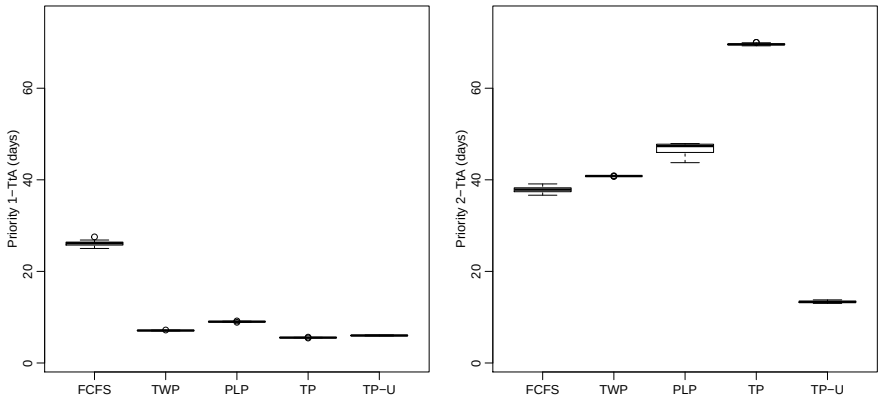


(e) Utilization of Regular Slots



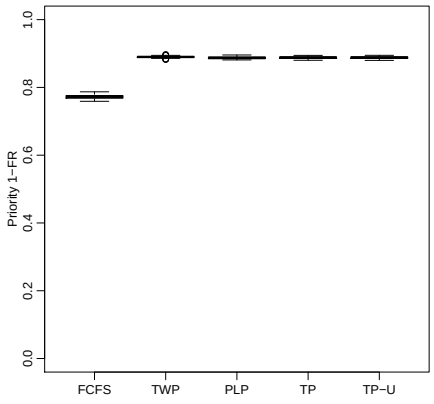
(f) Utilization of Overbook slots

Figure 3.3: Simulation Results for WtW case P, Arrival Case L1 ($\theta_{\max} = 5$)

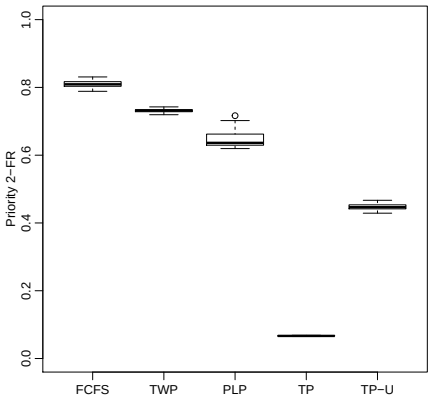


(a) Average TtA for Priority 1

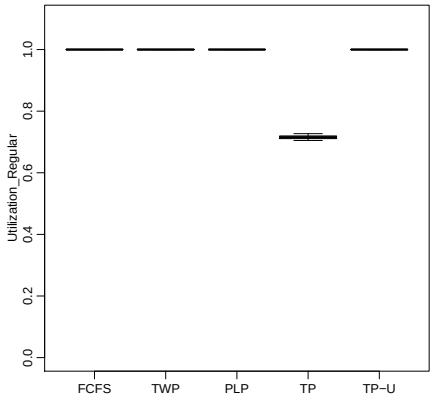
(b) Average TtA for Priority 2



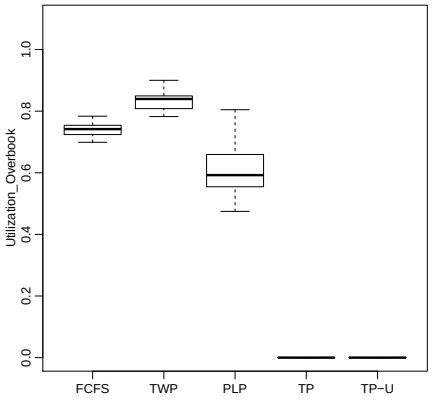
(c) Fill Rate for Priority 1



(d) Fill Rate for Priority 2

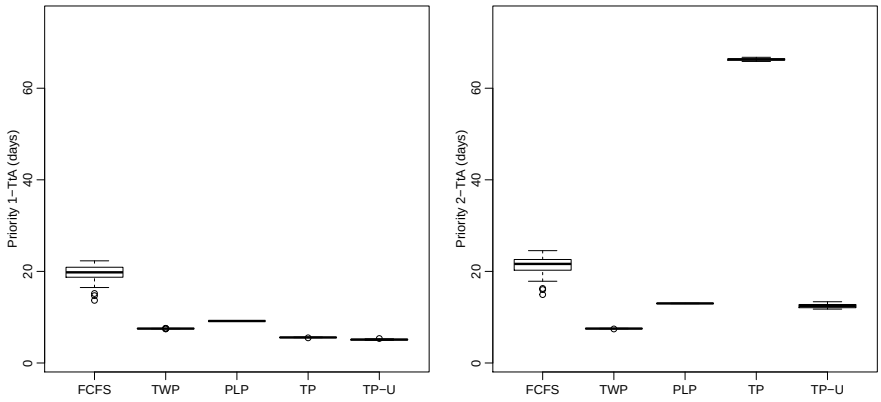


(e) Utilization of Regular Slots



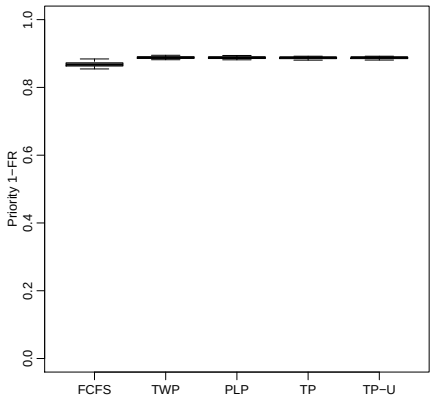
(f) Utilization of Overbook Slots

Figure 3.4: Simulation Results for WtW Case P, Arrival Case E2 ($\theta_{\max} = 5$)

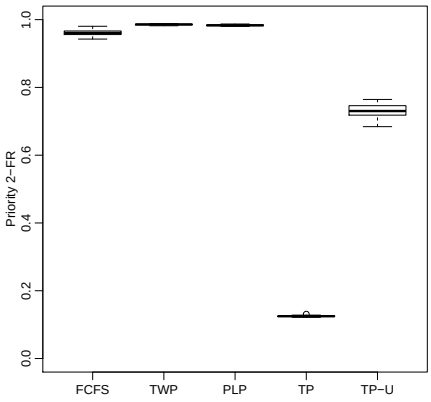


(a) Average TtA for Priority 1

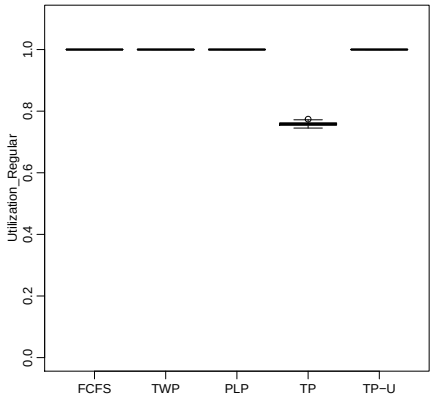
(b) Average TtA for Priority 2



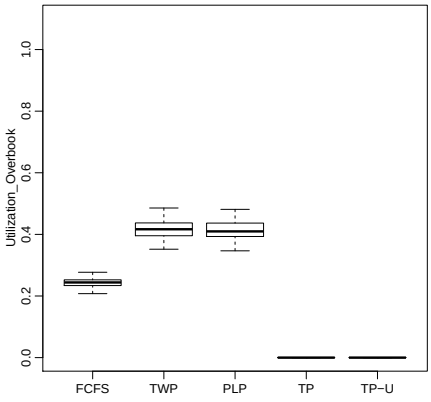
(c) Fill Rate for Priority 1



(d) Fill Rate for Priority 2



(e) Utilization of Regular Slots



(f) Utilization of Overbook Slots

Figure 3.5: Simulation Results for WtW Case P, Arrival Case H2 ($\theta_{\max} = 5$)

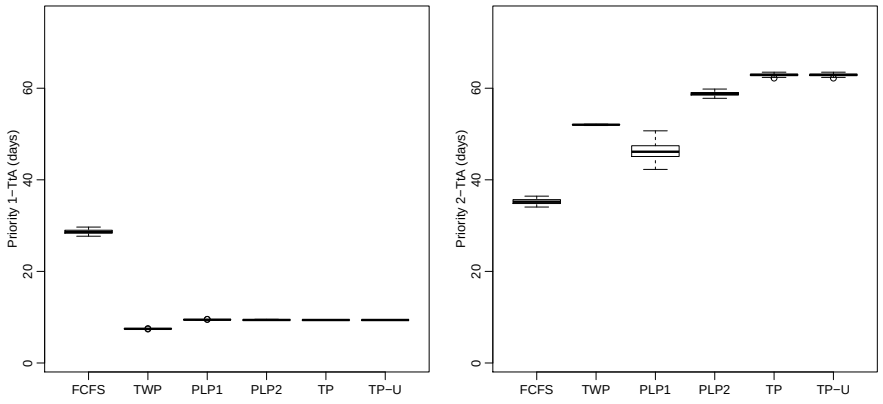
natural dilution from priority 1 patients under WtW cases A and P does not result in an arrival rate that can be satisfied with only two overbook slots. This result signals that decision maker should consider increasing the daily available capacity before using prioritization since even under the policies that uses strict prioritization, the highest priority patients cannot be served with minimum possible delay.

Generally speaking, we observe that TWP and PLP generate similar performance measures where TWP tends to utilize overbook slots more compared to PLP. While PLP is a good policy, it is not easy to maintain this policy since performance is highly dependent on the predetermined weekly protection levels. When we are determining the protection level for each case in our numerical experiments, we use the protection levels that result in the best performance measures since using a correct protection level is crucial for PLP's performance.

We consider a case to demonstrate the effect of protection levels on PLP's performance. We generate results for arrival case H3 with WtW case P and with $\theta_{\max} = 5$ in Figure 3.6. The results denoted as PLP2 depict the results of PLP with the best protection level while PLP1 is the result of PLP that protects one additional slot per week than the best protection level.

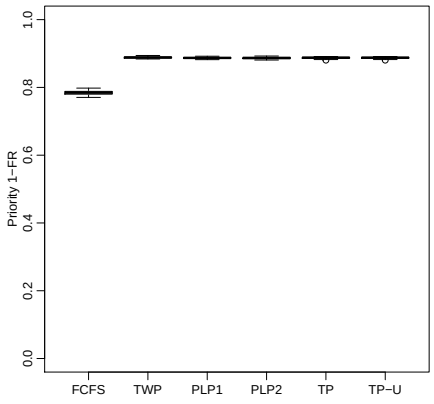
The results show that PLP1 results in much lower fill rate compared to PLP2 due to protecting one additional capacity for priority 1 patients. One interesting observation in here is that, since more patients are abandoning the system, the observed TtA(2) for priority 2 patients are lower while PLP2 utilizes more overbook slots. While PLP is an effective prioritization policy, even small changes in the predetermined protection level can lead to drastic changes in level of performance measures.

FCFS policy does not use patient priority classes, therefore, it results in higher average TtA(1) compared to other policies. On the other hand, since TP policy uses a strict template, and focusing on priority 1 patients, it results in higher average

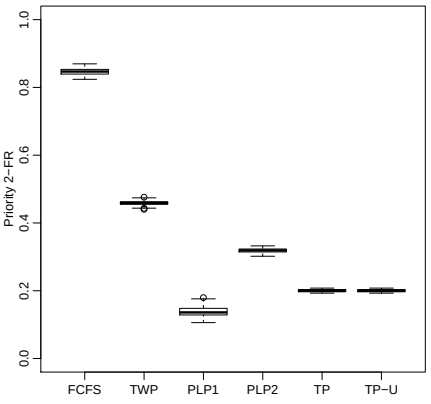


(a) Average TtA for Priority 1

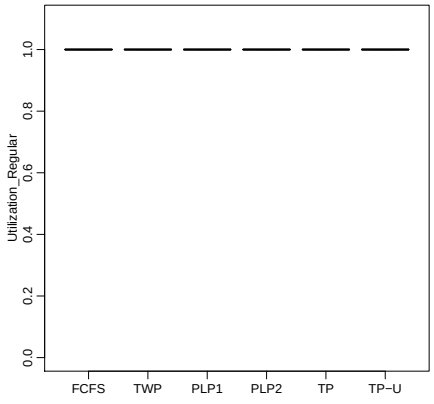
(b) Average TtA for Priority 2



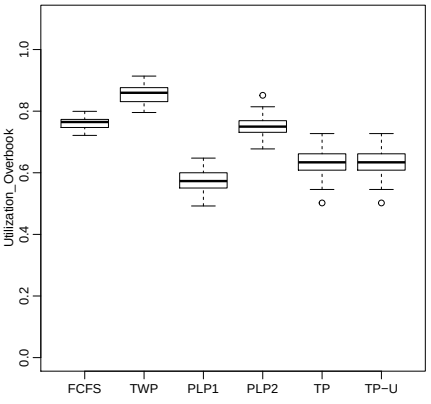
(c) Fill Rate for Priority 1



(d) Fill Rate for priority 2



(e) Utilization of Regular Slots



(f) Utilization of Overbook Slots

Figure 3.6: Simulation Results for WtW Case P, Arrival Case H3 ($\theta_{\max} = 5$)

TtA(2) along with a low utilization. This performance is expected from TP since when appointment requests from priority 1 patients are lower than the expected, the slots templated for priority 1 patients stay unutilized while priority 2 patients are experiencing unnecessary prolonged appointment delays. Since we are mainly focusing on heavily loaded cases, we observe that the utilization is close to 100% in all of the policies except TP. TP-U addresses this issue by allowing priority 2 patients to be scheduled to slots reserved for priority 1 patients if they are not used five days prior. Our results show that TP-U is performing similar to TWP and PLP in terms of average TtA measures and utilization. The main disadvantage of TP-U is that since it only allows slots to be used five days prior, it results in a higher variance in appointment delays that priority 2 patients experience. Even though the policy results in an average TtA(2) metric similar to TWP and PLP, it results in lower FR(2) due to the way it assigns patients to appointment slots.

Additionally, since H3 represents a case where the system is heavily loaded and the requests from priority 1 patients are higher than the capacity and much higher than the priority 2 requests, we observe that the utilization generated from all policies are equal to 100%, even the one from TP. We also observe that in this case, TP and TP-U results are identical since the high number of requests from priority 1 implies no slots assigned to priority 1 being available five days prior to appointment date.

3.4 Simulation Study on Real Data

So far in our analyses, we have used generated data to test the performance of our proposed policy in comparison to other policies that are commonly used in practice. Our simulation cases reflect the stylized case where higher priority patients are more sensitive to appointment delays and patients leave without booking an appointment if the offered appointment delay is more than their WtW. While a stylized simulation

model and generated data allow us to explore a variety of cases, our assumptions might not be correct in the real life. To this end, we utilize a dataset from a healthcare institution that provides destination medicine to patients.

Our data is from a pulmonary subspecialty; we use almost three years of data on booked appointments. The dataset includes appointment request dates, appointment dates, and referral type. The dataset also includes disposition codes, which indicate whether the booked appointment resulted in patient being seen by the provider, whether the appointment was canceled (C), rescheduled (RS), or patient did not show up to appointment (NS), and date of the cancellations and rescheduling. The data is specifically on booked consultation appointments of 60 minutes duration and consider three patient priority classes based on patient's referral type. Specifically, we use the appointment requests from three patient classes, which are new patients (NCON), external referrals (ECON), and internal referrals (ICON).

NCON patients new to the institution and trying to enter the system by self referral while ECON patients have a referral from their providers from another healthcare institution, and ICON patients are the ones that are referred to the subspecialty from another specialty within the institution. We use the referral based priorities as NCON patients are priority 1, and ICON patients are the last priority. The reason behind this prioritization scheme is that NCON patients are in the need specifically from the subspecialty while ICON patients are the ones who enter the system towards another clinic which they are visiting for their primary complaints, for instance from Internal Medicine specialty. Therefore, in terms of medical needs, getting an appointment from pulmonary clinic can be considered as an additional visit for ICON patients while it is the main destination of NCON and ECON patients. The distinction between NCON and ECON patients is due to the institution's goal of attracting new patients.

We analyze the available data under the stated three priority classes. We observe that 19% of the patient population that books an appointment at pulmonary clinic is priority 1 patients while it is 16% and 65% for priority 2 and priority 3 patients, respectively. The average offered appointment delays are observed as 47 days for priority 1 patients, 41 days for priority 2 patients, and 33 days for priority 3 patients. This observation shows that an appointment policy that utilizes prioritization is needed for improving the performance of the system since the current performance measures do not prioritize the patients in the desired order.

In our modeling efforts, we assume that the patients abandon the system if the offered delay is higher than their WtW. However, booked appointment data, it is not possible for us to observe this lost demand where we can only observe the booked appointments that are not realized, the ones with disposition code C/RS/NS. Due to only having access to booked appointment data, we only consider the dilution from demand due to patients C/RS/NS their booked appointments. In here, we assume that all C/RS/NS events occur due to patients' sensitivity to appointment delay. The main difference between patients abandoning the system at the time of the appointment request and C/RS/NS is that patients who C/RS/NS occupy an appointment slot until a certain point. In most cases, patients who NS or cancel on the day of appointment cause appointment slots being left unutilized since they occupy the slot until the very last moment.

In order to estimate patient WtW, we utilize the survival model that was introduced in Chapter 2. For each patient class, we separately use survival model on appointment requests from that class to estimate WtW. We limit the maximum delay that can be observed to 90 days and analyze the data accordingly. The estimated realization probabilities are shown in Figure 3.7.

Figure 3.7 shows that patient WtW is not in the same order with priority classes.

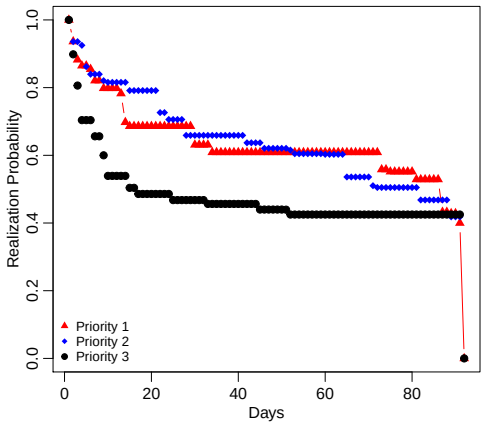


Figure 3.7: Estimated realization probabilities

We observe that priority 3 patients have higher sensitivity to waiting times, while priority 1 and priority 2 patients have higher WtW with similar characteristics. Since we focus on an institution that provides destination medicine, a large portion of the patients are usually not local patients. ICON patients are usually already visiting the hospital for another clinic due to their chief complaint at the time of the appointment request. For those patients, it is not desirable to wait for an additional appointment. Therefore, instead of waiting, these patients tend to cancel the booked appointments or try to reschedule them to an earlier date while they are already on the hospital campus. On the other hand, the requested appointment is more crucial for ECON and NCON patients considering that they want to enter the system, therefore, these patients prefer waiting for their booked appointments. Additionally, since ECON and NCON patients are externally referred or self referred, they are usually subjected to additional triage and testing, which lead to higher appointment delays being offered to those patients.

We design our simulation model to reflect the above-mentioned dynamics of the

system. The patient flow is modeled as follows. Upon an appointment request arrival, patient's priority class is assigned. Based on this priority and scheduling policy, an appointment with a certain delay is offered to the patient. The patient assesses this delay based on his WtW. If delay is less than the patient's WtW, appointment is booked and patient fulfills the appointment. If the delay is more than the WtW, patient still books it then subsequently cancels or not show up at the time of the appointment. We consider reschedules as cancellations and new requests since our main focus is the appointment requests and in most of the cases, patients are rescheduled to an earlier appointment. We label the patients that have WtW less than the delay as cancellation patients and no-show patients based on the patient class and delay value with the proportions estimated from the data. Additionally, for each delay value and for each patient class, we estimate a day of cancellation from the data if the patient is determined as a cancellation patient.

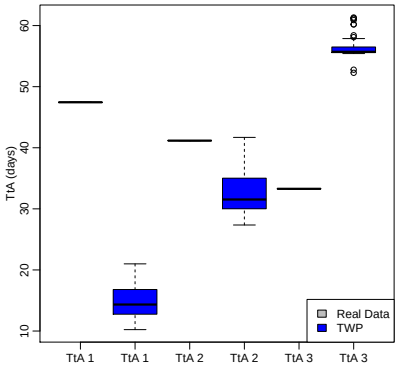
One of the issues of the dataset is that we do not have access to exact appointment calendar, therefore, we cannot identify the available capacity on each day and cannot calculate the exact utilization. To resolve this issue, we first calculate the average number of appointments that are either realized, no-show, or late cancelled (cancellation on the day of the appointment) to have an estimate on the capacity assuming no overbooks are used and no last minute patient is scheduled to late cancellation slots. After obtaining the estimate for available capacity, we replicate the real data with a simulation. Since we do not observe the exact dynamics of the scheduling policy, we use the exact data to replicate arrivals per day and delays offered to each patient. However, instead of directly replicating the C/RS/NS when we are considering the patient responses to delays, we use the realization probabilities as well as the cancellation probabilities and day of cancellation proportions that are estimated from the data. The reason we use simulation in this way is to estimate slot utilization and

make the results comparable to those from time window based policy. Notice that the average TtA values from the simulation will be the exactly the same as the real data while fill rates will be estimates. We run the simulation model for each of the observations in the dataset with 100 replications. The average fill rate that we obtain from the simulation model is 63.4% for priority 1 patients, 66.3% for priority 2 patients, and 53% for priority 3 patients while the values are 62.8%, 66.1%, and 52.4% in the real data. We conclude that the simulation model produces results similar to real data and we use the WtW associated parameters that we estimated from the data as well as the estimated capacity in our simulation model to test the performance of time window based policy.

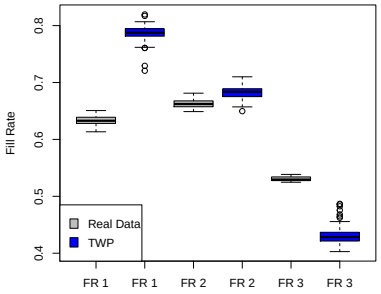
We first obtain time windows for the three priority classes, with booking horizon of length 90 days and $T_{\min} = 10$ and $T_{\max} = 30$. We set the available daily capacity to $C = 14$ where $\lambda_1 = 4$, $\lambda_2 = 3.4$, and $\lambda_3 = 14$. While determining the time windows, we only consider the results where each patient class can be served within the booking horizon with the available capacity. Therefore, while determining the time windows for higher priority classes, we ensure that the lower priorities can at least be served with minimum possible fill rate. The time windows that we obtain for priority 1 is $[1, 15]$ whereas the time windows for priority 2 and 3 are $[12, 41]$ and $[52, 62]$.

In real data, we cannot observe whether overbooking is used while assigning patients to appointment slots since we do not have an indicator in the dataset. Therefore, we do not include overbooks in our simulation model, we only utilize the regular slots. However, when we are setting the time windows, we set $\theta_{\max}$ to 0.5 since setting it to 0 results in low effective capacity. For instance, when $c = 14$, expected number of overbooks are close to 0 with effective capacity is 7.13 that does not allow all three classes to be served. Since we are not using overbooks in our simulation model, if there is no available slot available for patients within their time windows, the policy

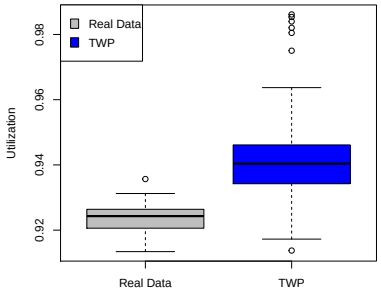
assigns patients to the first available slot beyond their time window.



(a) Average TtA



(b) Fill Rate



(c) Utilization

Figure 3.8: Simulation Results for TWP on Real Data

The results show TWP effectively delaying lower priority patients further into the future to improve performance measures for higher priority patients. Along with improving the TtA, TWP also improves FR for priority 1 and priority 2 patients by diluting more demand from priority 3 patients. Even though TWP results in lower FR for priority 3 compared to that observed in real data, we observe a slight increase in the slot utilization. This result might be due to the delay dependent day of cancellations for the patients who do not fulfill the booked appointments. From the real data, we observe that the patients who book appointments with lower delay values tend to

late cancel or reschedule, or cancel on a day closer to the appointment day compared to ones who experience higher delays. Therefore, it is expected to observe more late cancellations if priority 3 patients are booked earlier on the booking horizon, which can result in unutilized appointment slots.

Our analysis shows the effectiveness of using TWP using both artificially generated data and a real dataset from a real hospital. In our numerical experiments, we mainly focus on maximizing the proportion of higher priority patients with the effective capacity. While this is a valid objective for most cases, there might be other objectives such as minimizing the number of overbooks used under a given average TtA target. In those cases, objective function of our model can be changed to reflect decision maker's point of view.

While there are other effective policies to prioritize the patients, such as protection level policy, our results suggest that time windows based policies are performing well and easy to implement. From implementation point of view, time windows are guaranteeing a maximum appointment delay for each priority class which can be adapted based on patients' urgencies. Additionally, it is easy to maintain the time windows under the changes in patient arrival patterns and changes in patient mix since time window requires setting starting and ending days on the booking calendar for each priority class. With an effective method to identify the changes in arrival regimes, one can preemptively fine-tune the time windows easily and automatically update the policy without a significant effort.

As we discuss before, appointment office agents are acting as a gatekeeper to the system and the correct implementation of the appointment policy by the agents is essential for the success of the policy. Ideal way of implementing time windows is to design a software which only allows appointment office agent to see the associated part of the calendar at the time of the appointment request to avoid any human error

and provide a certain level of flexibility to both patients and appointment office agents by providing multiple possible appointment days and slots within the time window if the slots are available.

3.5 Compromised Prioritization

So far in this study, we have only focused on the case where strict prioritization is used. While using strict prioritization helps us to provide the highest level of service for the patients with the highest priority, it can result in some patient classes not being served due to serving higher priority classes with minimum level of dilution. In real life, not serving a patient class due to serving another class with higher fill rate is rarely observed. In some cases, serving all patient classes might be more desirable for decision makers rather than serving some classes with the highest possible fill rate. We call this scheme "compromised prioritization," where service provided for a higher priority class is reduced due to providing a higher service level to a lower priority class or to avoid rejecting all appointment requests from a lower priority class. For instance, in our analysis on real data, we relax strict prioritization to be able to serve patients from all three priority classes.

There are three main levers in providing care: fill rate, average TtA, and expected number of overbooks. Fill rate is a metric that represents the service level that are provided to patients while average TtA can be considered as the indicator of the quality of service provided. Average TtA metric for each patient class n for a time window $[B_n, E_n]$ can be defined as

$$\frac{B_n + E_n}{2}, \tag{3.20}$$

since we assume that patients are equally likely to be scheduled to any day within their time windows.

Especially for the patient populations that require urgent medical care, there might be certain time limits to provide care, we call these limits "safety fences." We denote these safety fences as τ. Since safety fences indicate medical urgency, this metric should be satisfied under any conditions. If the available capacity is not enough to provide care within the safety fence, the decision maker can either increase the regular capacity by hiring more providers or simply use overbooks as a temporary way of increasing the capacity.

While fill rate and average TtA metrics are related since higher fill rates are expected to be observed for lower average TtA values, they are not directly indicating the same results. Based on population WtW and medically-determined safety fences, one performance measure might limit the performance of the other. For instance, for patient populations that are more tolerant to appointment delays, high fill rates can be observed even for higher average TtA values which might exceed the medically determined safety fences.

The dynamic between the three levers that we describe and trade-off between providing care for patients from different priority classes can be captured by providing alternative set of solutions. These sets of solutions can be used by the decision maker by observing the service level and service quality that can be provided with the current capacity, determining the level of care that can to be provided to each priority group, and required capacity to reach certain service level and average TtA targets.

We use the real dataset that we utilize in Section 3.4 to provide set of alternative time windows that can used under compromised prioritization. While we are determining the time windows for each priority class, we take the hierarchy of priority levels into consideration. Specifically, as we did for strict prioritization, we first set the time windows for priority 1 patients and then calculate the possible time windows for priority 2 with the remaining capacity from priority 1. However, if the desired

service level or average TtA cannot be reached for priority 2 patients, the service level for priority 1 patient is lowered without violating the desired service level or average TtA for priority 1 patients. After fixing the time windows for priority 1 patients, same approach is used to analyze the trade-off between priority 2 and priority 3 patients. Note that at each step we consider the trade-off between two consecutive priority classes when determining the time windows for the higher priority class. Then continue with the remaining classes by setting the time windows for the higher priority class at each step.

In order to provide alternative set of solutions, we first determine relevant $\theta_{\max}$ values that can be used as in this setting. With the arrival parameters determined from the real life data, we calculate the maximum expected number of overbooks possible as 3.22 and the minimum as 0.125. We set the $\theta_{\max}$ values as 0.5, 1, 2, and 3 to observe the range of results that can be reached with different levels of available capacity. Under these $\theta_{\max}$ values, we first set targets on safety fences for each priority class and provide alternative solutions that satisfy these targets. We additionally consider the performance of strict prioritization. The set of targets are presented in Table 3.2.

Table 3.2: Targets Determined on Performance Measures

Target No.	τ_1	τ_2	τ_3
T1	Strict prioritization		
T2	6	10	15
T3	10	15	20
T4	10	30	45
T5	20	30	45

In T2, our goal is to observe the case where patients from each priority class are urgent, and targeted safety fences are relatively close. T3 considers a case similar to T2 with less strict safety fences. T4 covers the case where there are distinct targets, and T5 is the case that patients from all priority classes are less urgent compared to other cases.

Table 3.3 presents the time windows generated for each class under alternative targets and limit on expected number of overbooks. There are various insights that can be gained from the results in Table 3.3. Under $\theta_{\max}$ values 2 and 3, we observe that we can serve priority 1 and priority 2 patients with the natural dilution. Therefore, we do not observe any differences between the time windows set under different targets for those patients. T2 is the most restrictive target among the ones that we set. Under $\theta_{\max}$ values 0.5 and 1, there is no solution that can satisfy the safety fence targets for all patient classes. Therefore, we say that under T2, decision maker should increase the available care capacity by using more overbook slots to provide higher quality of care.

We can observe the trade-off between providing higher level of care to a higher priority and serving a lower priority patient clearly in $\theta_{\max} = 0.5$, under targets T4 and T5. While we cannot provide the required service quality under T4 to patients with the available capacity, with a small compromise from the service level provided for priority 1 patient, all patient classes can be served under T5. As it is shown in Table 3.3 for $\theta_{\max} = 0.5$, under targets T4 and T5, a minor reduction in FR_1 and in average 2 days of increase in TtA for priority 1 patients, priority 3 patients can be served without increasing the available capacity.

The time windows that are provided in Table 3.3 are not the unique solutions that satisfy the targets. For instance, under T3, we serve priority 1 patients with the earliest available time window while serving other priorities at their average TtA

Table 3.3: Time Windows Set Under Each Target

$\theta_{\max}$	Target No.	$[B_1, E_1]$	FR_1	$[B_2, E_2]$	FR_2	$[B_3, E_3]$	FR_3
$\theta_{\max} = 0.5$	T1	[1,10]	0.86	[1,10]	0.88	-	-
	T2	[1,11]	0.858	[5,15]	0.825	-	-
	T3	[3,17]	0.789	[5,25]	0.794	-	-
	T4	[3,17]	0.789	[25,35]	0.672	-	-
	T5	[6,18]	0.763	[25,35]	0.672	[33,57]	0.444
$\theta_{\max} = 1$	T1	[1,10]	0.86	[1,10]	0.88	[20,42]	0.467
	T2	[1,10]	0.86	[5,15]	0.825	-	-
	T3	[1,10]	0.86	[5,25]	0.794	[15,25]	0.488
	T4	[1,10]	0.86	[1,10]	0.88	[20,42]	0.467
	T5	[1,10]	0.86	[1,10]	0.88	[20,42]	0.467
$\theta_{\max} = 2$	T1	[1,10]	0.86	[1,10]	0.88	[5,14]	0.602
	T2	[1,11]	0.858	[1,10]	0.88	[5,14]	0.602
	T3	[1,10]	0.86	[1,10]	0.88	[5,14]	0.602
	T4	[1,10]	0.86	[1,10]	0.88	[5,14]	0.602
	T5	[1,10]	0.86	[1,10]	0.88	[5,14]	0.602
$\theta_{\max} = 3$	T1	[1,10]	0.86	[1,10]	0.88	[1,12]	0.696
	T2	[1,10]	0.86	[1,10]	0.88	[1,12]	0.696
	T3	[1,10]	0.86	[1,10]	0.88	[1,12]	0.696
	T4	[1,10]	0.86	[1,10]	0.88	[1,12]	0.696
	T5	[1,10]	0.86	[1,10]	0.88	[1,12]	0.696

target. Similarly, for $\theta_{\max} = 1$ under targets T4 and T5, we are providing care for priority 1 and priority 2 patients much earlier than their required safety fence. We can generate alternative solutions under these targets that allow lower priority patients to be serviced with higher service level by reducing the service level for higher priority patients without violating their targets. To observe this trade-off, we generate alternative solutions specifically for $\theta_{\max} = 1$, under targets T3 and T4.

Table 3.4: Alternative Time Windows

$\theta_{\max}$	Target No.	$[B_1, E_1]$	FR_1	$[B_2, E_2]$	FR_2	$[B_3, E_3]$	FR_3
$\theta_{\max} = 1$	T3	[4,14]	0.810	[5,25]	0.794	[8,32]	0.504
	T3	[5,14]	0.8038	[1,29]	0.806	[8,32]	0.504
	T4	[1,13]	0.848	[16,27]	0.752	[7,36]	0.502
	T4	[5,14]	0.803	[16,27]	0.752	[8,26]	0.515

When presenting the alternative results, we generate two alternative solutions where service level for priority 1 are set above 80%, above 75% for priority 2, and above 50% for priority 3 patients.

As we note before, average TtA and fill rate are similar metrics, however, one can be overcoming the other. For instance, to be able to satisfy the safety fence for priority 2 patients in T3, resulting fill rate should be higher than 75%. Therefore, TtA target overcomes the fill rate target for priority 2 patients in the cases we present in Table 3.4.

The decision maker can utilize results based on the system specific goals and additional preferences, and implement time windows accordingly. If the decision maker's main goal is to provide service to all patient types while using some form of strict prioritization that does not allow rejecting requests from a priority class, the

results that are presented in Table 3.3 are the time windows that should be used to prioritize patients.

If the decision maker's goal is to satisfy certain service level targets for each priority class under medically determined safety fences, compromising the performance of higher priority patients as we present in Table 3.4 is an effective approach to reach the targets. Among the results that are presented in Table 3.4, decision maker can make additional decisions on the features of the time windows. For instance, under target T3 and additional fill rate targets, decision maker can pick the window $[1, 29]$ to reduce the minimum possible delay or pick the window $[5, 25]$ to reduce the variance of the appointment delay experienced by priority 2 patients.

In a setting where patient urgencies are not the main concern of the decision maker while the service level is an important measure, the trade-off between service level and average overbook required should be analyzed. For instance, if the target service level is to provide 75% service level to priority 1 patients, 70% to priority 2, and 50% to priority 3 patients, the lowest level of expected overbooks that can be used to reach the service level targets is 0.762 while it is 1.587 for service level targets 80%, 75%, and 60% for priority 1, priority 2, and priority 3 patients, respectively. A small increase in the number of overbook slots used can result in significant improvements in service level as well as reduce average TtA for priority 1 patients from 15.5 to 9.5, from 29 to 20.5 for priority 2 patients, and from 21 to 12 for priority 3 patients.

The reason why we observe more significant changes in performance measures with a small increase in overbooks used is due to the characteristics of WtW distribution. Since patients are more sensitive to the increases in appointment delays for the lower delay values, we can obtain significant improvements in service levels by serving patients earlier in the booking horizon, especially for delay values lower than 20 days.

Another insight that we can get from the analysis is the effect of $\theta_{\max}$ value on the performance measures. As we note before, 3.22 is the maximum expected number of overbooks with the set parameters and the available capacity. Therefore, we can say that increasing $\theta_{\max}$ value beyond 3.22 does not lead any improvements in the performance measures. Additionally, we expect to observe that the relative improvement in performance measures diminishes as $\theta_{\max}$ value increases. This observation can be made from the results that we present in Table 3.3. From Table 3.3, we can observe that increasing the number of available overbook slots from 2 to 3 only changes the service level that can be provided to priority 3 patients while we can improve performance measures significantly by increasing $\theta_{\max}$ from 0.5 to 1.

If we have access to associated cost parameters or any sort of financial data, we can decide on ideal level of overbook capacity under the diminishing marginal return of the performance measures. We can also decide on the best solution that needs to be used among the alternatives by conducting a cost-benefit analysis. For instance, between the two alternative cases that are presented in Table 3.4 under T4, the main difference is the service level provided for priority 1 and priority 3. If serving additional 4.5% of priority 1 patients is more financially beneficial than serving 1.8% of priority 3 patients, one can choose the set of time windows in the first alternative presented in Table 3.4 under T4 than the second alternative. However, obtaining these cost parameters in this setting is not an easy task to conduct a financial analysis. Instead of cost components, one can use certain weights for performance measures that are determined by the decision maker to identify the optimal time windows to be chosen among the alternatives.

We generate additional data on patient arrivals and WtW distributions to further analyze compromised prioritization in two patient classes case. We generate trade-off curves under different system parameters on possible fill rate metrics for patients from

different priority classes. We construct the trade-off curves between the fill rates of priority 1 and priority 2 patients by evaluating the maximum β_2 value for each possible value of β_1 that satisfies inequality (3.11). Then, we generate managerial insights from the trade-off curves to assist decision maker to decide on the time windows that need to be used. We give further details in Appendix B.3.

3.6 Conclusion

Timely access is a critical component of quality of healthcare delivery. Throughout the years, healthcare systems facing issues regarding serving increasing demand to healthcare resources with limited clinical capacity. Due to this increasing demand, meeting right patients with right providers by allocating the available capacity becomes extremely important for both patient and provider satisfaction. While improving patient access can be considered as increasing the performance for overall system, the level of access necessary for patients are not identical considering responses to access delays, patient care needs, urgencies and priorities. Characterizing the patients' responses to delays and identifying patients' care needs are crucial in effectively utilizing the clinical capacity in providing timely access.

While the studies in the outpatient scheduling area is suggesting methods to utilize the available capacity in a way that improves the performance measures, addressing the mismatch between the clinical capacity and patient demand in a single perspective is not sufficient. An efficient method in improving patient access should consider the system as a whole rather than focusing on a single component and should not assume demand as an independent from the access policies used since patients react to the prolonged delays. While it is not possible to reduce the source of the demand, it is possible to manage the demand with an effective access policy by acknowledging the patients' inherent reactions to delays and using it as a lever.

Our main goal is to develop higher level access protocols that consider patient characteristics rather than determining slot level appointment decisions. We focus on developing such a policy that we refer as time window based policy which mainly identifies time intervals on the booking horizon that patients can be assigned to based on their priorities and their tolerance to appointment delays offered. The idea behind this policy is encouraging non-urgent, low priority patients to seek care at other institutions by offering appointments with higher delays and use appointment delay as a lever to control the patient demand by diluting the total demand in a way to match it with the available clinical capacity. This approach can be considered as weak rejection where patients from lower priorities are offered appointments that are further into future rather than directly getting rejected. The policy that we design utilizes patient WtW behavior and patient priorities to identify the time windows.

The main challenge in appointment systems is curse of dimensionality due to keeping track of the appointment calendar. To avoid curse of dimensionality, we consider a simplifying yet an effective approach that considers each time window as an uncapacitated bin. With the assumption that each patient is scheduled equally likely to any day within their respective bins, uncapacitated bin approach helps us to calculate the expected demand load as a function of the time windows allocated for each patient priority class and patient WtW without keeping track of the full appointment calendar.

Our suggested approach is similar to pricing strategies in revenue management which focus on determining set of fare classes to open at each point in time to discourage the customers that are not willing to pay the offered price. We use inherent patient WtW in a similar manner by offering higher delays to patients from lower priority classes. Time windows policy is a completely new perspective in patient access context since our approach mainly focusing on controlling the patient demand

through prioritization and improve access for the patients who requires timely access to care since concept of improving the access is not the same for every patient.

We specifically focus on the setting where patient priorities can be identified at the time of the appointment request and decision maker uses strict prioritization. Our study contributes to area of outpatient scheduling by bringing an innovative way to improve patient experience and provide care for the patients who are in dire need and sensitive to appointment delays. Via extensive simulation experiments, we show that the access policy that we propose performs better than the common policies that are used in practice. In addition to its performance, time windows based policy is also beneficial due to its ease of implementation in real life. It is easy to adjust based on the changes in total traffic and patient mix and these adjustments can be put in action without requiring significant effort and time.

We then utilize a dataset from a specialty clinic that patient priority classes can be identified from the available data to observe time window policy's performance in a real system. We show that the policy effectively improves performance measures in a real life system that does not completely satisfy our modeling assumptions by using a simulation model that is calibrated with real data. In the simulation model, we use realization probabilities for each delay value that represents the probability that a booked appointment will be realized to represent patient WtW. We estimate these probabilities from the available data by employing a statistical model designed for estimating patient WtW.

For our policy to be successfully applied in practice and improve patient access, it is necessary to approach the system in a comprehensive manner since the policy that we offer consider patient priorities and behavior in making decisions. The first step of this comprehensive approach should be identifying the patients who can benefit most from the provided service, i.e., identifying the *right* patients to prioritize. Establishing

the target patient population is critical to improve access since patients cannot be considered as equal in terms of access needs. In our study, we consider referral type based priorities that indicates a prioritization scheme considering patients need to for entering the system and attracting new patients. The nice feature of this prioritization scheme is that the priorities can be assigned at the time of appointment request and easy to identify.

Another way of identifying patient priorities can be considered as prioritizing patients in terms of medical necessity and urgency. While medical necessity and urgency are ideal characteristics to identify each patient's priority class, observing those characteristics at the time of the appointment request is a challenging task. In most cases, to identify patient's condition a preliminary triage is required by the medical providers. While pre-triage results in accurate priority assignments, it is costly and causes additional delays in access to care. An effective way of identifying need based priority levels is to utilize tools to detect prognostic evidence from available data. With the advancements in data mining and machine learning techniques, deriving critical information from Electronic Health Records (EHR) becomes more accessible. Developing a data-driven triage tool by employing machine learning techniques can be considered as a promising direction for a future study.

Another direction that we can follow for future study is identifying patient behavior in a more detailed way and reflecting it in our model in more details. For instance, as we observe in our case study not all patients who have WtW lower than the appointment delay balk without booking an appointment. As we indicate before, we observe that at many encounters, patients with WtW lower than the appointment delay can book the offered appointment but then either reschedule to an earlier date, cancel completely, or decide not to show up for the appointment. Notice that among all these disposition reasons only cancellations results in reduction in the system load.

However, this reduction does not observed instantly since patients who cancel first keep the appointment until a certain point in time then cancel before the time of appointment. Since we are specifically focusing on the total load on the first day of the appointment calendar, we can include the late cancellations and no-shows into our model to calculate the utilization and overbook slots utilized in more details.

Lastly, we can extend this study for different settings under different objectives. One objective can be determining the time windows to minimize the deviations from the available capacity under specific safety fences. We can also control the characteristics of time windows in terms of the variance of the appointment delays that are experienced by the patients from the same priority class.

Chapter 4

DYNAMIC ASSIGNMENT OF PATIENTS TO PRIMARY AND SECONDARY INPATIENT UNITS

4.1 Introduction

Over the last decade, hospital Emergency Department (ED) overcrowding has become a widely recognized problem in healthcare delivery in the U.S. and around the world. In a report to congress, the U.S. Government Accountability Office highlighted this problem, and emphasized that ED waiting times for the emergent patients exceeds the recommended time window for 50% of visits (GAO (2009)).

ED overcrowding may have dire consequences, including higher complication rates and even increased mortality (Bernstein *et al.* (2009), CNN (2008)). As overcrowding increases, patients are subject to higher dissatisfaction, impaired access, higher rates of leaving without being seen (LWBS), and decreased economic performance (Hoot and Aronsky (2008)).

One important factor associated with ED overcrowding is the prolonged ED boarding of patients admitted to inpatient units (GAO (2003)). ED boarding occurs when an admitted patient waits for transfer to an inpatient unit due to bed unavailability in a downstream unit. Although this may cause congestion and may block ED resources from being assigned to newly arrived patients (see, e.g., Saghafian *et al.* (2012), and the references therein for the so-called "bed-block" effect), boarding may be viewed in a positive light by some when it is done to ensure transfer to the most appropriate inpatient unit (rather than a secondary unit that has bed availability). This is due to a questionable yet prevalent belief that patience in transferring admitted

patients is always a virtue. This belief deserves further scrutiny, especially because it is well understood that prolonged boarding times have several negative consequences for patients, including an increased risk of adverse events (ROAE). [1]

The assignment of hospital beds to patients is a challenging task due to several complexities, including limited capacity of hospital beds, time-dependencies of bed request arrivals, and unique treatment requirements of patients (Proudlove *et al.* (2007)). These complexities force hospital administrators to incorporate various aspects of the operational status of their system (such as the current congestion level, time of the day, and discharge times in inpatient units) in their decision-making process. Nevertheless, from a medical standpoint, the ideal way of assigning a bed for a specific type of patient is directly related to the patient's medical diagnosis and treatment needs. However, to accommodate patient demands with the limited available hospital resources, hospital administrators may consider alternative assignment options. In particular, when there is no available bed in the ideal downstream unit (i.e., the patient's primary inpatient unit), the patient may be assigned to an alternative, secondary inpatient unit with an acceptable (if suboptimal) service capability and capacity. This practice of assigning patients to an alternative unit is known as "overflowing."

Overflowing is not a new concept in hospitals; however, in practice the overflow process is often controlled in a myopic manner without much attention to the needs of future patients. Instead, what is needed is a reasonable balance between the risk of keeping a patient in the ED (with the hope of a primary unit assignment) vs. that of assigning the patient to a secondary unit that has current bed availability. A careful consideration of these trade-offs might have a significant impact on both

[1] Patients boarded in the ED are sometimes kept on hallway beds, which raises additional concerns about whether they receive the care that is deemed necessary for them — the inpatient unit level of care.

Table 4.1: IWs and Their Sizes in MCA

IW Name	Abbrev.	Definition	No. of Beds
2 West	2W	Intensive Care Unit (ICU)	30
3 East	3E	Orthopedics and Urology Surgical Services	40
3 West	3W	Medical/Surgical Organ Transplant	36
4 East	4E	Hematology, Oncology	30
4 West	4W	Cardiology and Cardiothoracic Surgery	36
5 West	5W	Neurosciences and E.N.T.	36
7 East	7E	Palliative Care, General Surgery	36
7 West	7W	Hematology and Oncology patients with medical-surgical overflow	24

patient safety and operational efficiency of hospitals. Our goal in this chapter is to develop a systematic approach to facilitate better decision making with respect to inpatient unit assignments.

To gain insights, we explore these issues in our partner hospital, Mayo Clinic Arizona (MCA). There are eight inpatient wards (IWs) in MCA from which a bed can be requested for an admitted patient. A detailed description of these eight IWs are shown in Table 4.1. The data we have collected from MCA shows that the average ED boarding time (the average time between bed request and occupancy) at MCA is 111 minutes, with boarding times up to 150 minutes for some patients. Moreover, we observe from our data that about 30% of the patients admitted through the MCA ED are boarded for at least two hours. An average delay of 111 minutes is significant, especially when we consider that the average ED Length of Stay (LOS) for admitted patients in MCA is about 5 hours. This suggests that, on average, almost 37% of ED

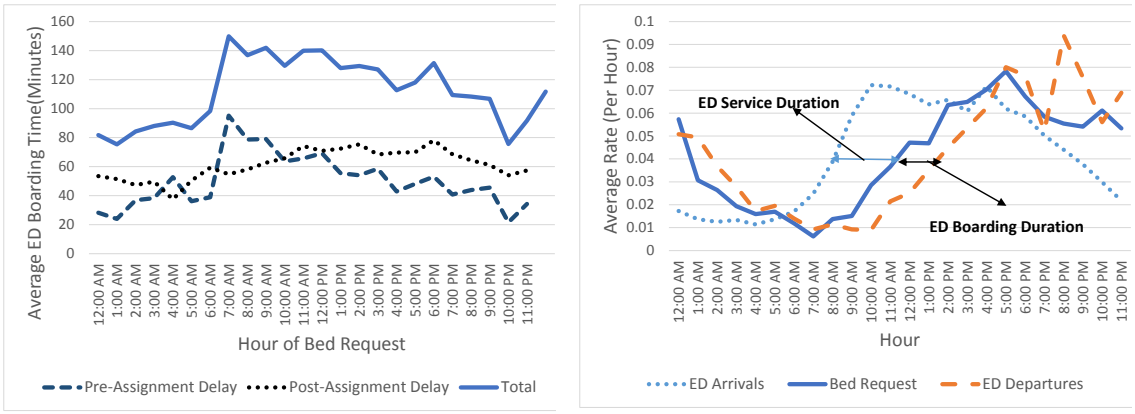
LOS is caused by boarding. [2] Furthermore, as is shown in Figure 4.1(a), we find that boarding duration is highly time-dependent. Therefore, even if the average waiting time is not extremely long, patients admitted through the ED experience different levels of delay based on the hour in which their inpatient bed is requested.

As we illustrate in Figure 4.1(a), boarding delays consists of two parts: Pre-Assignment and Post-Assignment. Pre-Assignment delay is the time between bed request and assignment of a suitable inpatient bed to the patient. Post-assignment delay is the time between bed assignment and bed occupancy. Our analyses of MCA data reveal that post-assignment delays are higher on average than pre-assignment delays (see Figure 4.1(a)). Additionally, as we show in Figure 4.1(b), there is a significant mismatch (i.e., time lag) between the hourly bed request pattern and the ED departure pattern (see, e.g., Shi *et al.* (2015), Armony *et al.* (2015), and Powell *et al.* (2012) for related results reported for other hospitals). The time between ED arrivals and ED departures in Figure 4.1(b) is defined as ED LOS, and the time between bed requests and ED departures represents the ED boarding time. As can be seen from this figure, the ratio of ED boarding time to ED LOS can be as high as 48% for some patients.

Effective assignment policies to primary and secondary inpatient units might significantly help hospitals such as MCA to improve their prolonged ED boarding times. In this study, we utilize a variety of analytical and simulation analyses calibrated with hospital data to gain insights into the structure of such policies as well as their achievable improvement magnitudes. In particular, we seek to answer the following questions:

- *Structure:* When should a patient be kept in the ED until a bed becomes

[2] See also, Carr *et al.* (2010) who report that 17% of the ED total LOS is caused by the ED boarding.



(a) Average Boarding Time

(b) Non-stationary Arrival, Bed Request, and Departure Rates

Figure 4.1: ED Boarding Times Based on Collected Data from Our Partner Hospital

available in his/her primary inpatient unit instead of being quickly assigned to a secondary unit with current bed availability?

- *Magnitude:* How much improvement can be achieved if a hospital adopts an effective policy for dynamically assigning ED admitted patients to their primary or secondary inpatient units?

To gain insights and answer these questions, we start by utilizing a Markov decision process (MDP) and modeling the flow process as a multi-class queueing network problem with "flexible" servers. In this model, the servers are defined as the downstream inpatient unit beds that are "flexible," in that they can serve different classes of patients. The literature on hospital-like multi-class queueing systems with flexible servers that can address the appropriateness of bed assignment decisions is not vast. We contribute to this literature by considering (a) a stochastic penalty cost that reflects the reduction in service quality when a patient is assigned to a secondary inpatient unit, and (b) stochastic risk of adverse events that can occur due to prolonged

ED boarding times. By analyzing our MDP setting, we find that the optimal assignment policy is a state-dependent threshold-type policy: keeping patients in the ED for their primary inpatient unit to become available pays off, but only up to a certain threshold that depends on the number and status of outstanding ED bed requests. That is, *patience is a virtue, but only up to a point.*

Our findings and results regarding the structure of the optimal policy can help hospitals to make better bed assignment decisions, particularly as we shed light on some guidelines that can strike a better balance between patient safety, quality of care, and operational efficiency. However, we note that the optimal policy generated by our model is complex to use in practice, since it is highly dependent upon the system state (e.g., the number of patients of different types boarded in the ED). Therefore, based on the properties of the optimal policy, we develop two heuristic policies which are simple to implement and effective. We test these heuristic policies by comparing their performance with the optimal policy using a detailed patient flow simulation model calibrated with hospital data. We find that implementing our proposed assignment policy would reduce the average ED boarding time by 10 minutes per patient (a 9% improvement). Moreover, our analysis suggests that our proposed policy would improve a combined measure of patient safety and quality of care metrics by 14%, and would decrease the percentage of patients with more than two hours of boarding by 2%.

We also use our simulation framework to generate insights into hospital conditions under which such improvements can be most significant. Our results suggest that hospitals with higher congestion levels (e.g., busy teaching hospitals) would benefit more than other hospitals (e.g., less busy community hospitals) from utilizing our proposed policy as a way to strike a better balance between patient safety, quality of care, and operational efficiency. Our results also suggest that, under specific conditions on ad-

verse event rates and number of patients boarded in the ED, keeping an inpatient bed idle for potential future bed requests is beneficial. This practice of intentional bed idling is currently used in some inpatient units such as the ICU. However, our results provide support for implementation across a wider range of inpatient units, and reveal that bed idling should be used more broadly in hospitals.

The main contributions of this chapter are four-fold: (1) We generate insights into effective bed assignment policies by developing a model that considers the trade-offs between risk of adverse events that may occur while a patient is boarded in ED, and a potentially lower quality of care that might be provided if the patient is routed to a secondary unit. (2) We develop an easy-to-implement and yet effective policy for bed assignment in hospitals that considers multiple inpatient units, multiple patient types, time-dependent bed request arrivals, and dynamic ED and inpatient unit congestion levels. (3) By making use of some laboratory findings, and testing our proposed bed assignment policy via a detailed simulation model calibrated with hospital data, we generate various insights for hospitals. For example, we find that our proposed policy is more effective in reducing ED boarding times for patients that are less sensitive to assignment to a secondary inpatient unit. Examples of such patients include those without an elevated serum troponin (Tn) level among chest pain (CP) patients, or those with a B-type natriuretic peptide (BNP) less than 4,000 pg/ml among congestive heart failure (CHF) patients. (4) We also shed light on various hospital-dependent conditions under which our proposed policy is reasonably effective, thereby discussing the suitability of our proposed policy for implementation in a wide range of hospitals.

The rest of this chapter is organized as follows. Section 4.2 reviews the related studies on patient flow and dynamic assignment policies. Section 4.3 presents a model of patient flow, and develops an MDP framework that captures the trade-offs in the

model. Section 4.4 identifies the structure of the optimal policy. In Section 4.5, we describe our proposed heuristic bed assignment policy, and compare its performance with the optimal policy. In Section 4.6, we describe our detailed simulation model of patient flow, and use it to perform various sensitivity analyses. Finally, we present our concluding remarks in Section 4.7. All proofs are provided in Appendix C.2.

4.2 Literature Review

In this section, we briefly review studies that are related to our work. We divide such studies to two categories: (a) related studies on ED patient flow, and (b) related studies on dynamic assignment and routing in queueing systems.

4.2.1 Related Studies on ED Patient Flow

ED patient flow studies can be found in both the medical and operations research/management science literature. Such studies typically focus on patient flow either into the ED, within the ED, or out of the ED. An extensive review of operations research/management science contributions to these three elements can be found in Saghafian *et al.* (2015). Our work in this chapter focuses on patient flow out of the ED. Research on this last part of flow includes studies on effective ways for improving the process for those who are admitted to the hospital through the ED as well as those discharged to go home. Our study contributes to the former, and hence, we discuss only the relevant studies within that literature.

Harrison *et al.* (2005) use discrete-event simulation to analyze the effect of bed capacity on overflow rates. The authors indicate that seasonality of arrivals is one of the main triggers of overflow in hospitals. Thompson *et al.* (2009) study a capacity utilization-based patient allocation problem. In their model, patients may be transferred between different units to minimize the total cost under a preemptive

service policy assumption, where assignment to each unit is accompanied by a reward/cost. Similar to Thompson *et al.* (2009), we consider different levels of quality of care that can be provided in different inpatient units. However, unlike that study, we also model the risk of adverse events (ROAE) that can occur because of prolonged waiting in the ED. This allows us to provide a system-wide view that, in addition to operational efficiency, considers both patient safety and quality of care concerns. Another related study is Mandelbaum *et al.* (2012), which considers the fair routing of patients to inpatient units, where fair routing means targeting the same level of idleness among all servers. Unlike Mandelbaum *et al.* (2012), we consider patient routing as a mechanism to eliminate prolonged ED boarding times. Furthermore, the study of Mandelbaum *et al.* (2012) analyzes a model with a single customer class, whereas we consider heterogeneous patient classes in order to gain insights into the questions we raised in the Introduction.

Teow *et al.* (2012) use data mining techniques to identify factors that trigger overflow decisions. Unlike Teow *et al.* (2012), our study attempts to identify conditions under which it is optimal to overflow a patient to a secondary inpatient unit. Shi *et al.* (2015) focus on patient flow from ED to inpatient units, and propose early discharge policies in inpatients units as a mechanism to reduce and flatten ED boarding times. Our study focuses on a similar patient flow from the ED to inpatient units; however, unlike the predetermined trigger times in Shi *et al.* (2015), we (a) optimize bed assignment decisions based on the number of boarded patients in the ED, and (b) consider both patient safety and quality of care metrics. Furthermore, a policy of changing physician discharge routines that is described in Shi *et al.* (2015) might be hard to implement in many hospitals due to cultural issues such as difference in physicians' preferences. Our study offers guidelines on alternative ways of improving the patient flow.

Similar to our study, Griffin (2012) develops a patient flow model to improve bed assignment by maximizing the suitability of patient assignments and minimizing ED boarding times. The author evaluates five dynamic bed assignment algorithms to aid decision makers. Due to the large dimension of the state and action spaces, Griffin (2012) cannot identify the exact structure of the optimal assignment policy. In our study, we first gain insights into the structure of the optimal policy by using a stylized model of patient flow with two inpatient units and two patient types. We then make use of these insights to develop a heuristic policy. Using realistic simulations calibrated with hospital data, we next examine the performance of this heuristic policy in a realistic setting. This combination of analytical and simulation analyses allows us to fully address the questions we raised in the Introduction. In addition, instead of assuming that all inpatient units can serve as a potential secondary unit for all patients (as is assumed in the majority of the above-mentioned studies), we use historical hospital data, laboratory findings, and physicians' opinion to determine specific secondary inpatient units for each patient type.

4.2.2 Related Studies on Dynamic Assignment and Routing in Queueing Systems

Our model captures the system characteristics as a multi-class queueing system where the bed requests for ED admitted patients are considered as arrivals, and inpatient unit beds are considered as servers. In multi-class queueing systems, the customers can be differentiated based on service rates, holding costs, arrival rates, or service requirements. Under an average holding cost objective, Cox and Smith (1991) demonstrate that the widely-used $c\mu$ policy is optimal for both preemptive and non-preemptive cases service protocols. The $c\mu$ policy is also shown to be the optimal policy in various more complex queueing networks (see, e.g., Kakalik and Little (1971), Buyukkoc *et al.* (1985), and Walrand (1988)). A version of the $c\mu$ rule,

generalized $c\mu$, is proved to be the optimal policy for different queueing structures under heavy traffic (see, e.g.,(Van Mieghem, 1995; Mandelbaum and Stolyar, 2004)). Saghafian and Veatch (2016) establish the optimality of the $c\mu$ rule for queueing systems with flexible servers and two tier structures, where one tier is served by one server while the second tier can be served by all the servers.

In Lin and Kumar (1984), the authors show that when two types of servers with different service speeds are available—a setting termed "slow server problem"—the optimal assignment policy is a threshold-type policy: customers/jobs are assigned to the slow server whenever the queue length reaches a certain threshold. Our model resembles similar characteristics to the "slow server problem," because (a) patient service times in inpatient units are not identical, and (b) there is some flexibility in assignments (for some patients). However, instead of heterogeneous servers, we consider heterogeneous patient types with different service rates, since it better matches the hospital patient flow we study. This differentiates our study from the above-mentioned studies in the literature since in such studies the resulted optimal policy typically depends on the difference between service rates of servers (see, e.g., Bell and Williams (2001)). However, our data analysis shows that service durations in primary and secondary units are not statistically different (for patients of the same type).

Dynamic assignment problems in queueing networks are extensively analyzed in the literature (see, e.g., (Mandelbaum *et al.*, 2012; Meyn, 2001, 2003), and Palmer and Mitrani (2004)). Armony and Bambos (2003) and Dai and Lin (2005) study dynamic assignment problems considering a throughput maximization objective. Andradóttir *et al.* (2007) and Saghafian *et al.* (2011) allow for server disruptions and repairs in systems with heterogeneous flexible servers, and De Véricourt and Zhou (2005) study a call center setting where agents are heterogeneous in terms of both service rate and quality of service (see also Zhan and Ward (2013)). Another related stream of

literature that considers flexible servers is the "skill-based routing" literature, where the customers are routed to the servers that have the appropriate skill sets (similar to the routing of patients to primary vs. secondary units in our study). However, unlike our work, the focus of those studies are mostly on settings where (a) servers have multiple skills (e.g., call center agents), and (b) staffing decisions are the primary concerns (see, e.g., Garnett and Mandelbaum (2000), Gans *et al.* (2003), Wallace and Whitt (2005)). There are also various other studies on routing policies in multi-server, multi-class settings (see, e.g., Gurvich and Whitt (2009), Tezcan and Dai (2010), Armony and Ward (2010), Gurvich and Perry (2012)). However, in these studies only costs related to waiting and losing customers are considered, whereas we focus on the trade-off between waiting and overflows. Moreover, we note that the majority of the above-mentioned studies focus on heavy traffic settings. Unlike them, we seek to address the questions we raised in the Introduction under practical hospital congestion levels. To this end, we do not impose any heavy traffic assumption, and instead make use of actual hospital bed census data as the basis of our analytical and simulation analysis.

4.3 The Model

A general representation of patient flow through the ED and hospital inpatient wards (IWs) is presented in Figure 4.2. A patient that arrives to the ED goes through the triage stage, and is assigned an Emergency Severity Index (ESI). If there is an examination room available, the patient immediately starts the ED service; otherwise, he/she will have to wait in a designated ED waiting area. Once the ED treatment is done, the patient is either discharged home or is admitted to the hospital. For an admitted patient, if there is a bed available in his/her primary IW (or the secondary IW if applicable), the patient is transferred out of the ED; otherwise, he/she is kept

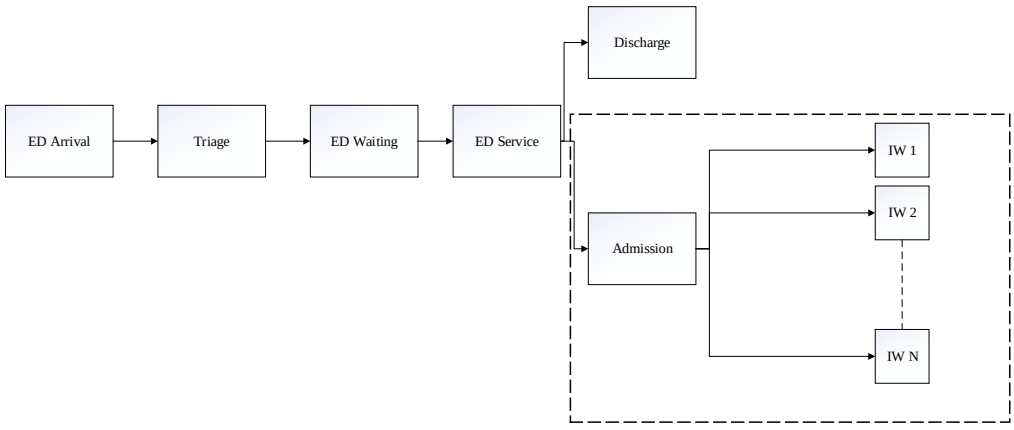


Figure 4.2: General Flow of Patients with the Dotted Area Representing the Focus of This Chapter (IW: Inpatient Ward)

in the ED until a bed becomes available. For the goals of this study, we focus on the patient flow within the dashed area of Figure 4.2. [3]

To gain insights into the questions we raised in the Introduction, we start by modeling the patient flow as a multi-class queueing system with IWs as flexible servers, and analyze it by using an MDP. The patients in the system are classified based on their primary IW, i.e., where they can be best served from a medical standpoint. Ward-level placement is typically determined by a bed placement coordinator, sometimes in consultation with the ED or the admitting physician. Once a patient is moved to an IW, the IW bed is considered as unavailable until the patient is done with the inpatient unit service, and hence, the service processes in IWs are typically non-preemptive. To gain some high level insights, we start by considering each of the IWs as a single "super server," which represents the capacity of the IW as a whole. This pooling of beds within each IW allows us to keep track of availability of capacity

[3]Thus, we do not consider measures related to events that occur outside this flow. For instance, an important measures for EDs is the percentage of patients who leave without being seen. But this occurs almost always from the waiting room of EDs (i.e., before the ED service starts), which is outside the dashed area in Figure 2.

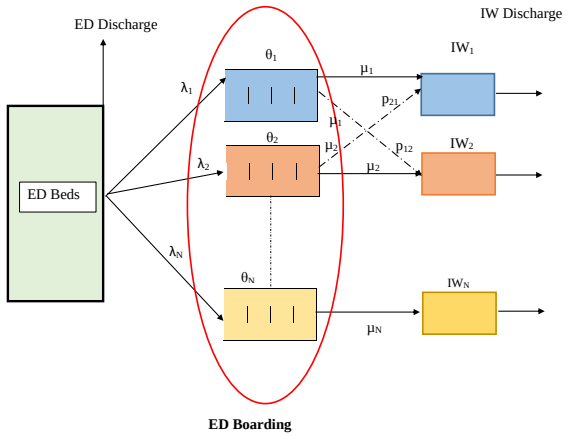


Figure 4.3: A Queueing Representation of the Patient Flow

in IWs in a computationally tractable way. However, to test the insights we gain from this simplifying assumption, we relax it in Section 4.6, and consider each IW bed as a server. Similarly, we start by considering the arrival process as a stationary Poisson Process, and assume IW service times are exponential. In Section 4.6, we also relax these simplifying assumptions by using empirical distributions (for both interarrival and service times) that we have estimated based on our data.

Figure 4.3 illustrates the patient flow under consideration as a queueing system. Our discussions with medical providers revealed that, for the vast majority of patients, only one IW can be considered as a secondary IW. [4] Hence, as illustrated in Figure 4.3, the system consists of multiple primary-secondary pairs, where each patient type has only one primary IW and only one secondary IW.

To analyze the patient flow depicted in Figure 4.3, we let N_p and N_s denote the set of patient classes and servers (IWs), respectively. For $i \in N_p$, we denote by λ_i the arrival (i.e., bed request) rate of class i patients. We model the service

[4] We also note that some patients can only be served in their primary unit (e.g., ICU patients). We still consider a primary-secondary pair for such patients, but disallow for service in the secondary IW by considering a high penalty cost for care delivery in the secondary IW.

process in IWs with class-dependent service rates μ_i where $i \in N_p$. We also let $X_i(t)$ denote the number of class i patients boarded in the ED at time t, and define $\underline{X}(t) = (X_i(t) : i \in N_p)$ as the vector of the number of all such patients. Moreover, for $i \in N_p$ and $j \in N_s$, we let $a_{ij}(t) = 1$ if IW j is serving a class i patient at time t, and $a_{ij}(t) = 0$ otherwise. We model the potential occurrence of adverse events that might occur for patients boarded in the ED (awaiting transfer to an inpatient unit) by class-dependent Poisson process, In particular, we let $\bar{\theta}_i$ denote the per unit of time risk of adverse events (i.e., the rate of the underlying Poisson process) that can occur for a class i patient boarded in the ED, denote by c_i the associated cost per adverse event, define $\theta_i = c_i\bar{\theta}_i$, and let $\underline{\theta} = (\theta_i : i \in N_p)$. In this setting, θ_i plays the role of "expected holding cost" for a patient of class i, and is accrued per unit of time boarding in the ED. However, the actual "holding cost" is random and depends on stochastic deteriorations in the patient's conditions. Similarly, for assignments to secondary inpatient units, we let p_{ij} denote the expected value of a non-negative "penalty cost" (which is random in nature due to its dependency to various patient and provider-dependent conditions) that is accrued due to a lower-than-desired quality of care when a patient of class i is assigned to IW j ($p_{ij} = 0$ if $i = j$). [5]

The objective is to find an optimal assignment policy to control the patient flow in order to minimize the expected total long-run average sum of (a) adverse events (a patient safety concern), and (b) the penalties accrued due to placement in secondary units (a quality of care concern). [6] This optimal objective can be calculated as:

[5] In Section 4.6, we will discuss how we have used a year of data on patients with chest pain (CP) or congestive heart failure (CHF) to estimate all the parameters required for our model.

[6] We may refer to these as "costs" for simplicity. However, it should be noted that these are general, and may include various negative consequences of undesirable outcomes with respect to patient safety and/or quality of care caused by patient flow decisions. We refer interested readers to empirical studies such as Kuntz *et al.* (2014), Berry Jaeker and Tucker (2016), Chan *et al.* (2016)), and the references therein for further examples of such outcomes.

$$Z^* = \inf_{\pi \in \Pi} Z^\pi = \inf_{\pi \in \Pi} \left[\sum_{i \in N_p} \sum_{j \in N_s} p_{ij} O_{ij}^\pi + \sum_{i \in N_p} \theta_i L_i^\pi \right], \tag{4.1}$$

where Π is the set of admissible (non-preemptive, non-collaborative, and non-anticipative [7]) policies, Z^π is the long-run average objective under policy $\pi \in \Pi$, L_i^π denotes the long-run average number of class i patients in the queue (i.e., boarded in the ED) under policy $\pi \in \Pi$, and O_{ij}^π denotes the long-run average number of class i patients overflowed to IW j under policy $\pi \in \Pi$. In this setting:

$$L_i^\pi = \limsup_{T \to \infty} \frac{1}{T} \int_0^T E\left[X_i^\pi(s)\right] ds, \tag{4.2}$$

$$O_{ij}^\pi = \limsup_{T \to \infty} \frac{A_{ij}^\pi(T)}{T}, \tag{4.3}$$

where $A_{ij}^\pi(T)$ is the cumulative number of times up to time T that IW j has been assigned to a class i patient under policy $\pi \in \Pi$ (i.e., a counting process associated with $a_{ij}(t) = 1$).

4.3.1 A Markov Decision Process Formulation

As mentioned earlier, our partner hospital has eight main IWs (see Table 4.1), and hence, $|N_p| = |N_s| = 8$. However, as noted earlier, because each patient type has only one primary and one secondary IW, the hospital can be viewed as multiple primary-secondary IW pairs. Hence, we expect the insights generated by focusing on a single primary-secondary pair to be useful for the whole hospital system. For this reason, and to gain some clear insights into effective patient flow control policies, we start by considering the simplest case where $N_p = N_s = \{1, 2\}$, and later test the insights gained via simulations calibrated with data for a larger system. We let $\underline{a_1} = (a_{11}, a_{21})$ and $\underline{a_2} = (a_{12}, a_{22})$, where $a_{ij} = 1$ if server j is busy with a patient

[7] The reason we focus on non-anticipative policies is that even when the providers have a rough estimate on the discharge times of their patients, the exact discharge time is unknown and can be affected by several factors. Similarly, the exact timing of future bed requests are not known.

of class i. We assume that all the underlying processes are memoryless, and require that at any point in time $\sum_{i \in N_p} a_{ij} \leq 1 \ \ (\forall j \in N_s)$. With these, we define the system state as $\tilde{X} = (\underline{X}, \underline{a_1}, \underline{a_2})$ with state space $\mathcal{S} = \mathbb{Z}_+^2 \times \{0,1\}^2 \times \{0,1\}^2$. [8] We then use uniformization to transfer the underlying continuous-time Markov chain (CTMC) to a discrete-time Markov chain (DTMC). Let $\psi = \lambda_1 + \lambda_2 + 2\max\{\mu_1, \mu_2\}$ be the uniformization factor. Then, the long-run average cost optimality equation for the DTMC can be written as:

$$
\begin{aligned}
J(\tilde{X}) + \hat{Z}^* = \frac{1}{\psi} \Bigg[& \underline{\theta}\, \underline{X}^T + \min_{u = \underline{u_{ij}} \in \mathcal{U}(\tilde{X})} \Bigg\{ \sum_{i \in N_p} \sum_{j \in N_s} \lambda_i T^{\underline{u_{ij}}} J(\underline{X} + e_i, \underline{a_j}) \\
& + \sum_{i \in N_p} \sum_{j \in N_s} \sum_{k \in N_p} a_{kj} \mu_k T^{\underline{u_{ij}}} J(\underline{X}, \underline{a_j} - e_k) \\
& + \left(\psi - \sum_{i \in N_p} \lambda_i - \sum_{k \in N_p} \sum_{j \in N_s} a_{kj} \mu_k \right) J(\tilde{X}) \Bigg\} \Bigg],
\end{aligned} \tag{4.4}
$$

where $J(\tilde{X})$ is a relative cost function defined as the difference between the total expected cost of starting from state $\tilde{X}$ and a reference state (state $\underline{0}$), $\hat{Z}^*$ is the optimal average cost per uniformized period, the notation "^{T}" represents the transpose operator, and $T^{\underline{u_{ij}}}$ is a functional operator that depends on action vector $\underline{u_{ij}}$ and is defined in Appendix C.1. In optimality equation (4.4), e_i is a vector with the same dimensions as $\underline{X}$ containing a one in the ith position and zeros elsewhere. Thus, the first line inside the minimization in (4.4) is due to inpatient bed request arrivals from the ED, which occur with rate λ_i for patients of class i. Similarly, the second line in (4.4) is due to discharges of patients from IWs, and the last line in (4.4) is due to the self-loop in the underlying DTMC. The control actions u_{ij} in (4.4) are taken so as to minimize the long-run average cost, where the set of admissible actions is:

[8] Since we do not allow preemptions to better reflect the actual practice, it is necessary to keep track of the IWs' availabilities $(\underline{a_1}, \underline{a_2})$ as a part of the state.

$$\mathcal{U}(\tilde{X}) = \Big\{ u = (u_{ij})_{i\in N_p, j\in N_s} \text{ s.t. : } u_{ij} \in \{0,1\}, \tag{4.5}$$
$$\sum_{i\in N_p} u_{ij} \leq (1 - \sum_{i\in N_p} a_{ij}) \ \ \forall j \in N_s, \sum_{j\in N_s} u_{ij} \leq X_i \ \ \ \forall i \in N_p \Big\}.$$

That is, a patient cannot be assigned to IW j, if IW j is busy or if the number of patients boarded in the ED is insufficient.

4.4 The Optimal Patient-IW Assignment Policy

In Appendix C.2, we show that we can restrict our attention to policies that do not allow idling an IW $j \in N_s$ when there is a patient with IW j as his/her primary IW boarded in ED (See Proposition 2 in Appendix C.2). [9] Although this is an expected result in service systems in which preemption is allowed, we note that in non-preemptive services such as the one we model, this insight can be counter intuitive. To establish this non-idling result under our non-preemptive assumption, we first demonstrate a monotonicity property in Appendix C.2 (see, Lemma 2). Here, we seek to answer the questions we raised in the Section 4.1, and generate insights into conditions under which patients should be forced to wait in the ED until a bed in their primary inpatient unit becomes available (rather than being transferred to a secondary unit with current bed availability). We start by establishing the following result.

Proposition 1 (Optimality of an Index-Based Priority Rule) *If $p_{ij} = 0$ for all $i \in N_p$ and $j \in N_s$, it is optimal for each IW to give strict priority to the patient class that has the highest $\theta_i \mu_i$ except to avoid idling, regardless of the status or allocation of other IWs.*

[9] Note that this result is only on idling when a primary patient exists, and does not mean idling IW beds cannot be optimal in general (see, e.g., Theorem 1).

Strict priority rules are typically suboptimal in non-preemptive service environments such as the one we study. Interestingly, however, Proposition 1 provides a sufficient condition under which inpatient units should give strict priority to serving the patient class with the highest $\theta\mu$ value: when reduction in quality of care is not a main concern, or similarly when the differences in service qualities between primary and secondary inpatient units are negligible. Labeling the class with the highest value of $\theta\mu$ as Class 1, this means that although care delivery of patients cannot be preempted to accommodate a new bed request, in order to merely minimize the risk of adverse events, IWs should always prioritize serving Class 1 patients when at least one such patient is boarded in the ED and the inpatient unit has some available capacity. The implication of Proposition 1 for a hospital bed manager is important and is as follows. If there is a Class 1 patient boarded in ED that is not expected to experience a reduction in quality of care from an alternative IW assignment, the bed manager should prioritize assigning him/her to a bed as soon as one becomes available in either his/her primary or secondary IW: *patience is not a virtue* in this case.

But what if in addition to the risk of adverse events (a patient safety concern), the bed manager is also concerned about the quality of care? Our numerical results suggest that the optimal policy in such a situation is a state-dependent threshold-type policy, where the threshold is on the number of patients boarded in the ED. We will discuss this in detail in the remainder of this section. However, to gain some initial analytical insights, we first focus on the patient flow to IW 1. This allows us show that when we introduce non-zero overflow penalty costs in our model, the primary unit of Class 1 patients (IW 1) prioritizes Class 1 patients under the optimal policy whenever it has some available capacity, and idles when $X_2 < \bar{X}_2$ where $\bar{X}_2$ is a threshold level. Thus, Class 2 patients should be kept boarded in the ED rather than

being overflowed to IW 1 when $X_2 < \bar{X}_2$. Hence, in this case, we find that *patience is a virtue, but only up to a point.*

Theorem 1 (Threshold-Based Idling) *There exists an optimal stationary policy which is of a threshold type: IW 1 (i) serves its secondary patients when the number of such patients boarded in the ED reaches a state-dependent threshold level and has no primary patient boarded in the ED, (ii) serves its primary patients whenever such patients are boarded in the ED, and (iii) idles otherwise.*

As is specified in Theorem 1, it is optimal to idle IW 1 when there is no Class 1 patient available and the number of Class 2 patients waiting for an inpatient bed is below a threshold level. This is due the non-zero overflow penalty, which represent the reduction in quality of care when a patient is assigned to a secondary unit. In fact, when p_{ij} is high enough, the optimal policy idles IW j whenever it does not have any primary patient boarded in the ED. For non-extreme overflow penalty cases, when IW 1 does not have a primary patient boarded in the ED, it first idles until the number of Class 2 patients boarded in the ED reaches a certain level, then prioritize Class 2 patients until either a Class 1 patient starts to board in the ED, or the number of Class 2 patients falls below the threshold. For hospital bed managers, Theorem 1 implies that when a bed becomes available in IW 1, Class 1 patients should be assigned to that IW if there are Class 1 patients boarded in the ED. Otherwise, Class 2 patients should be assigned to IW 1, but only if the number of Class 2 patients boarded in the ED is higher than a certain level. This insight is important, because it sheds light on the fact that an IW 1 bed can be left idle under the optimal policy depending on the congestion level of the ED. By idling such a bed and asking Class 2 patients to continue boarding in the ED, the hospital bed manager can avoid a potential reduction in quality of care, and also prevent a future arriving Class 1 patient from

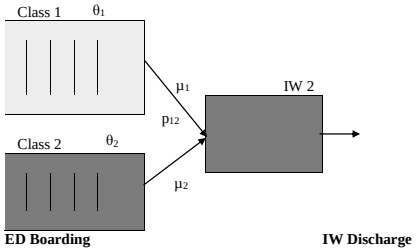


Figure 4.4: A Queueing Representation of the Simplified System

prolonged ED boarding, which in turn may have significant patient safety related consequences.

4.4.1 Patient Flow to IW 2

To gain further insights into the structure of effective patient-IW assignment policies, we now turn our attention to IW 2, and consider the simplified model illustrated in Figure 4.4. Recall that IW 2 is the primary IW for Class 2 patients, and the secondary IW for Class 1 patients, where we labeled classes (without loss of generality) such that $\theta_1\mu_1 \geq \theta_2\mu_2$. Thus, IW 2 prefers to serve Class 1 with respect to the $\theta\mu$ index, but Class 2 with respect to the overflow penalty cost parameters. As we will see, understanding the main trade-offs in this simplified model is essential for answering the questions we raised in the Introduction. Put differently, although the model presented in Figure 4.4 is a stylized version of the complex patient flow in hospitals, it allows us to gain useful insights that we can further test via realistic simulations.

We further simplify our analysis here by assuming that the service process is preemptive. [10] This allows us to consider $\underline{Y} = (Y_1, Y_2)$ as the system's state, where Y_i represents the number of Class i patients in the system, the state space is $\mathcal{S} = \mathbb{Z}_+^2$,

[10] We realize that allowing service preemption is not fully realistic; however, this assumption is useful for tractability and for gaining sharp insights. We relax this assumption in Section 4.6, and utilize real-world data along with simulation analyses to verify the insights gained.

and the set of admissible actions is:

$$\mathcal{U}(\underline{Y}) = \Big\{ u = (u_{i2})_{i\in\{1,2\}} \ \text{ s.t. : } u_{i2} \in \{0,1\}, u_{i2} \le Y_i, \sum_{i\in\{1,2\}} u_{i2} \le 1 \ \ \forall i \in \{1,2\} \Big\}. \tag{4.6}$$

Since the optimal policy and performance under long-run average setting can be obtained by using limit arguments over the infinite-horizon (see, e.g., Linn (1999)), we start by considering the system in infinite horizon. The infinite-horizon optimality equation for this simplified model can be written as:

$$J(\underline{Y}) = \underline{\theta}\,\underline{Y}^T + \beta \min_{u\in\mathcal{U}(\underline{Y})} \Bigg\{ \sum_{i\in\{1,2\}} \tilde{\lambda}_i\, J(\underline{Y} + e_i) + \sum_{i\in\{1,2\}} \tilde{\mu}_i u_{i2} \left(p_{i2} + J(\underline{Y} - e_i)\right) + \left(1 - \Lambda - \sum_{i\in\{1,2\}} \tilde{\mu}_i u_{i2}\right) J(\underline{Y}) \Bigg\}, \tag{4.7}$$

where β is the discount factor per uniformized period, the overflow penalty cost parameters p_{12} and p_{22} are scaled so that $p_{22} = 0$, and the vector $\underline{\theta}$ is scaled so that $\underline{\theta}\,\underline{Y}^T$ represents the expected cost per uniformized period when the system is at state $\underline{Y}$. Moreover, in (4.7), the uniformization rate is $\bar{\psi} = \lambda_1 + \lambda_2 + \max\{\mu_1, \mu_2\}$, where $\tilde{\mu}_i = \dfrac{\mu_i}{\bar{\psi}}$, $\tilde{\lambda}_i = \dfrac{\lambda_i}{\bar{\psi}}$, and $\Lambda = \tilde{\lambda}_1 + \tilde{\lambda}_2$. Next, we define the functional operators T_a, T_u and T_* (see, e.g, Saghafian and Veatch (2016) for the use of similar operators in a

different queueing structure) as:

$$T_\theta J(\underline{Y}) = \underline{\theta}\, \underline{Y}^T, \tag{4.8}$$

$$T_a J(\underline{Y}) = \sum_{i\in\{1,2\}} \tilde{\lambda}_i\, J(\underline{Y} + e_i), \tag{4.9}$$

$$\begin{aligned} T_u J(\underline{Y}) &= \sum_{i\in\{1,2\}} \tilde{\mu}_i u_{i2} \left(p_{i2} + J(\underline{Y} - e_i)\right), \left(1 - \Lambda - \sum_{i\in\{1,2\}} \tilde{\mu}_i u_{i2}\right) J(\underline{Y}), \\ &= (1-\Lambda) J(\underline{Y}) - \sum_{i\in\{1,2\}} \tilde{\mu}_i u_{i2} \left(\Delta_i J(\underline{Y} - e_i) - p_{i2}\right), \end{aligned} \tag{4.10}$$

$$T_* J(\underline{Y}) = \min_{u\in\mathcal{U}(\underline{Y})} T_u J(\underline{Y}), \tag{4.11}$$

$$T J(\underline{Y}) = T_\theta J(\underline{Y}) + \beta \left(T_a J(\underline{Y}) + T_* J(\underline{Y})\right), \tag{4.12}$$

where $\Delta_i J(\underline{Y}) = J(\underline{Y} + e_i) - J(\underline{Y})$. Using these functional operators, we can simply write the infinite-horizon optimality equation (4.7) as

$$J(\underline{Y}) = T J(\underline{Y}). \tag{4.13}$$

The average cost and finite-horizon cost equations can be obtained in a similar manner. Specifically, the finite-horizon cost satisfies $J_{n+1}(\underline{Y}) = T J_n(\underline{Y})$, and the average cost can be calculated as $\lim_{\beta\to 1^-}(1-\beta)J(\underline{Y})$ (see, e.g., Linn (1999) Corollary 7.5.10 for further discussion).

Using the above-mentioned setting, we next consider the following two properties for all $\underline{Y} \geq (1,1)$

$$(i)\ \ \mu_1\Delta_1 J(\underline{Y}) - \mu_2\Delta_2 J(\underline{Y} + e_1 - e_2) \geq \mu_1\Delta_1 J(\underline{Y} - e_1) - \mu_2\Delta_2 J(\underline{Y} - e_2), \tag{4.14}$$

$$(ii)\ \ \mu_1\Delta_1 J(\underline{Y} - e_1) - \mu_2\Delta_2 J(\underline{Y} - e_2) \geq \mu_1\Delta_1 J(\underline{Y} + e_2 - e_1) - \mu_2\Delta_2 J(\underline{Y}). \tag{4.15}$$

Property *(i)* implies that assigning Class 1 patients to IW 2 becomes more desirable as the number of boarded Class 1 patients increases, and property *(ii)* implies that assigning Class 2 patients to IW 2 becomes more desirable as the number of boarded

Class 2 patients increases. Let $\mathcal{F}$ be the set of real-valued functions defined on $\mathcal{S} = \mathbb{Z}_+^2$ such that if $F \in \mathcal{F}$ then F satisfies properties (4.14)-(4.15). The following lemma shows that, if $\theta_1\mu_1 \geq \theta_2\mu_2$, the functional operator T defined in (4.12) preserves properties (4.14)-(4.15).

Lemma 1 (Preservation) *If $\theta_1\mu_1 \geq \theta_2\mu_2$ and $J \in \mathcal{F}$, then $TJ \in \mathcal{F}$.*

Utilizing Lemma 1, we can establish the following result.

Theorem 2 (Optimality of a Threshold-Type Policy) *If $\theta_1\mu_1 \geq \theta_2\mu_2$, then the optimal policy obtained from (4.7) is of a threshold type: IW 2 should prioritize its primary patients until the number of Class 1 patients boarded in the ED reaches a threshold that depends on the number of Class 2 patients still waiting for a bed assignment.*

The optimal policy described in Theorem 2 is a threshold-based "primary-then-$c\mu$" rule: IW 2 serves its primary patients up to a point, and switches to the $c\mu$ rule ($\theta\mu$ in our notation) afterwards. Note that when $p_{12} = 0$, the optimal assignment policy is the well-known $c\mu$ rule (see, e.g., Buyukkoc *et al.* (1985) and Saghafian and Veatch (2016)), because the threshold becomes zero. However, when we consider a non-zero penalty cost in the model, under the optimal policy, IW 2 first serves its primary patients until the marginal benefit of serving a primary patient versus a secondary one reaches the value of the penalty that might be accrued due to the reduction in quality of care. This suggests that, when the number of boarded patients is low, EDs should try to match their patients with their primary units to ensure the highest quality of care. However, once the number of boarded patients passes a specific threshold, the focus should shift from concerns about decrements in quality of care to concerns about the risk of adverse events that can occur due to prolonged boarding.

Thus, we again observe that *patience (for a primary unit assignment) is a virtue, but only up to a point.*

Hospital bed managers can use our results in various ways when deciding on which patient class to assign to an IW 2 bed that has just become available. For instance, when a bed becomes available in IW 2, and they do no expect any near-term bed availability in IW 1, Theorem 2 suggests that bed managers should consider the number of both Class 1 and 2 patients boarded in ED and prioritize the primary patient type (Class 2) until the number of Class 1 patients boarded in ED reaches a certain level. From then on, they should start prioritizing Class 1 patients until the number of Class 1 patients boarded in ED drops below that certain level. However, the bed manager should be aware that this level is highly dependent on the number of patients from both classes in the ED as well as estimation of parameters related to (a) reduction in quality of care when a secondary inpatient unit is used (p_{ij}), (b) risk of adverse events for both classes (θ_i), and (c) average length of stay for both classes ($\frac{1}{\mu_i}$). Thus, the decision should be made in a careful way and only after performing sensitivity analysis. To further assist hospital bed managers in making such decisions, we utilize the insights we gained from analyzing the optimal policy of our simplified models, and develop effective bed assignment heuristics in the next section. We then use a variety of simulation experiments (calibrated with hospital data that we have collected) to evaluate their effectiveness under realistic conditions, and generate more detailed insights for hospital administrators via sensitivity analyses.

4.5 Heuristic Policies

When we consider non-preemptive service policies (which better represent the current practice in most hospitals) under the general system structure discussed in Section 4.3, our numerical computations show that the optimal policy is complex: it

has a state-dependent threshold that depends on all the elements in the system state, including IW bed availabilities. Our numerical results also show that the optimal policy has a structure similar to the optimal control of the "N" structure queueing network, where one server works as a *shared* server while the other works as a *dedicated* server. That is, the primary unit of Class 1 patients (IW 1) typically prioritizes its primary patients (i.e., works as a dedicated unit whenever its queue of boarded patients is not empty), and primary unit of Class 2 patients (IW 2) typically first serves its primary patients until the number of Class 1 patients boarded in ED exceeds a threshold, then helps IW 1 by serving Class 1 patients (see Appendix C.2 for some numerical experiments supporting this observation). In what follows, we take advantage of this (as well as our earlier findings) to develop easy-to-implement heuristic policies for use in hospitals.

4.5.1 A Birth-and-Death Process to Approximate the Optimal Threshold

To develop a heuristic that is easy to implement, we start by considering the optimal policy of an "N" queueing network by assuming that IW 2 can serve patients from both types (a shared server) while IW 1 can only serve Class 1 patients (a dedicated server). We use a birth-and-death process for this system to estimate the optimal threshold level on the number of Class 1 patients boarded in the ED above which IW 2 starts helping IW 1 by serving Class 1 patients. In particular, assuming that the threshold level is some number T, we can approximate the Class 1 queueing dynamics via the birth-and-death process depicted in Figure C.6 (see Appendix C.3). When the number of patients in the Class 1 queue, X_1, is smaller than the threshold level, only IW 1 will serve Class 1 patients which will occur with rate μ_1. However, when X_1 is larger than the threshold, both IW 1 and IW 2 will serve Class 1 patients, and hence, the death rate becomes $2\mu_1$.

We use a separate birth-and-death process to approximate the dynamics for Class 2 patients (see Figure C.7 in Appendix C.3). Let $P^2(T)$ be the steady-state fraction of time that IW 2 serves Class 2 patients. Then, the service rate for Class 2 patients is $P^2(T)\mu_2$. Let $L^1(T)$ and $L^2(T)$ denote the long-run average queue length (i.e., number of patients boarded in the ED) of Class 1 and Class 2 patients, respectively. Assuming that $O^1(T)$ denotes the average number of Class 1 patients served by IW 2, and $Z(T)$ denotes the long-run average system cost under threshold level T, we can calculate $Z(T)$ as:

$$Z(T) = \theta_1 L^1(T) + \theta_2 L^2(T) + p_{12} O^1(T). \tag{4.16}$$

The objective is to find the value of T that minimizes $Z(T)$. To calculate (4.16), we use the above-mentioned birth-and-death processes to estimate $L^1(T)$, $L^2(T)$, and $O_1(T)$. To this end, we first need to obtain the steady-state probability P_i^j which is the probability that the length of queue $j \in \{1, 2\}$ equals to $i \geq 0$. From the balance equations, we have:

$$P_i^1 = (\frac{\lambda_1}{\mu_1})^i P_0^1 \qquad \forall \ i \leq T, \tag{4.17}$$

$$P_i^1 = (\frac{\lambda_1}{\mu_1})^T (\frac{\lambda_1}{2\mu_1})^{i-T} P_0^1, \quad \forall \ i > T. \tag{4.18}$$

By using the fact that these probabilities must sum to 1, we find P_0^1 as:

$$P_0^1(T) = \frac{(1-\rho_1)(1-\rho_2)}{\rho_1^T(\rho_2 - \rho_1) + (1-\rho_2)}, \tag{4.19}$$

where $\rho_1 = \dfrac{\lambda_1}{\mu_1}$ and $\rho_2 = \dfrac{\lambda_1}{2\mu_1}$. By using these probabilities, we can obtain the average queue length for Class 1 patients, $L^1(T)$:

$$L^1(T) = \sum_{i=0}^{T} i\rho_1^i P_0^1(T) + \sum_{i=T+1}^{\infty} i\rho_1^T \rho_2^{i-T} P_0^1(T). \tag{4.20}$$

Also, $O^1(T) = \frac{1}{2}\sum_{i=T+1}^{\infty} i\rho_1^T \rho_2^{i-T} P_0^1(T)$ by assuming that Class 1 patients in the queue

will be served equally by IW 1 and IW 2 after the number of Class 1 patients boarded in ED reaches the threshold. [11] To calculate the average queue length of Class 2 patients, $L_2(T)$, we first calculate the following:

$$P^2(T) = P(x_1 \leq T) = P_0^1(T)\frac{1-\rho_1^{T+1}}{1-\rho_1}. \tag{4.21}$$

The average queue length for Class 2 patients is then:

$$L^2(T) = \frac{\lambda_2}{P^2(T)\mu_2 - \lambda_2}. \tag{4.22}$$

These allow us to calculate $Z(T)$ via (4.16), and find the optimal threshold value $T^* = \arg\min_{T\geq 0} Z(T)$. However, the threshold level T^* does not have a closed-form solution, and the function $Z(T)$ can be non-convex in general. Nevertheless, we can utilize numerical approaches (e.g., bisection search) to find the value that minimizes (4.16). We term the heuristic policy that controls the patient flow based on this threshold as the *birth-and-death threshold (BDT)* policy.

4.5.2 Penalty-Adjusted Largest Expected Workload Cost Policy (LEWC-p)

Our results in Section 4.4 reveal that there exists a threshold type optimal policy that optimizes performance by following the primary-then-$c\mu$ rule (see, e.g., Theorem 2). This policy tends to serve the primary patient type with the lower $c\mu$ value until the cost differences of serving the secondary patients exceeds the overflow penalty cost (see the discussion in Appendix C.2, proof of Lemma 1). This insight suggests that instead of using a heuristic policy to directly approximate the threshold—the idea behind the BDT policy—there might be value in following a heuristic that balances the costs associated with different queues. Thus, as our second heuristics, we develop

[11] This is not a strong assumption, because the service rates are patient class dependent not IW dependent.

a modified version of the Largest Expected Workload Cost (LEWC) policy proposed by Saghafian *et al.* (2011) for general parallel queueing systems. The LEWC policy dynamically balances the expected workload cost of queues by prioritizing the queue with the largest expected workload cost (ROAE in our setting). [12] In order to also incorporate the additional penalty cost of serving patients in their secondary IW—a main factor for the patient flow focus of this study— we propose a penalty-adjusted version of LEWC, which we term *LEWC-p*. To this end, similar to Saghafian *et al.* (2011), we first use the following Linear Program (LP). In this LP, the objective is to find the optimal server allocations to maximize the minimum percentage excess capacity among all patient types:

$$\text{Max} \quad \tau \tag{4.23}$$

Subject to:

$$\sum_{j \in N_s} y_{ij}\mu_i \geq \lambda_i(1+\tau) \qquad \forall i \in N_p, \tag{4.24}$$

$$\sum_{i \in N_p} y_{ij} \leq 1 \qquad \forall j \in N_s, \tag{4.25}$$

$$y_{ij} \geq 0 \qquad \forall i \in N_p, \forall j \in N_s. \tag{4.26}$$

In this LP, y_{ij} is the decision variable that represents the long-run proportion of time that IW j serves patient class i. Constraint (4.24) ensures that the objective function maximize the minimum excess capacity among all patient classes. Constraint (4.25) guarantees that the total proportion of time for each IW does not exceed 1, and Constraint (4.26) enforces the proportions to be non-negative.

Next, when a bed in IW j becomes available, we calculate an index, $I_{ij}(x_i)$, for each queue $i \in N_p$ (class of patients boarded in the ED) to approximate the penalty-

[12] LEWC is a dynamic policy, because it prescribes different actions based on the system state.

adjusted expected workload cost of that queue given that its current length is x_i:

$$I_{ij}(x_i) = \frac{\theta_i x_i}{\sum_{j\in N_s} y^*_{ij}\mu_i} - p_{ij}\frac{x_i y^*_{ij}}{\sum_{j\in N_s} y^*_{ij}}, \tag{4.27}$$

where y^*_{ij}'s are the solution to LP (4.23)-(4.26). The first part of the index approximates the cost associated with risk of adverse events for class i patients: since there are x_i patients in the queue, it will take approximately $\frac{x_i}{\sum_{j\in N_s} y^*_{ij}\mu_i}$ units of time to serve them, and the cost due to adverse events is θ_i per unit of time per patient boarded. The second part of the index approximates the associated penalty cost. In this term, $\frac{y^*_{ij}}{\sum_{j\in N_s} y^*_{ij}}$ represents the proportion of patients of class i served by IW j. [13]

With these, the penalty-adjusted LEWC policy (LEWC-p) is as follows:

1. Solve LP (4.23)-(4.26) to derive optimal allocations y^*_{ij}.

2. Whenever a patient arrives or IW j becomes available, compute indices $I_{ij}(x_i)$ for all patient classes ($i \in N_p$), then assign the bed to patient class $k = \arg\max_{i\in N_p} I_{ij}(x_i)$. If the primary and secondary queues of IW j have the same index, break the tie by assigning the bed to the primary queue. If the primary queue of IW j is empty, and its secondary queue has a negative index, keep the bed in IW j idle.

4.5.3 Comparison of the Proposed Heuristic Policies

We now compare the performance of the proposed BDT and LEWC-p heuristic policies with the optimal policy. As a benchmark, we also use the generalized $c\mu$

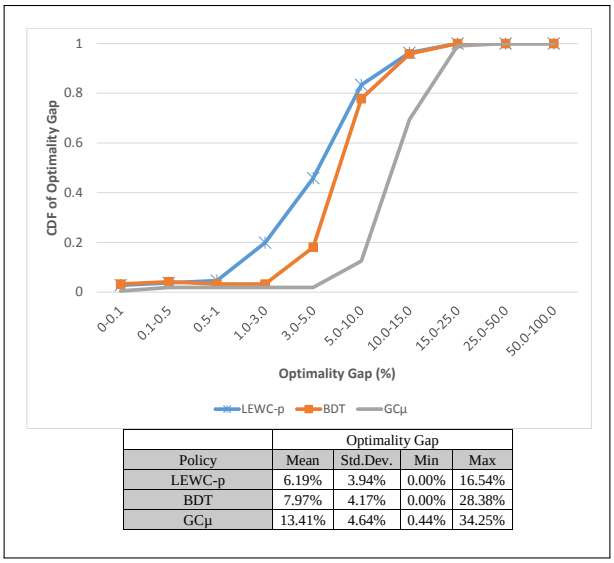
[13] In using (4.27), we assume that LP (4.23)-(4.26) has a unique optimal solution with $y^*_{ij} \neq 0$ whenever $i \neq j$. For systems in which this solution is not unique (e.g., balanced systems where $\frac{\lambda_i}{\mu_i} = \kappa, \quad \forall i \in N_p$) ties need to be broken based on cost parameters.

(G$c\mu$) rule. Under the G$c\mu$ policy, the available bed in IW j is assigned to the class that has the highest $\theta_i \mu_i x_i$ value. We use this policy as a benchmark since it (a) takes the queue lengths into account, and (b) is known to work well in a variety of queueing systems. [14]

To compare these policies (BDT, LEWC-p, and G$c\mu$), we create a large test suite which covers various combinations of parameters (e.g., costs associated with risk of adverse events and reduction in quality of care, arrival rates, service rates, etc.). Tables C.4-C.6 in Appendix C.4 summarize the parameter combinations in this test suite, which generate a total of 216 problem instances. To find the optimal policy for each problem instance, we use the well-known value-iteration algorithm to solve our MDP formulation. This allows us to report optimality gaps for each of the policies under consideration.

Figure 4.5 illustrates our computational results over the test suite by constructing the empirical Cumulative Distribution Function (CDF) for the percentage optimality gap of each of the non-optimal policies (BDT, LEWC-p, and G$c\mu$). The results presented in this figure show that LEWC-p and BDT policies can both be considered as "nearly-optimal" policies. However, the mean and standard deviation of LEWC-p optimality gap is smaller than that of the BDT policy, so we can conclude that it is the better policy. The performance of $Gc\mu$ is, however, significantly worse than both the LEWC-p and BDT policies. This is mainly because $Gc\mu$ does not consider penalties associated with secondary unit assignments. However, even when the underlying penalty parameter is zero, we observe that $Gc\mu$ is not the best policy for all cases. When the penalty parameter is zero, both of the proposed heuristic policies (BDT and LEWC-p) perform close to each other while BDT performs slightly better due to

[14] This is especially the case in systems with quadratic holding costs and in systems that face heavy traffic. Our system does not meet any of these conditions. However, we still use the optimality gap of the (G$c\mu$) rule to better gage the optimality gap of our proposed heuristics.



	Optimality Gap			
Policy	Mean	Std.Dev.	Min	Max
LEWC-p	6.19%	3.94%	0.00%	16.54%
BDT	7.97%	4.17%	0.00%	28.38%
GCμ	13.41%	4.64%	0.44%	34.25%

Figure 4.5: Performance of LEWC-p, BDT, and $Gc\mu$ Relative to the Optimal Policy

the assumption that IW 1 only serves Class 1 patients (under the $c\mu$ policy both of the IWs serve Class 1 whenever feasible).

Table 4.2 compares the optimality gap of LEWC-p, BDT, and $Gc\mu$ policies for various congestion levels in the system. All of the policies show a smaller mean optimality gap in moderate to high congestion levels than in the low congestion level. This observation suggests that implementing them in crowded systems (e.g., in busy teaching hospitals) is better than doing so in less crowded systems (e.g., in less busy urban hospitals). Finally, Table 4.3 compares the policies based on various penalty parameter settings and shows that all policies perform best when the underlying penalty parameter is high. Moreover, LEWC-p is more robust than the BDT policy to changes in the penalty parameter. This is intuitive, since the BDT policy only uses one threshold level, while the LEWC-p policy dynamically adjusts the assignments based on the number of patients of different classes that are boarded in the ED.

Table 4.2: Optimality Gap of Policies for Various Congestion Levels

Congestion Level	Policy	Mean	Min	Max
Low: $\rho \leq 0.5$	LEWC-p	7.01 %	0.00 %	14.03 %
	BDT	8.66 %	0.00 %	28.38 %
	Gcμ	16.90 %	11.90 %	34.25 %
Moderate: $\rho = 0.7$	LEWC-p	5.83 %	1.29 %	16.54 %
	BDT	8.34 %	2.61%	15.47 %
	Gcμ	13.47 %	8.66 %	17.17 %
High: $\rho \geq 0.9$	LEWC-p	5.68 %	1.74 %	9.20 %
	BDT	7.07 %	3.82 %	14.42 %
	Gcμ	10.49 %	0.09 %	15.28 %

4.6 Simulation Analysis Using Hospital Data

To gain more insights into effective policies for assigning ED patients to their primary or secondary inpatient units, we use a discrete-event simulation model of ED patient flow, and calibrate it with a year of hospital data that we have from our partner hospital. This enables us to relax some of the assumptions we made earlier (e.g., exponential service times, Poisson arrivals, etc.), and also shed light on the magnitude of achievable benefits for EDs as well as hospital conditions under which our proposed assignment policy (LEWC-p) will work well. To this end, we first describe the admission sources in our partner hospital. We then describe the arrival process from such sources. Finally, we discuss the service process as well as the empirical length of stay (LOS) distributions and other parameters that we have estimated from our data set.

Table 4.3: Optimality Gap of Policies for Various Penalty Cost Parameters

Penalty Cost	Policy	Mean	Min	Max
Low: $p_{12} = p_{21}$=1	LEWC-p	7.02 %	1.29 %	16.54 %
	BDT	9.87 %	6.00 %	28.38%
	Gcμ	14.37 %	8.85 %	21.78 %
Moderate: $p_{12} = p_{21}$=10	LEWC-p	5.82 %	0.00 %	15.06 %
	BDT	6.17 %	0.00%	15.73 %
	Gcμ	12.92 %	0.09 %	22.40 %
High: $p_{12} = p_{21}$=100	LEWC-p	4.68 %	0.00 %	13.24 %
	BDT	4.68 %	0.00 %	13.92 %
	Gcμ	11.73 %	0.44 %	34.25 %
Low-High: $p_{12} = 1$, p_{21}=100	LEWC-p	4.55 %	2.78 %	4.73 %
	BDT	5.21 %	4.59 %	7.24 %
	Gcμ	11.68 %	10.49 %	13.05 %
High-Low: $p_{12} = 100$, p_{21}=1	LEWC-p	7.96 %	5.74 %	10.15 %
	BDT	10.22 %	8.74 %	12.53 %
	Gcμ	15.42 %	11.14 %	16.41 %

4.6.1 Patient Flow and IWs in Our Partner Hospital

Admission Sources. Patients are admitted to IWs from three main sources. We categorize admitted patients based on their source of admission in three groups: *ED admits*, *direct admits*, and *Operating Room (OR) admits*. ED admits are patients who finish their treatment with ED and receive an admit decision from an ED physician. Direct admits are the ones directly admitted to an IW without any preceding visits. OR admits are the patients who initially receive a surgery from the hospital and are

subsequently admitted to an IW.

IWs. Patients from the three admission sources described above require a bed from one of the eight inpatient units based on their diagnosis. The name of IWs, their descriptions, and number of beds in each of them in our partner hospital can be found in Table 4.1 (see Section 4.1).

Patient Types. To gain clear insights into effective assignment policies, we focus on patients who were admitted via the ED of our partner hospital with an admission diagnosis of either *chest pain* (CP) or *congestive heart failure* (CHF). These patients are often assigned to a secondary IW; the primary IW for both CP and CHF patients is 4 West (4W), and their secondary IW is 5 West (5W) (see Table 4.1 for more information regarding these IWs). There are two types of CP and CHF patients: Type 1 patients are those considered to be more sensitive to a secondary bed assignment (i.e., are subject to higher reduction in quality of care if assigned to a secondary inpatient unit). Type 2 patients are those who are less sensitive to a secondary bed assignment. We develop a classification scheme using simple laboratory findings and based on our discussions with medical experts at our partner hospital. We define Type 1 CP patients as those who have an elevated serum *troponin* (Tn) level, and Type 2 CP patients as those who have a normal troponin level. We define Type 1 CHF patients as those who have a *B-type natriuretic peptide* (BNP) level of 4,000 pg/ml or greater, and Type 2 CHF patients as those with BNP levels below 4,000 pg/ml. Our empirical analyses show that, among patients of same type, there is no statistically significant difference in the mean IW service time between primary and secondary units (see Table C.7 in Appendix C.5).

Arrival Process. We use bed-request times as the "arrival" time of each patient to our system. We observe from our data set that, for each of the three arrival sources (ED admit, direct admit, OR admit), the arrival rate is highly time-dependent. Fur-

thermore, we observe that the arrival process for each arrival source and for each IW can be modeled as a nonhomogeneous Poisson Process with a rate that is constant during one-hour time blocks. In addition to hour-of-day dependent arrival rates, we observe day-of-week dependency in arrival rates for ED admits. We simulate the patient flow assuming that the arrival process is cyclo-stationary with one week as the cycle. We do not consider the rare transfers between inpatient units, since (a) our focus in this chapter is on the patient flow between ED and IWs, (b) these transfers do not have any significant effect on the optimal policy, and (c) based on our data set, the rate of such transfers is negligible compared to the arrival rate of ED admits, direct admits, and OR admits.

Service Process. In our simulation model, we consider the beds in IWs as servers. Based on our data, the service rates depend on patient type and admission source but not the IW (see Table C.7 in Appendix C.5 for p-values on the equality of means of service times for primary and secondary IWs for different patient types). Table C.8 in Appendix C.5 shows the average service time (in days) for each IW based on the admission source. Our statistical analyses suggest that we can use lognormal distributions as service time distributions. [15]

Costs. Penalty costs are assigned based on the patient type (Type 1 and Type 2 discussed above). The average penalty cost for Type 1 patients are always higher than that of Type 2 patients, since Type 1 patients are more sensitive to a secondary bed assignment. However, due to current lack of data on quality of care and patient safety, estimating cost parameters is inherently subject to error, and necessitates performing various sensitivity analyses. To perform such sensitivity analyses, we consider a wide range of parameters for both penalty costs and costs associated with risk of

[15] Lognormal distribution as a service time distribution is not unique to hospitals. For instance, Brown *et al.* (2005) show similar characteristic of the service time distribution in call centers.

adverse events (see Appendix C.6 for more information). This range of parameters are provided by our physician collaborators, and are intended to represent values that are realistic while covering possible differences among hospitals.

Performance Measures. In addition to the overall objective we introduced in Section 4.3, we use the overflow proportion (the ratio of patients assigned to a secondary IW to the total number of patients of same type served) and the average ED boarding time (the average time between a request and bed occupancy) as other performance measures. We also use the 2-hour boarding rate (the fraction of patients that are boarded for two hours or more) [16] as another performance measure. We do so because reducing excessive boarding times (and not just average boarding times) is also important for most EDs.

Priorities and Runs. We use the first-in-first-out (FIFO) priority rule for each IW regardless of the admission source of patients. Each simulation observation is obtained for 1,000 replications with a replication length of one year. The number of replications is chosen so as to enforce tight confidence intervals, enabling us to represent simulation confidence intervals with their midpoint in all of our graphs. This warm-up period is determined through the Welch method (see, e.g., Welch (1983)).

Base Case Scenario. We consider the base case scenario to be a reflection of the current system in our partner hospital based on a year of data that we have collected. We use this scenario as a benchmark to analyze the potential changes that may occur due to implementing our proposed policies. Thus, we use the level of performance measures in the base case scenario (e.g., 2-hour boarding rate, average boarding time in the ED, etc.) for CP and CHF patients as a point of reference, and compare the results of our proposed policy with those metrics. To this end, we focus on patient

[16] As we discussed in Section 4.1, the current 2-hour boarding rate at our partner hospital based on our data set is around 30%.

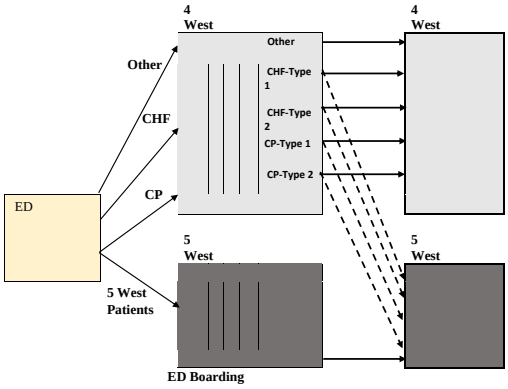
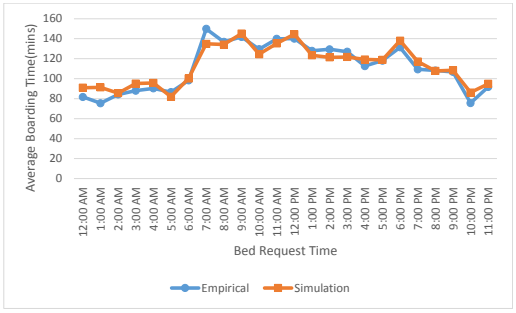


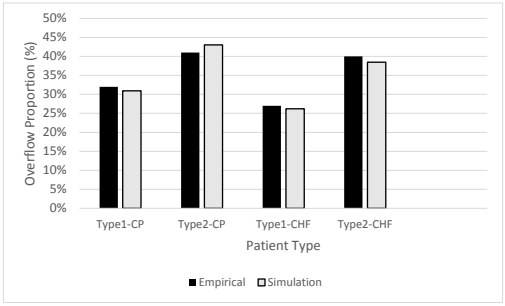
Figure 4.6: Patient Flow in the Simulation Model

flow from ED to the two IWs that can serve CP and CHF patients: 4 West and 5 West. In addition to CP and CHF patients, we simulate the flow of other patients that require a bed from 4 West or 5 West, but note that these patients are not eligible for overflows, and can only be assigned to their primary inpatient units. We include these patients in our simulation model to represent the capacity utilization in 4 West and 5 West more accurately, thereby increasing the fidelity of our simulations. Figure 4.6 illustrates the patient flow under consideration. [17] The dashed lines in Figure 4.6 show assignments of patients to secondary IWs (overflows that incur a penalty cost) while the solid lines show assignments to primary IWs. In the current practice, there is no specific rule for assigning patients to their primary vs secondary units. Thus, for our base case scenario, we use the FIFO rule for the primary bed assignments, and model the overflows to secondary IWs by using the proportions that are obtained from our data analyses.

[17] In Appendix C.7, we extend our simulation analysis to the whole patient flow depicted in Figure 4.3 with all the 8 IWs listed in Table 4.1. However, since this requires estimating various parameters for each and every patient type served in the hospital, our simulations lose fidelity. Thus, here we stay with CP and CHF patients (i.e., patients for which we have more accurate data).



(a) Hourly Average ED Boarding Time



(b) Overflow Proportion

Figure 4.7: Validating the Simulation Model

4.6.2 Validating the Simulation Model

To validate our simulation model, we compare our empirical results obtained directly from our data set with those obtained from our simulation model. Figures 4.7(a) and 4.7(b) compare the resulting time-dependent boarding time of patients as well as the resulted overflow rates of the simulation model with that of the empirical data. Using the t-test for the equality of means, we observe no statistical difference between outputs of our simulation model and those from empirical data (p-value = 0.412). Similarly, using Kolmogorov-Smirnov tests for comparing the distributions of outputs (e.g., boarding time distributions) with the empirical distributions from our data, we do not observe any significant mismatch. These results give us confidence that our simulation model is relatively of high fidelity, and accurately matches the current practice.

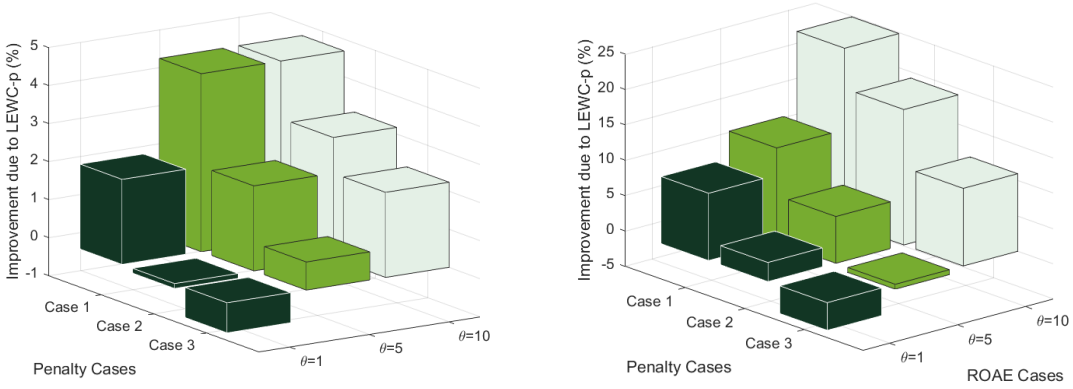
4.6.3 Performance of the Proposed LEWC-p Policy

We now use our simulation model for CP and CHF patients to investigate the impact of implementing our proposed LEWC-p policy. Based on our results, we make the following observation:

Observation 1 (Benefits of LEWC-p) *Implementing LEWC-p for assigning CP and CHF patients to their IWs improves the total average cost by* 14%*, the 2-hour boarding rate by* 2%*, and the average boarding time by* 9% *(10 minutes/patient). Also, compared to current practice, these improvements due to implementing LEWC-p are all statistically significant (the p-value on the difference is 0.00018, 0.022, and 0.001, respectively).*

We next test the sensitivity of the gained benefits to the penalty costs and costs associated with adverse events. As we increase the latter, the improvement in the 2-hour boarding rate and the average ED boarding time increases (see Figures 4.8(a) and 4.8(b)). Furthermore, we observe that as we increase the penalty cost, IWs start to work as dedicated units tending to only serve their primary patients. Hence, after increasing the penalty cost, we observe improvements in overflow proportions, but the average ED boarding time and the costs associated with adverse events increase. This result is similar to what we observed from the optimal policy of the analytical model: as we increase (decrease) the penalty cost, the LEWC-p policy mimics the optimal policy by decreasing (increasing) the assignments to secondary IWs. Similarly, as we increase (decrease) costs associated with adverse events that may occur during ED boarding, the LEWC-p policy mimics the optimal policy by increasing (decreasing) the assignments to secondary IWs.

Figures 4.9(a) and 4.9(b) illustrate the change in total number of boarded patients in the ED and overflow proportion as ROAE and penalty cost parameters change under the LEWC-p policy. As the ROAE cost increases, the proposed policy starts to assign patients to their secondary IW more aggressively. This leads to lower average ED boarding times, and suggests that utilizing a secondary IW is a more attractive option for patients who have a higher ROAE (e.g., those in need of timely care following their ED service). Another implication of Figures 4.9(a) and 4.9(b) is that



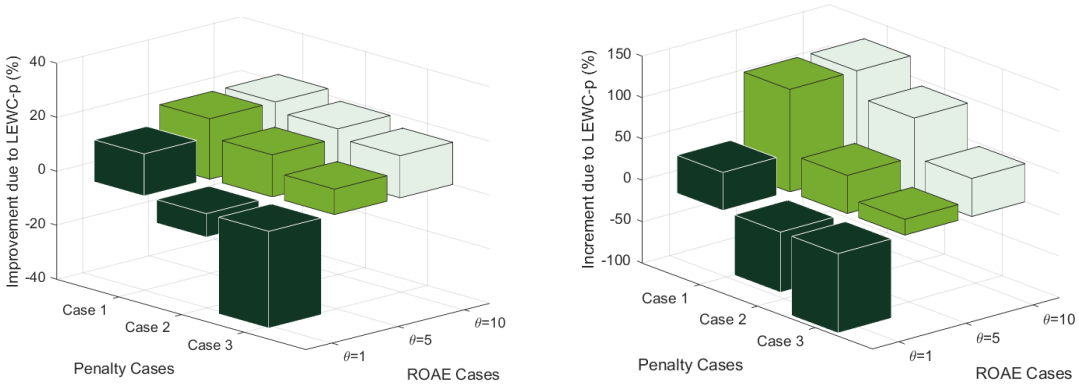
(a) Improvement in 2-Hour Boarding Rate (b) Improvement in Average Boarding Time

Figure 4.8: Improvement Due to LEWC-p Compared to Current Practice for Various Penalty and ROAE Parameters

assigning patients to their secondary IWs has a minimal effect on the average ED boarding time when the penalty cost parameter is high. This suggests that hospital administrators should be more patient in assigning beds for patients who are more sensitive to a secondary IW assignment (e.g. Type 1 patients as opposed to Type 2 patients): *the virtue of patience is dependent on patient type.*

Our proposed policy allows idling IW beds (in anticipation of future needs) even when there are patients boarded in the ED who need them. However, hospital beds are valuable assets, and keeping them idle while patients are waiting for them might not be perceived as attractive by hospital administrators. To gain some insights into the impact of idling, we modify our policy by assigning 4 West patients to 5 West when there is no 5 West patient boarded in the ED (disallowing idling of 5 West beds). From our results on the performance of LEWC-p policy with and without idling, we can make the following observation:

Observation 2 (Nonidling Policy) *Non-idling flow policies increase the number*



(a) Average Number of Patients Boarded in ED

(b) Overflow Proportion

Figure 4.9: The Effect of ROAE and Penalty Cost Parameters on the Average Number of Patients Boarded and Overflow Proportion Due to LEWC-p

of patients overflowed, but does not significantly change the average number of patients boarded in the ED, the average boarding time, and the 2-hour boarding rate.

The above observation captures one of the most fundamental trade-offs in our study. Prohibiting idling 5 West beds increases the number of 4 West patients assigned to 5 West while reducing the number of 4 West patients boarded in ED who are eligible for a secondary unit assignment. However, these assignments result in blocking the access of future arriving 5 West patients to 5 West beds, which increases the number of 5 West patients boarded in ED. As a result, the average number of patients boarded in ED, the average boarding time, and the 2-hour boarding rate do not change significantly. These outcomes contradict the prevalent perception among hospital administrators that beds should not be idled intentionally. We note that this perception might be correct when the ROAE among different patient groups (in our

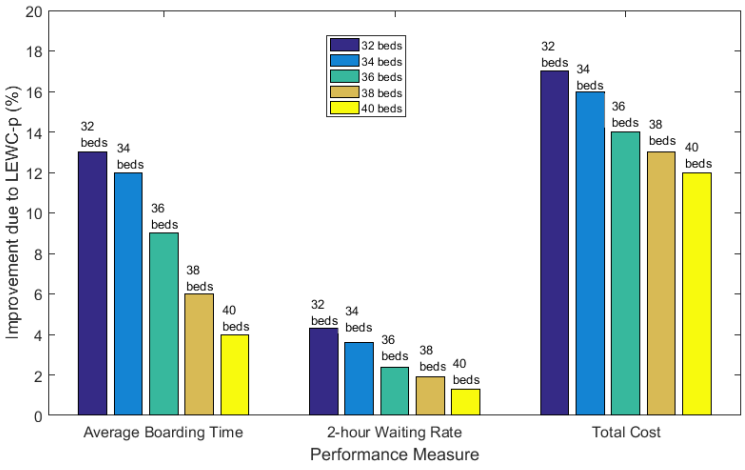


Figure 4.10: Effect of Inpatient Bed Capacity on the Improvements Due to LEWC-p

case, 4 West and 5 West patients) is significantly different. [18] However, our results suggest that hospitals should typically refrain from prohibiting idling: *idling IW beds can be beneficial*.

In our previous simulation experiments, we used the current bed capacity of IWs in our partner hospital (see, Table 4.1). To gain more insights for other hospitals which might have higher or lower capacities, we now provide sensitivity analysis by altering the number of beds in IWs (both 4 West and 5 West). Figure 4.10 illustrates the results, and enables us to make the following observation:

Observation 3 (Effect of Inpatient Bed Capacity) *The achievable improvements due to implementing LEWC-p on the performance measures are greater in hospitals with lower inpatient bed capacity (all else equal).*

This observation suggests that hospitals that lack enough inpatient bed capacity (e.g., busy teaching hospitals) will benefit more from implementing the LEWC-p

[18] If the ROAE for eligible 4 West patients is much larger than that of the 5 West patients, prohibiting idling can be beneficial in terms of the total average cost metric and average boarding time for 4 West patients.

policy. Thus, instead of investing in increasing their capacity—a challenging and extraordinarily expensive undertaking that often requires a certification of need—they can benefit from better bed assignment policies such as LEWC-p, which requires only a minimal investment.

Another related issue in understanding the effect of inpatient bed capacity is the practice of "bed reservation." Unlike our partner hospital, some hospitals reserve a portion of their IW capacity for their primary patients so as to reduce the effect of overflows on those patients. It is clear that as the number of beds usable for overflows decreases, the number of patients that are assigned to a secondary IW decreases, which in turn results in a lower total penalty cost. However, the impact of this practice on other performance measures such as the average boarding time, the average number of patients boarded, and the 2-hour boarding is not obvious. To observe the effect of this practice, we consider three cases by assuming that only 25% (9 beds), 50% (18 beds), and 75% (27 beds) of the beds in 5 West can be used for accommodating CP and CHF patients. From this analysis, we make the following observation:

Observation 4 (Effect of Restricting Bed Capacity) *Under the proposed LEWC-p policy, restricting the bed capacity for overflow patients significantly increases the average boarding time, the average number of patients boarded, and the 2-hour boarding rate, while decreasing the number of overflows. However, the relative impact of this practice is not statistically significant between cases with* 25% *and* 50% *restriction, or with* 50% *and* 75% *restriction.*

Our results suggest that, when the number of beds usable for overflows decreases, the number of overflows decreases (as expected). Since our proposed policy captures the trade-off between the ROAE and the quality of care, reduction in overflows leads to an increase in the number of patients boarded in the ED. However, the changes in

performance measures are not significant when we either drop to 25% bed capacity from 50% bed capacity, or drop to 50% bed capacity from 75% bed capacity. For hospitals with similar characteristics to our partner hospital (in terms of bed request arrivals, inpatient LOS, etc.) this suggests that, to make a statistically significant impact on the performance measures, hospital administrators should consider dramatic changes in the number of beds to be used for overflows.

Overflow trigger times are often used in practice (see, e.g., Shi *et al.* (2015)) where a patient is overflowed to a secondary IW only when the boarding time of the patient exceeds a predetermined trigger time. We next investigate how our proposed policy performs when the hospital employs an overflow trigger time. To this end, we assume that a patient can be overflowed either when his/her boarding time exceeds the trigger time, or when the LEWC-p policy assigns him/her to a secondary IW. We analyze the performance of this modified policy by considering various trigger times.

Observation 5 (Overflow Trigger Time) *Imposing overflow trigger times typically increases the penalty costs accrued due to lower levels of quality of care. However, regardless of the level of the trigger time, the relative improvement in the costs associated with adverse events is not high enough to yield an overall improvement in the aggregate cost measure. In addition, the impact of imposing a trigger time on the average boarding time and the 2-hour boarding rate is significant only when the trigger time is no more than two hours.*

This observation suggests that imposing an overflow trigger time that is higher than two hours does not change the performance of our proposed LEWC-p policy. In fact, using a trigger time that is more than two hours typically adds complexity in assignment decisions without any significant change in performance measures. As we noted earlier, our proposed LEWC-p policy improves the average boarding time

by approximately 10 minutes per patient compared to the current practice. This results in approximately 100 minutes of average boarding time and 29% of 2-hour boarding rate in the improved system. Setting a trigger time that is lower than two hours can affect more than 70% of the patient population (since 2-hour boarding rate is 29%), which may result in improvements in the average boarding time and the 2-hour boarding rate. However, we find that adding trigger times to LEWC-p does not lead to improvements in the aggregate cost measure, regardless of the level of the trigger time. This further confirms that the proposed LEWC-p policy already strikes a strong balance between concerns related to prolonged ED boarding times and those related to overflows.

4.7 Conclusion

We study the dynamic assignment of ED admitted patients to hospital IWs. We utilize a queueing framework and an MDP model to gain insights into effective mechanisms to minimize the risk of adverse events (a patient safety concern) while reducing the number of secondary inpatient unit assignments (a quality of care concern).

Our results for a simplified model with two patient classes and two IWs suggest that the optimal policy is a threshold-type policy, where the threshold depends on the number of patients boarded in the ED. Under this policy, the primary unit of Class 1 patients (i.e., patients that have a higher $\theta\mu$ value) typically works as a dedicated unit that serves its primary patients whenever such a patient is boarded in the ED. Moreover, the primary unit of Class 2 patients serves them before helping IW 1 on Class 1 patients, and switches to serving Class 1 patients once the number of Class 1 patients boarded in the ED reaches a threshold. These suggest that patience in transferring ED admitted patients to IWs is a virtue, but only up to a point. Contrary to the prevalent perception among hospital administrators, we also find that idling

IW beds can be beneficial. In particular, while idling is used in some hospitals and some specific inpatient units, our results indicate evidence for wider implementation of idling policies. We also show that, when the penalties that represent the reduction in quality of care in secondary units are negligible, the optimal policy is a strict priority rule in which both IWs prioritize serving Class 1 patients in order to myopically decrease the risk of adverse of events for patients boarded in the ED.

Our analyses show that the optimal policy is complex in general, and may not be suitable for implementation in practice. Therefore, we use the insights gained from analyzing our simplified models to develop two heuristic policies that are easy to implement. We first use a birth-and-death process to approximate the threshold level that minimizes an aggregate measure of both patient safety and quality of care. Then, we propose a modified version of the LEWC heuristic termed LEWC-p that enables us to dynamically strike a balance between concerns of patient safety and quality of care. The results show that LEWC-p significantly outperforms other policies, and is also more robust than them in that it has a lower standard deviation of optimality gap. Thus, an important contribution of this study is to introduce LEWC-p as a simple but effective policy that can be implemented in hospitals.

We then investigate the achievable gains due to implementing LEWC-p by using a simulation model that we calibrated with a year of data collected from our partner hospital. By using this simulation model, we are able to reflect the realistic features of the hospital patient flow, and test the insights gained from our analytical models. To gain clear results, we focus on chest pain (CP) and congestive heart failure (CHF) patients. Furthermore, by utilizing laboratory findings to separate patients based on the level of Tn for CP patients and BNP for CHF patients, we classify these patients as Type 1 and Type 2. Our analyses on CP and CHF patients indicate that LEWC-p can yield significant improvements compared to the current practice by striking a

better balance between patient safety and quality of care metrics. We also shed light on various hospital characteristics that will make the use of our proposed policy more beneficial.

We suspect that the proposed model and policy on patient flow from the ED to IWs can be extended to other areas of the hospital. Similar to the bed-block phenomenon in the ED, operating rooms (ORs) experience problems due to bed shortages in the post-anesthesia care unit (PACU). Our model and analyses can be used in those areas of a hospital to provide insights into the trade-off between waiting to be assigned to an appropriate bed versus a quick overflow to a less appropriate bed.

Our model can be also extended in various other ways. First, our model focuses on the patient flow between the ED to IWs without including the transfer between IWs (the requirements for such a flow would be different from our focus in this chapter). However, an extension of our model can be used to study patient flow between IWs, and hence, may provide other ways to further reduce ED boarding times. Second, our model considers IW beds as servers, although in the actual system the transfer process from ED to IWs is more complicated. For instance, in many hospitals, the nurses' availabilities often affect the patient flow, as nurses are responsible for transferring patients from the ED to IWs. Future research can expand our study by considering such more complex scenarios. Finally, in our objective function, we focus on the risk of adverse events and quality of care of patients admitted through the ED. Future research can extend our objective function by incorporating other concerns such as the LOS and waiting times for ED patients who are discharged home after their ED visit.

Chapter 5

CONCLUSION

In this chapter, we summarize the contributions of the study presented in this dissertation and list the possible directions for future research. The main contribution of this dissertation is introducing access management concept and analyzing patient access problem and providing guidelines to improve patient access to healthcare resources considering the unique features of different healthcare settings. Each chapter is motivated by real life problems associated with patient access and addresses those problems by employing analytical and simulation models utilizing real life data. The summary of all contributions with their associated chapters is presented below.

5.1 Summary of Contributions

In this dissertation, we focus on two different healthcare settings which are outpatient appointment scheduling and ED. In each chapter, we analyze specific issues regarding patient access. Our contributions in Chapters 2, 3, and 4 are summarized as follows.

We analyze an outpatient appointment setting and introduce access management in Chapter 2 and Chapter 3. Access management can be considered as a comprehensive approach that oversees the whole system and utilizes system specific characteristics to improve patient experience. This framework considers either medically determined or institution based priority classes along with patients' sensitivity to appointment delay and designs effective access protocols considering patient priority and sensitivity to appointment delays.

In Chapter 2, we focus on developing statistical models to estimate patients' sen-

sitivity to the delays, willingness to wait behavior, that are offered to patients at the time of appointment request from transactional appointment data. This study is essential in improving patient access important since the behavior has a significant impact on patients' experience and is not possible to be observed directly from the data. We introduce two statistical models, which are the survival model and rank-based choice model, which both estimate patient WtW through estimating the probability of realizing an appointment with a certain delay. This dissertation is the first study that utilizes a survival model to explain patient behavior. We conduct extensive numerical studies and show that both models are effective in identifying patient behavior from data. We use the proposed models on data from a real hospital clinic to gain managerial insights and suggest possible ways of utilizing those insights in clinical practice.

When we consider patient access, it does not simply refers to ability to access healthcare resources for all of the patients but acknowledging the differences in patient needs and service level expectations. Prioritization is an essential tool for responding these differences in patient needs and improving overall patient access by responding the needs of "right" patients by offering them "right" appointment delay. In Chapter 2, we show that patients' sensitivity to wait can be estimated by using available data. In Chapter 3, we introduce the idea of using appointment delay as a lever to control patient demand to address the mismatch between available clinical capacity and patient demand. We develop a time window based policy for patient access problem which is a unique approach in allocating available capacity to the patients from different priority groups. To the best of our knowledge, none of the studies in the literature focus on controlling the patient demand and allocating the available capacity considering patient behavior. Time window based policy is a practical approach that utilizes prioritization and easy to implement in real life settings.

The study presented in Chapter 4 is focusing on the second setting and considering an important problem in ED context. Bed block problem has been studied extensively in the literature, unlike those studies, we generate insights into effective ways of improving patient flow by analyzing the trade-off between patient safety and quality of care by considering number of boarded ED patients, risk of adverse events, and potential reduction of quality of care due to assignment to an alternative unit. Additionally, by using a stylized model we are able to gain some insights into the structure of the optimal policy and use those insights to develop easy to implement heuristic policies that are effective for bed assignment. Lastly, we develop a detailed simulation model calibrated with hospital data to assess the performance of our proposed policy and conduct sensitivity analyses. We show that system performance can be improved by utilizing overflow strategies that we suggested and we generate managerial insights to help bed managers in giving bed assignment decisions.

5.2 Future Work

One direction for future research is extending the study that is presented in Chapter 2 to include details on patients' reactions to specific offered delays and cancellation behavior in addition to WtW to fully characterize patient behavior along with specific slot based expectations. Additionally, one important patient characteristics that can be significant in patients' appointment fulfilling behavior is patients' location. Since the focus of our study is a destination clinic, for some patients, the earliest appointment slot might not be the best option for the patient due to required travel time. To able to conduct such analysis, detailed data collection and access to that data are required. This study can lead to an accurate realization probability estimation and let decision maker to decide on possible overbooks slots more effectively. This also can help us to develop time windows in a more detailed way since they are directly

associated with the estimated WtW distribution.

Our model in Chapter 3 is developed for a setting that focuses on serving patients from different priorities in a hierarchical manner. The model can be extended to cover alternative objectives such as minimizing the deviation from a targeted patient mix under fixed average TtA targets. Alternatively, one can focus on identifying the required minimum capacity to serve patients under certain fill rate expectations. The settings under which time window based policies are possibly effective can be studied under these alternative objectives as a possible future direction.

In Chapter 3, Section 3.5, we briefly introduce how to utilize trade-off curves to develop a decision making framework. While we are constructing those trade-off curves, we only consider two priority classes and mainly focus on fill rate as performance measure. Our approach can be extended to cover alternative performance measures simultaneously on the trade-off curves. The numerical results can be extended to gain more insights into effective strategies under different objectives when more than two priority classes present.

In all of our analyses, we assume that there exists a tool or set of rules to identify patient priorities. One way to prioritize patients is doing it based on medical needs which is not easy to identify at the time of the appointment request. A direction for future study can be developing prognostic tools by utilizing the clinical data to identify patients' conditions and prioritize them based on medical necessity and urgency of each condition. These tools can be set of questions that can be answered by the patients about their symptoms and medical history, and their previous test and imaging results, if they are relevant in determining patient needs.

Ahmadi-Javid, A., Z. Jalali and K. J. Klassen, "Outpatient appointment systems in healthcare: A review of optimization studies", European Journal of Operational Research **258**, 1, 3–34 (2017).

Andersen, P. K. and B. B. Ronn, "A nonparametric test for comparing two samples where all observations are either left-or right-censored", Biometrics pp. 323–329 (1995).

Anderson-Bergman, C., "An efficient implementation of the emicm algorithm for the interval censored npmle", Journal of Computational and Graphical Statistics **26**, 2, 463–467 (2017a).

Anderson-Bergman, C., "icenreg: Regression models for interval censored data in r", Journal of Statistical Software **81**, 12 (2017b).

Andradóttir, S., H. Ayhan and D. G. Down, "Compensating for failures with flexible servers", Operations Research **55**, 4, 753–768 (2007).

Armony, M. and N. Bambos, "Queueing dynamics and maximal throughput scheduling in switched processing systems", Queueing systems **44**, 3, 209–252 (2003).

Armony, M., S. Israelit, A. Mandelbaum, Y. N. Marmor, Y. Tseytlin and G. B. Yom-Tov, "On patient flow in hospitals: A data-based queueing-science perspective", Stochastic Systems **5**, 1, 146–194 (2015).

Armony, M. and A. R. Ward, "Fair dynamic routing in large-scale heterogeneous-server systems", Operations Research **58**, 3, 624–637 (2010).

Astaraky, D. and J. Patrick, "A simulation based approximate dynamic programming approach to multi-class, multi-resource surgical scheduling", European Journal of Operational Research **245**, 1, 309–319 (2015).

Ayvaz, N. and W. T. Huh, "Allocation of hospital capacity to multiple types of patients", Journal of Revenue and Pricing Management **9**, 5, 386–398 (2010).

Baker, D. W., C. D. Stevens and R. H. Brook, "Patients who leave a public hospital emergency department without being seen by a physician: causes and consequences", Jama **266**, 8, 1085–1090 (1991).

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

Bell, S. L. and R. J. Williams, "Dynamic scheduling of a system with two parallel servers in heavy traffic with resource pooling: asymptotic optimality of a threshold policy", Annals of Applied Probability pp. 608–649 (2001).

Belobaba, P. P., "Survey paperairline yield management an overview of seat inventory control", Transportation science **21**, 2, 63–73 (1987).

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

Berry Jaeker, J. A. and A. L. Tucker, "Past the point of speeding up: The negative effects of workload saturation on efficiency and patient severity", Management Science **63**, 4, 1042–1062 (2016).

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

Brown, L. D., T. T. Cai and A. DasGupta, "Interval estimation for a binomial proportion", Statistical science pp. 101–117 (2001).

Buyukkoc, C., P. Variaya and J. Walrand, "c mu rule revisited.", Adv. Appl. Prob. **17**, 1, 237–238 (1985).

Carr, B. G., J. E. Hollander, W. G. Baxt, E. M. Datner and J. M. Pines, "Trends in boarding of admitted patients in us emergency departments 2003–2005", Journal of Emergency Medicine **39**, 4, 506–511 (2010).

Cayirli, T. and E. Veral, "Outpatient scheduling in health care: a review of literature", Production and operations management **12**, 4, 519–549 (2003).

Cayirli, T., E. Veral and H. Rosen, "Assessment of patient classification in appointment system design", Production and Operations Management **17**, 3, 338–353 (2008).

Chan, C. W., V. F. Farias and G. J. Escobar, "The impact of delays on service times in the intensive care unit", Management Science **63**, 7, 2049–2072 (2016).

CNN, U., "Tape shows woman dying on waiting room floor", Updated July (2008).

Cox, D. R. and W. Smith, *Queues*, vol. 2 (CRC Press, 1991).

Dai, J. G. and W. Lin, "Maximum pressure policies in stochastic processing networks", Operations Research **53**, 2, 197–218 (2005).

De Véricourt, F. and Y.-P. Zhou, "Managing response time in a call-routing problem with service failure", Operations Research **53**, 6, 968–981 (2005).

Defife, J. A., C. Z. Conklin, J. M. Smith and J. Poole, "Psychotherapy appointment no-shows: Rates and reasons.", Psychotherapy: Theory, Research, Practice, Training **47**, 3, 413 (2010).

Dempster, A. P., N. M. Laird and D. B. Rubin, "Maximum likelihood from incomplete data via the em algorithm", Journal of the royal statistical society. Series B (methodological) pp. 1–38 (1977).

Dreiher, J., M. Froimovici, Y. Bibi, D. A. Vardy, A. Cicurel and A. D. Cohen, "Nonattendance in obstetrics and gynecology patients", Gynecologic and obstetric investigation **66**, 1, 40–43 (2008).

Erdelyi, A. and H. Topaloglu, "Computing protection level policies for dynamic capacity allocation problems by using stochastic approximation methods", Iie Transactions **41**, 6, 498–510 (2009).

Farias, V. F., S. Jagabathula and D. Shah, "A nonparametric approach to modeling choice with limited data", Management science **59**, 2, 305–322 (2013).

Fleming, T. R. and D. Lin, "Survival analysis in clinical trials: past developments and future directions", Biometrics **56**, 4, 971–983 (2000).

Gallucci, G., W. Swartz and F. Hackerman, "Impact of the wait for an initial appointment on the rate of kept appointments at a mental health center", Psychiatric Services **56**, 3, 344–346 (2005).

Gans, N., G. Koole and A. Mandelbaum, "Telephone call centers: Tutorial, review, and research prospects", Manufacturing & Service Operations Management **5**, 2, 79–141 (2003).

GAO, G. A. O., "Hospital emergency departments: crowding vary among hospitals and communities", URL http://www.gao.gov/new.items/d03460.pdf (2003).

GAO, G. A. O., "Hospital emergency departments: crowding continues to occur, and some patients wait longer than recommended time frames", URL http://www.gao.gov/new.items/d09347.pdf (2009).

Garnett, O. and A. Mandelbaum, "An introduction to skills-based routing and its operational complexities", Teaching notes **114** (2000).

Gentleman, R. and C. J. Geyer, "Maximum likelihood for interval censored data: Consistency and computation", Biometrika **81**, 3, 618–623 (1994).

Griffin, J. A., *Improving health care delivery through multi-objective resource allocation*, Ph.D. thesis, Georgia Institute of Technology (2012).

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

Gurvich, I. and O. Perry, "Overflow networks: Approximations and implications to call center outsourcing", Operations research **60**, 4, 996–1009 (2012).

Gurvich, I. and W. Whitt, "Scheduling flexible servers with convex delay costs in many-server service systems", Manufacturing & Service Operations Management **11**, 2, 237–253 (2009).

Harrison, G. W., A. Shafer and M. Mackay, "Modelling variability in hospital bed occupancy", Health Care Management Science **8**, 4, 325–334 (2005).

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

Kakalik, J. S. and J. D. Little, "Optimal service policy for the m/g/1 queue with multiple classes of arrivals.", Tech. rep., RAND CORP SANTA MONICA CALIF (1971).

Kaplan, E. L. and P. Meier, "Nonparametric estimation from incomplete observations", Journal of the American statistical association **53**, 282, 457–481 (1958).

Kazemian, P., M. Y. Sir, M. P. Van Oyen, J. K. Lovely, D. W. Larson and K. S. Pasupathy, "Coordinating clinic and surgery appointments to meet access service levels for elective surgery", Journal of biomedical informatics **66**, 105–115 (2017).

Kessler, R. C., A. Sonnega, E. Bromet, M. Hughes and C. B. Nelson, "Posttraumatic stress disorder in the national comorbidity survey", Archives of general psychiatry **52**, 12, 1048–1060 (1995).

Klassen, K. J. and T. R. Rohleder, "Scheduling outpatient appointments in a dynamic environment", Journal of operations Management **14**, 2, 83–101 (1996).

Klassen, K. J. and T. R. Rohleder, "Outpatient appointment scheduling with urgent clients in a dynamic, multi-period environment", International Journal of Service Industry Management **15**, 2, 167–186 (2004).

Kuntz, L., R. Mennicken and S. Scholtes, "Stress on the ward: Evidence of safety tipping points in hospitals", Management Science **61**, 4, 754–771 (2014).

LaGanga, L. R. and S. R. Lawrence, "Clinic overbooking to improve patient access and increase provider productivity", Decision Sciences **38**, 2, 251–276 (2007).

Levesque, J.-F., M. F. Harris and G. Russell, "Patient-centred access to health care: conceptualising access at the interface of health systems and populations", International journal for equity in health **12**, 1, 18 (2013).

Lin, W. and P. Kumar, "Optimal control of a queueing system with two heterogeneous servers", IEEE Transactions on Automatic control **29**, 8, 696–703 (1984).

Linn, I. S., "Stochastic dynamic programming and control of queueing systems", (1999).

Liu, N., S. R. Finkelstein, M. E. Kruk and D. Rosenthal, "When waiting to see a doctor is less irritating: Understanding patient preferences and choice behavior in appointment scheduling", Management Science **64**, 5, 1975–1996 (2017).

Liu, N. and S. Ziya, "Panel size and overbooking decisions for appointment-based services under patient no-shows", Production and Operations Management **23**, 12, 2209–2223 (2014).

Liu, N., S. Ziya and V. G. Kulkarni, "Dynamic scheduling of outpatient appointments under patient no-shows and cancellations", Manufacturing & Service Operations Management **12**, 2, 347–364 (2010).

Lucas, J., R. J. Batt and O. A. Soremekun, "Setting wait times to achieve targeted left-without-being-seen rates", The American journal of emergency medicine **32**, 4, 342–345 (2014).

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

Mandelbaum, A. and A. L. Stolyar, "Scheduling flexible servers with convex delay costs: Heavy-traffic optimality of the generalized cμ-rule", Operations Research **52**, 6, 836–855 (2004).

Merrit-Hawkins, "2017 survey of physician appointment wait times and medicaid and medicare acceptance rates", URL https://www.merritthawkins.com/uploadedFiles/MerrittHawkins/Pdf/mha2017waittimesurveyPDF.pdf (2017).

Meyn, S. P., "Sequencing and routing in multiclass queueing networks part i: Feedback regulation", SIAM Journal on Control and Optimization **40**, 3, 741–776 (2001).

Meyn, S. P., "Sequencing and routing in multiclass queueing networks part ii: Workload relaxations", SIAM Journal on Control and Optimization **42**, 1, 178–217 (2003).

Murray, M. and D. M. Berwick, "Advanced access: reducing waiting and delays in primary care", Jama **289**, 8, 1035–1040 (2003).

Norris, J. B., C. Kumar, S. Chand, H. Moskowitz, S. A. Shade and D. R. Willis, "An empirical investigation into factors affecting patient cancellations and no-shows at outpatient clinics", Decision Support Systems **57**, 428–443 (2014).

Osadchiy, N. and D. Kc, "Are patients patient? the role of time to appointment in patient flow", Production and Operations Management **26**, 3, 469–490 (2017).

Palmer, J. and I. Mitrani, "Optimal server allocation in reconfigurable clusters with multiple job types", in "International Conference on Computational Science and Its Applications", pp. 76–86 (Springer, 2004).

Parizi, M. S. and A. Ghate, "Multi-class, multi-resource advance scheduling with no-shows, cancellations and overbooking", Computers & Operations Research **67**, 90–101 (2016).

Patrick, J. and M. L. Puterman, "Improving resource utilization for diagnostic services through flexible inpatient scheduling: A method for improving resource utilization", Journal of the Operational Research Society **58**, 2, 235–245 (2007).

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

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

Proudlove, N., R. Boaden and J. Jorgensen, "Developing bed managers: the why and the how", Journal of nursing management **15**, 1, 34–42 (2007).

Rust, C. T., N. H. Gallups, W. S. Clark, D. S. Jones and W. D. Wilcox, "Patient appointment failures in pediatric resident continuity clinics", Archives of pediatrics & adolescent medicine **149**, 6, 693–695 (1995).

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

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

Saghafian, S., M. P. Van Oyen and B. Kolfal, "The w network and the dynamic control of unreliable flexible servers", IIE Transactions **43**, 12, 893–907 (2011).

Saghafian, S. and M. H. Veatch, "A c-mu rule for two-tiered parallel servers", IEEE Transactions on Automatic Control **61**, 4, 1046–1050 (2016).

Saultz, J. W., "Defining and measuring interpersonal continuity of care", The Annals of Family Medicine **1**, 3, 134–143 (2003).

Saure, A., J. Patrick, S. Tyldesley and M. L. Puterman, "Dynamic multi-appointment patient scheduling for radiation therapy", European Journal of Operational Research **223**, 2, 573–584 (2012).

Shi, P., M. C. Chou, J. Dai, D. Ding and J. Sim, "Models and insights for hospital inpatient operations: Time-dependent ed boarding time", Management Science **62**, 1, 1–28 (2015).

Smith, B. C., J. F. Leimkuhler and R. M. Darrow, "Yield management at american airlines", interfaces **22**, 1, 8–31 (1992).

Sun, J. and J. D. Kalbfleisch, "The analysis of current status data on point processes", Journal of the American Statistical Association **88**, 424, 1449–1454 (1993).

Teow, K. L., E. El-Darzi, C. Foo, X. Jin and J. Sim, "Intelligent analysis of acute bed overflow in a tertiary hospital in singapore", Journal of medical systems **36**, 3, 1873–1882 (2012).

Tezcan, T. and J. Dai, "Dynamic control of n-systems with many servers: Asymptotic optimality of a static priority policy in heavy traffic", Operations Research **58**, 1, 94–110 (2010).

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

Truong, V.-A., "Optimal advance scheduling", Management Science **61**, 7, 1584–1597 (2015).

Turnbull, B. W., "The empirical distribution function with arbitrarily grouped, censored and truncated data", Journal of the Royal Statistical Society. Series B (Methodological) pp. 290–295 (1976).

Van Mieghem, J. A., "Dynamic scheduling with convex delay costs: The generalized c— mu rule", The Annals of Applied Probability pp. 809–833 (1995).

van Ryzin, G. and G. Vulcano, "A market discovery algorithm to estimate a general class of nonparametric choice models", Management Science **61**, 2, 281–300 (2014).

van Ryzin, G. and G. Vulcano, "An expectation-maximization method to estimate a rank-based choice model of demand", Operations Research **65**, 2, 396–407 (2017).

Wallace, R. B. and W. Whitt, "A staffing algorithm for call centers with skill-based routing", Manufacturing & Service Operations Management **7**, 4, 276–294 (2005).

Walrand, J., *An introduction to queueing networks* (Prentice Hall, 1988).

Wang, W.-Y. and D. Gupta, "Adaptive appointment systems with patient preferences", Manufacturing & Service Operations Management **13**, 3, 373–389 (2011).

Wei, L.-J., "The accelerated failure time model: a useful alternative to the cox regression model in survival analysis", Statistics in medicine **11**, 14-15, 1871–1879 (1992).

Welch, P. D., "The statistical analysis of simulation results", The computer performance modeling handbook **22**, 268–328 (1983).

Wellner, J. A. and Y. Zhan, "A hybrid algorithm for computation of the nonparametric maximum likelihood estimator from censored data", Journal of the American Statistical Association **92**, 439, 945–959 (1997).

Zacharias, C. and M. Pinedo, "Appointment scheduling with no-shows and overbooking", Production and Operations Management **23**, 5, 788–801 (2014).

Zhan, D. and A. R. Ward, "Threshold routing to trade off waiting and call resolution in call centers", Manufacturing & Service Operations Management **16**, 2, 220–237 (2013).

Zhang, Z. and J. Sun, "Interval censoring", Statistical Methods in Medical Research **19**, 1, 53–70 (2010).

APPENDIX A

APPENDICES OF CHAPTER 2

A.1 Expectation Maximization (EM) Algorithm

The EM method starts with an initial estimate of the $\mathbf{p}$ vector. We use $1/(Z+1)$ for each WtW group. Then, with each instance of appointment booking, we compute the conditional expected value of the log-likelihood function by using the current estimates (expectation step). This expected value is then maximized to generate the new estimates (maximization step). The procedure is repeated until convergence is achieved according to a given stopping criterion; we use difference between two estimates of probabilities less than 10^{-4}. We now describe the two steps of expectation and maximization in further detail.

Expectation Step: In expectation step in iteration $l+1$, we calculate the expected values of the number of encounters in each WtW group, $\hat{m}_k^{l+1}$, with the estimates of probabilities of being in WtW group k denoted as $\hat{p}_k^{l+1}$. In order to calculate the estimate for the number of arrivals for each patient WtW group k at iteration $l+1$, denoted as $\hat{m}_k^{l+1}$, we first need to calculate an estimate for a_t , which we denote as $\hat{a}_t^{l+1}$, for the time buckets with no event (when $t \in \mathcal{S} \cup \mathcal{A}$, $a_t = 1$.)

We denote the probability of a patient arriving in time bucket t belonging to WtW group k considering the offered delay is w_t and the appointment offered is fulfilled as $P(G_t = k | t \in \mathcal{S}, W_t = w)$. Similarly, $P(G_t = k | t \in \mathcal{A}, W_t = w)$ is defined as the probability of a patient arriving in time bucket t belonging to WtW group k given that the booked appointment at t with delay $W_t = w$ is not fulfilled. At each time bucket t, we update the probability mass function (pmf) for each of these probabilities by using Bayes theorem based on the data which shows the set that the time bucket belongs to and the offered delay W_t at time bucket t, and current estimates of probabilities from the previous iteration (estimates obtained as a result of iteration l), $\hat{\mathbf{p}}^l$.

If the patient fulfills an appointment that is booked in time bucket t, we update

the pmf for any $k \in \{k \in \{w, \ldots, Z\}$:

$$P(G_t = k | t \in \mathcal{S}, W_t = w, \hat{\mathbf{p}}^l) = \frac{\mathbb{P}(k|\hat{\mathbf{p}}^l)}{P(t \in \mathcal{S}|W_t = w, \hat{\mathbf{p}}^l)} = \frac{\hat{p}_k^l}{\sum_{i=w}^{Z} \hat{p}_i^l}, \tag{A.1}$$

where $P(t \in \mathcal{S}|W_t = w, \hat{\mathbf{p}}^l)$ is the conditional probability that the appointment that is offered at time t being fulfilled with given delay $W_t = w$ and current probability estimates $\hat{\mathbf{p}}^l$ and indicator $\mathbb{1}_{k \in \{w,\ldots,Z\}}$ shows whether the appointment is offered to from a patient with WtW group k that has WtW greater than or equal to the offered delay. Similarly, if the booked appointment is C/RS/NS, we update the pmf for any $k \in \{0, \ldots, w-1\}$:

$$P(G_t = k | t \in \mathcal{A}, W_t = w, \hat{\mathbf{p}}^l) = \frac{\mathbb{P}(k|\hat{\mathbf{p}}^l)}{P(t \in \mathcal{A}|W_t = w, \hat{\mathbf{p}}^l)} = \frac{\hat{p}_k^l}{\sum_{i=0}^{w-1} \hat{p}_i^l}. \tag{A.2}$$

On the other hand, a time bucket $t \in \mathcal{B}$, we update the estimate $\hat{a}_t^{l+1}$ for the time bucket considering the appointment delay $W_t = w$ that can be offered were an an arrival to occur.

$$\begin{aligned} \hat{a}_t^{l+1} &= P(a_t = 1 | t \in \mathcal{B}, W_t = w, \hat{\mathbf{p}}^l) \\ &= \frac{P(t \in \mathcal{B}|W_t = w, a_t = 1, \hat{\mathbf{p}}^l) P(a_t = 1)}{P(t \in \mathcal{B}|W_t = w, \hat{\mathbf{p}}^l)} \\ &= \frac{P(t \in \mathcal{W}|W_t = w, \hat{\mathbf{p}}^l)\lambda}{\mathbb{P}(t \in \mathcal{W}|W_t = w, \hat{\mathbf{p}}^l)\lambda + (1-\lambda)} \\ &= \frac{\alpha \sum_{i=0}^{w-1} \hat{p}_i^l \lambda}{\alpha \sum_{i=0}^{w-1} \hat{p}_i^l \lambda + (1-\lambda)}. \end{aligned} \tag{A.3}$$

Notice that the probability that an appointment request occurs at a time bucket t is recorded as a no event bucket is the probability that an appointment request arrives at t with probability λ and patient decides to PLWBA. Then similar to (A.2), we calculate the probability $P\ (G_t = \ k|t \ \in \ \mathcal{B}, W_t = w, \hat{p})$, an arriving patient at a no event time bucket being from group $k \in \{0, \ldots, w-1\}$:

$$\mathbb{P}(G_t = k | t \in \mathcal{B}, W_t = w, \hat{\mathbf{p}}^l) \ = \frac{\hat{p}_k^l}{\sum_{i=0}^{w-1} \hat{p}_i^l}. \tag{A.4}$$

since an arriving patient decides to PLWBA since the offered delay at t, $W_t = w$ is exceeding patient's WtW. We then calculate the estimates of $\hat{m}_k$ from (A.3) as:

$$\hat{m}_k^{l+1} = \sum_{t\in\mathcal{S}} P(G_t = k|W_t = w, \hat{\mathbf{p}}^l) + \sum_{t\in\mathcal{A}} P(G_t = k|W_t = w, \hat{\mathbf{p}}^l) \tag{A.5}$$
$$+ \sum_{t\in\mathcal{B}} \hat{a}_t^{l+1} P(G_t = k|W_t = w, \hat{\mathbf{p}}^l).$$

Then the expected log-likelihood function becomes:

$$\mathsf{E}\left[\mathcal{L}(\mathbf{p})|\hat{\mathbf{p}}^l\right] = \sum_{i=0}^{Z} \hat{m}_i^{l+1} \log p_i^{l+1}. \tag{A.6}$$

Maximization Step: We calculate the maximizer $\hat{p}_k^{l+1}$ from:

$$\hat{p}_k^{l+1} = \frac{\hat{m}_k^{l+1}}{\sum_{i=0}^{Z} \hat{m}_i^{l+1}}. \tag{A.7}$$

We repeat this procedure until $\hat{\mathbf{p}}^{l+1} - \hat{\mathbf{p}}^l < 10^-4$.

A.1.1 Pseudocode for EM algorithm

Algorithm 1 EM algorithm for estimating realization probabilities

Input data: Maximum WtW T, number of time-buckets H, λ is estimated as dividing the total number of time buckets that an appointment request occurs to total number of time buckets, H, imputed booked appointment data matrix $I - BAD_t$ for each time-bucket t showing whether a booking occurred in the time-bucket $I - BAD_t(1) = 1$ or $I - BAD_t(1) = 0$ for no bookings, if the booked appointment is realized $I - BAD_t(2) = 1$, $I - BAD_t(2) = 0$, otherwise, and offered appointment delay $I - BAD_t(3) = w_t$, (1-α) is the misclassification error estimated from data.

Initialization: Set $p_k = 1/(Z+1)$, $a_t = 0 \quad \forall t \in \{1, 2, \ldots, H\}$

Repeat:

  $m_k := 0, p_{kt} := 0 \quad \forall t \in \{1, 2, \ldots, H\}, k \in \{0, 1, \ldots, Z\}$

  Expectation Step:

  for $t \in \{1, 2, \ldots, H\}$ **do**

    if $I - BAD_t(1) = 1$ **then**

      Set $a_t = 1$.

      if $I - BAD_t(2) = 1$ **then**

        for $k \in \{w_t, \ldots, Z\}$ **do**

          Set $p_{kt} = p_k / \sum_{i=w_t}^{Z} p_i$.

      else

        for $k \in \{0, \ldots, w_t - 1\}$ **do**

          Set $p_{kt} = p_k / \sum_{i=0}^{w_t - 1} p_i$.

15: **else**

16: Set $a_t = \dfrac{\lambda\alpha \sum_{i=0}^{w_t-1} p_i}{(1-\lambda) + \lambda\alpha \sum_{i=0}^{w_t-1} p_i}$

17: Update estimates for m_k

18: **for** $k \in \{0, 1, \ldots, T\}$ **do**

19: **for** $t \in \{1, 2, \ldots, H\}$ **do**

20: $m_k = m_k + a_t p_{kt}$

21: **Maximization Step**

22: **for** $k \in \{0, 1, \ldots, Z\}$ **do**

23: $p_k = m_k / (\sum_{t=1}^{H} a_t)$

24: **Until** Stopping defined criterion is met.

APPENDIX B

APPENDICES OF CHAPTER 3

B.1 Pseudocode for the Solution Approach

Algorithm 2 Algorithm for calculating time windows

Input data: Number of priority classes N, $T_{\min}$, $T_{\max}$, length of the booking horizon T and daily capacity C. Case specific inputs λ_i for all patient classes $i \in \{1, 2, \ldots, N\}$, overbook capacity $\theta_{\max}$, and set of realization probabilities p_{ik} for all patient classes $i \in \{1, 2, \ldots, N\}$ and for all delay values $k \in \{0, 1, \ldots, T\}$.

Construct Dilution Tables:

Create a matrix DT^i to store the results.

Calculation of possible $[B_i, E_i]$ pairs and set count of possible time windows $c_i = 0 \quad \forall \quad i \in \{1, 2, \ldots, N\}$:

for $i \in \{1, 2, \ldots, N\}$ **do**

  for $B_i \in \{1, 2, \ldots, T - T_{\min}\}$ **do**

    Increase window count by 1: $c_i = c_i + 1$

    Set E_i to any $B_i + T_{\min} - 1 \leq E_i \leq B_i + T_{\max} - 1$

    Set $TtA_{c_i} = \frac{B_i + E_i}{2}$

    Set $\mathrm{DT}^i[c_i, 1] = B_i, \mathrm{DT}^i[c_i, 2] = E_i, \mathrm{DT}^i[c_i, 3] = TtA_{c_i}$.

    Return DT^i.

  Initialize dominating solutions matrix DS^i having size $2(T - T\mathrm{min}) + 1 \times 4$ where $\mathrm{DS}^i[, 1]$ lists all possible $TtA \in \{\frac{T_{\min}}{2}, \frac{T_{\min}+1}{2}, \ldots, \frac{2T - T_{\min}}{2}\}$ and rest as 0.

  Calculate dilution DL_j for each window $j \in \{1, 2, \ldots, \text{length of } \mathrm{DT}^i\}$:

  for $j \in \{1, 2, \ldots, \text{length of } \mathrm{DT}^i\}$ **do**

Set $DL_j = \frac{1}{E_j - B_j + 1} \sum_{k=B_j}^{E_j} p_{ik}$.

$\mathrm{DT}^i[j,4] = DL_j$

Find l where $\mathrm{DT}^i[j,3] = \mathrm{DS}^i[l,1]$ and $l \in \{1, 2, \ldots, 2(T - T\min) + 1\}$

if $\mathrm{DT}^i[j,3] > \mathrm{DS}^i[l,4]$ **then**

Set $\mathrm{DS}^i[l,2] = \mathrm{DT}^i[j,1], \mathrm{DS}^i[l,3] = \mathrm{DT}^i[j,2]$ and $\mathrm{DS}^i[l,4] = \mathrm{DT}^i[j,3]$

Construct a table for possible Λ along with associated expected number of overbooks from expression (3.9) under capacity C and identify Λ^* for overbook capacity $\theta_{\max}$. Assign $\Lambda_{\text{remaining}} = \Lambda^*$.

for $i \in \{1, 2, \ldots, N\}$ **do**

if $\min(\mathrm{DS}^i[,4])\lambda_i \geq \Lambda_{\text{remaining}}$ **then**

Patient priority classes $q \in \{i, i+1, \ldots, N\}$ cannot be served with available capacity.

BREAK

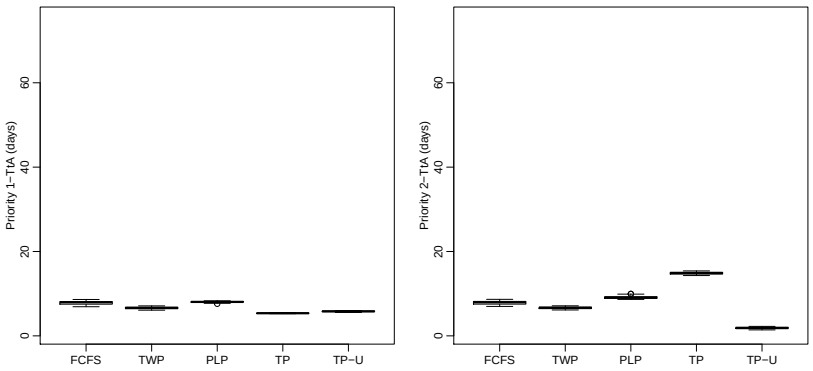
else

$\Lambda_i^* = \arg\min_{\mathrm{DS}^i[,4]}(\Lambda_{\text{remaining}} - (\mathrm{DS}^i[,4])\lambda_i)\lambda_i$

$\Lambda_{\text{remaining}} = \Lambda_{\text{remaining}} - \Lambda_i^*$

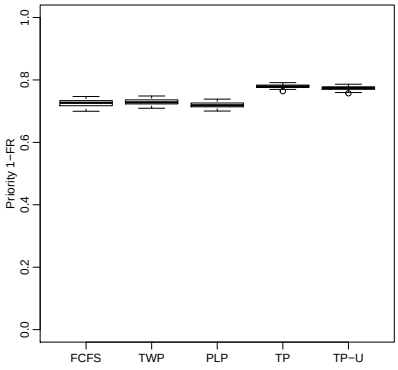
Find index, x, where $\mathrm{DS}^i[x,4] = \arg\min_{\mathrm{DS}^i[,4]}(\Lambda_{\text{remaining}} - (\mathrm{DS}^i[,4])\lambda_i)$

$B_i^* = \mathrm{DS}^i[x,2], E_i^* = \mathrm{DS}^i[x,3]$

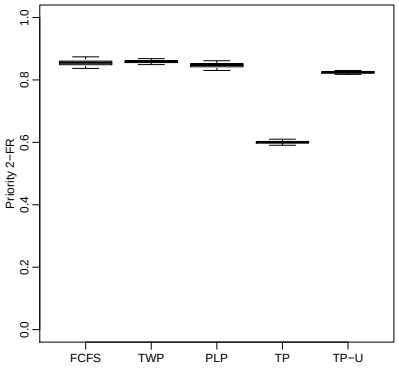


(a) Average TtA for Priority 1

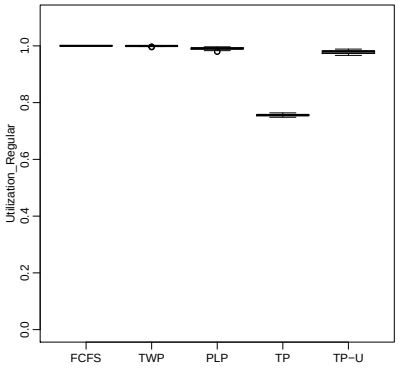
(b) Average TtA for Priority 2



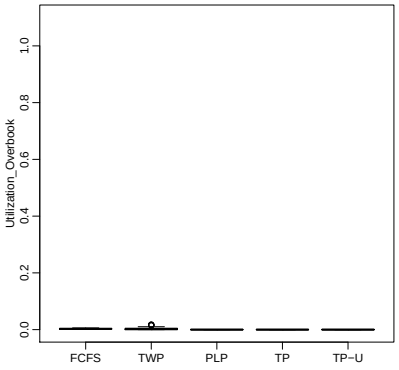
(c) Fill Rate for Priority 1



(d) Fill Rate for Priority 2

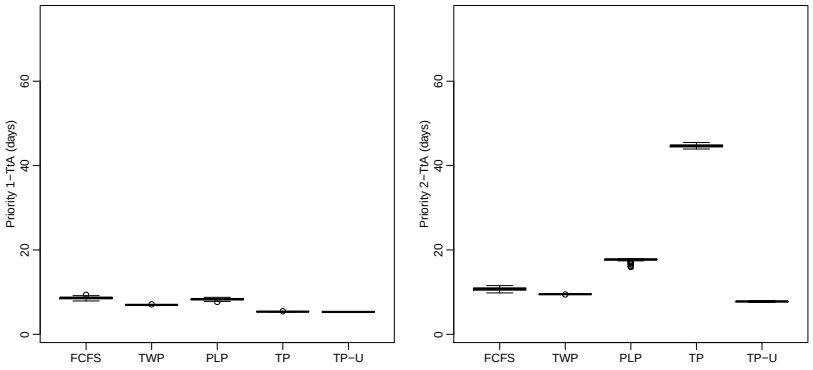


(e) Utilization of Regular Slots



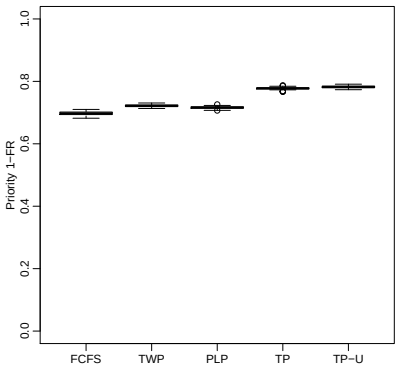
(f) Utilization of Overbook

Figure B.1: Simulation Results for WtW Case I, Arrival Case L1 ($\theta_{\max} = 5$)

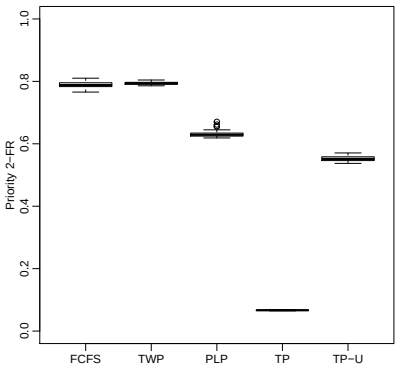


(a) Average TtA for Priority 1

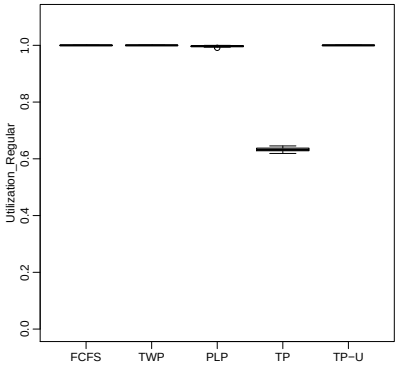
(b) Average TtA for Priority 2



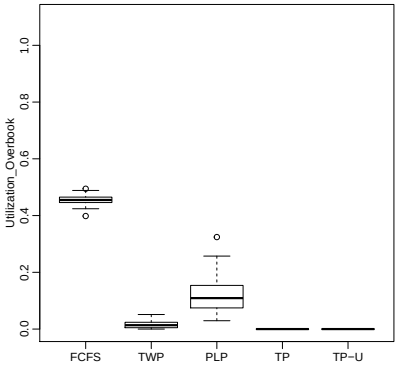
(c) Fill Rate for Priority 1



(d) Fill Rate for Priority 2

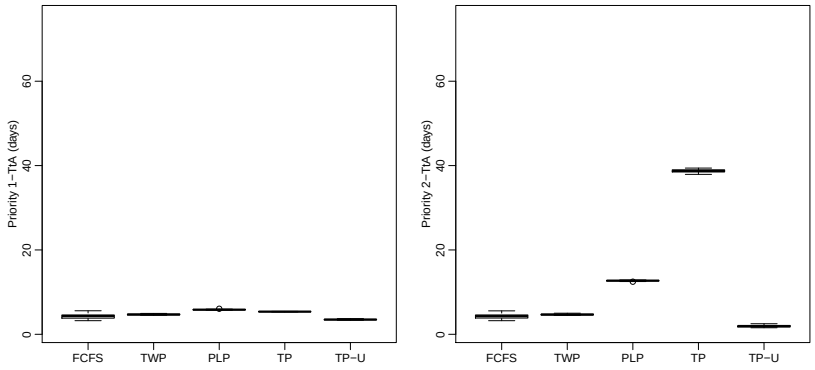


(e) Utilization of Regular Slots



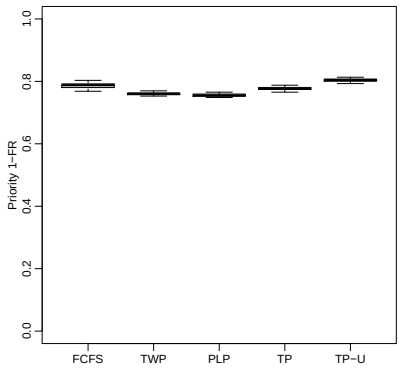
(f) Utilization of Overbook

Figure B.2: Simulation Results for WtW Case I, Arrival Case E2 ($\theta_{\max} = 5$)

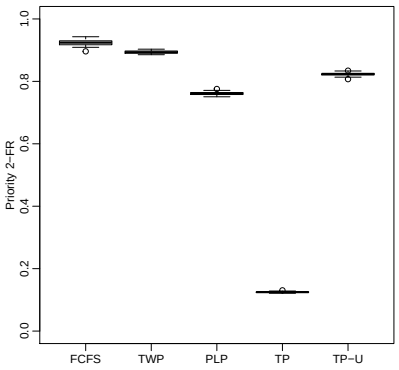


(a) Average TtA for Priority 1

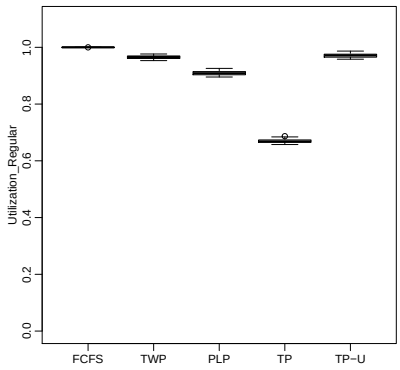
(b) Average TtA for Priority 2



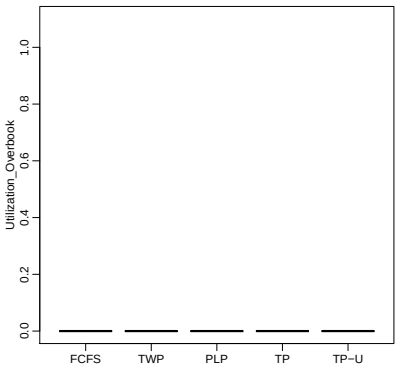
(c) Fill Rate for Priority 1



(d) Fill Rate for Priority 2

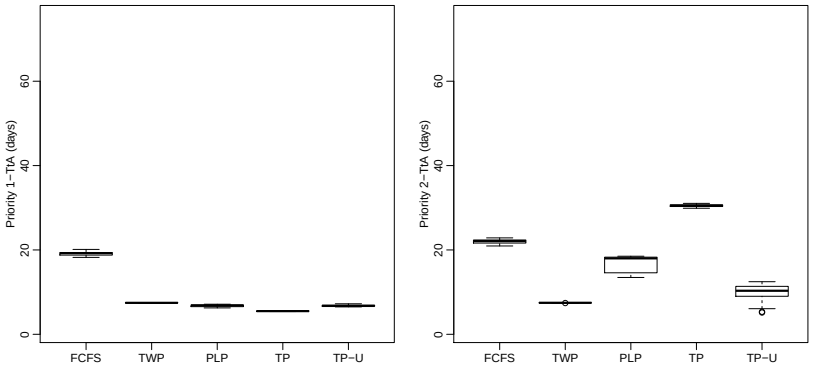


(e) Utilization of Regular Slots



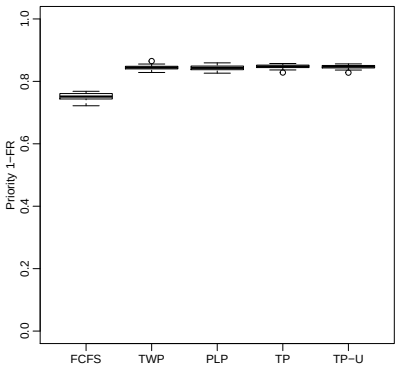
(f) Utilization of Overbook

Figure B.3: Simulation Results for WtW Case I, Arrival Case H2 ($\theta_{\max} = 5$)

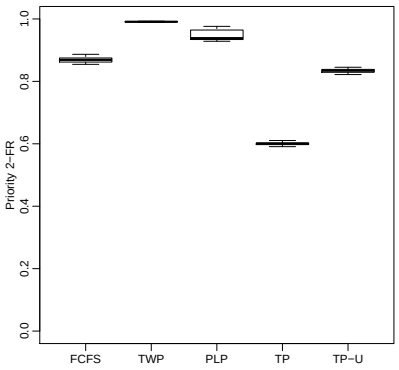


(a) Average TtA for Priority 1

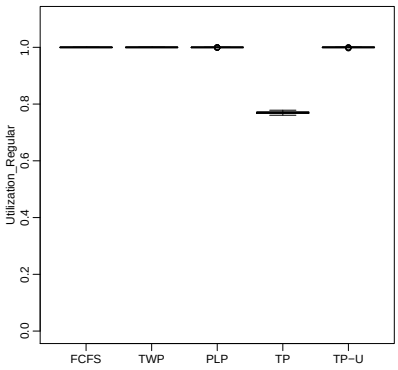
(b) Average TtA for Priority 2



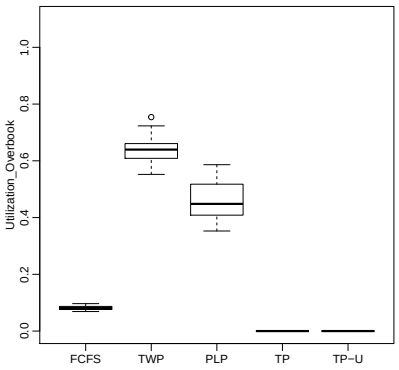
(c) Fill Rate for Priority 1



(d) Fill Rate for Priority 2

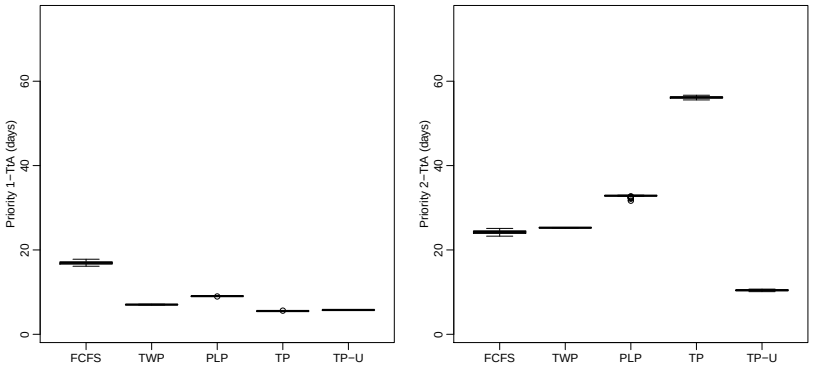


(e) Utilization of Regular Slots



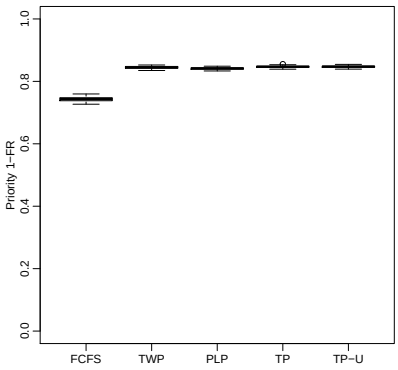
(f) Utilization of Overbook

Figure B.4: Simulation Results for WtW Case A, Arrival Case L1 ($\theta_{\max} = 5$)

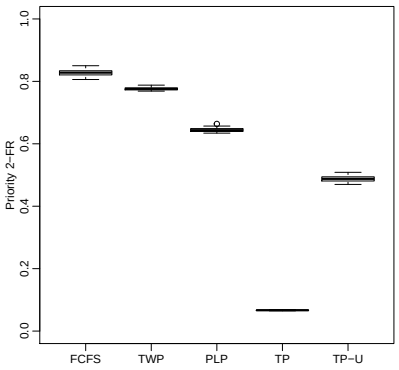


(a) Average TtA for Priority 1

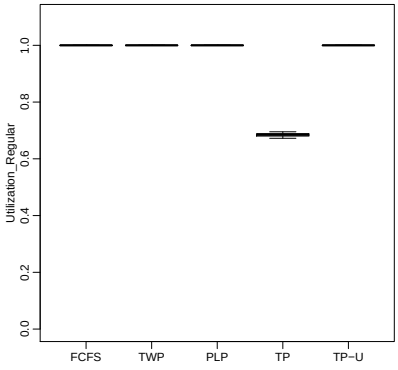
(b) Average TtA for Priority 2



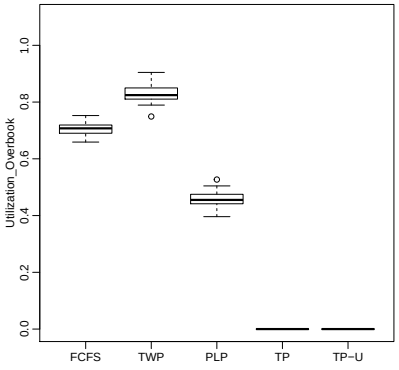
(c) Fill Rate for Priority 1



(d) Fill Rate for Priority 2

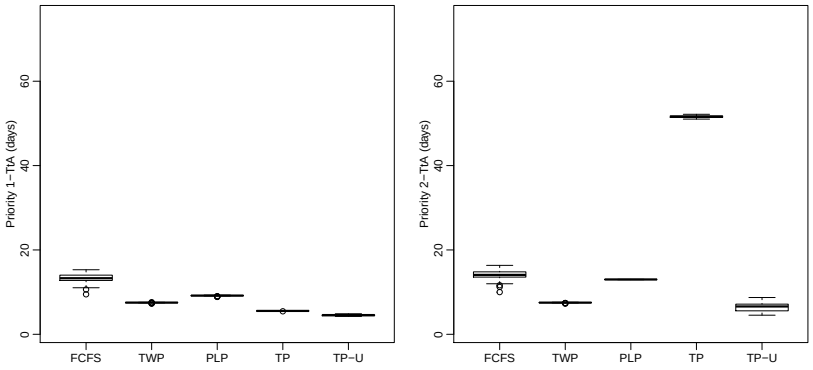


(e) Utilization of Regular Slots



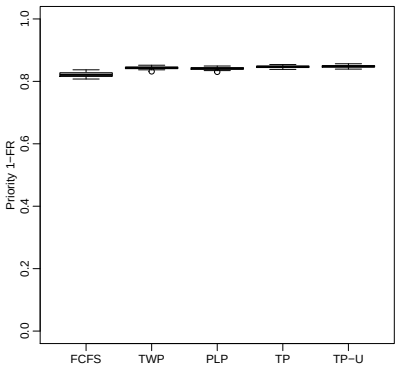
(f) Utilization of Overbook

Figure B.5: Simulation Results for WtW Case A, Arrival Case E2 ($\theta_{\max} = 5$)

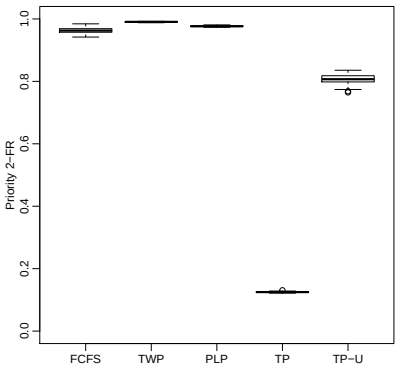


(a) Average TtA for Priority 1

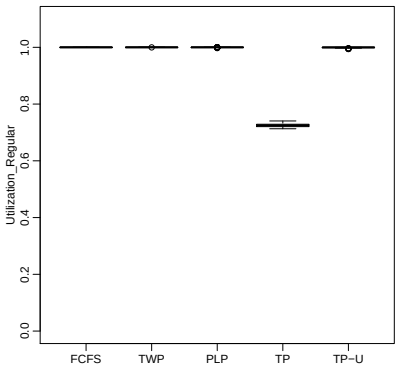
(b) Average TtA for Priority 2



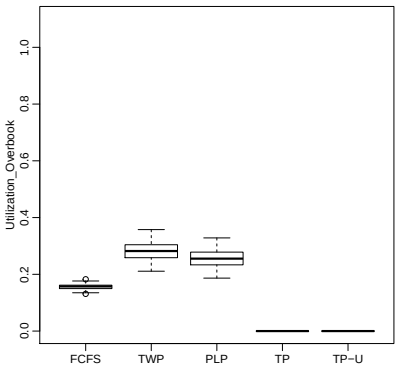
(c) Fill Rate for Priority 1



(d) Fill Rate for Priority 2



(e) Utilization of Regular Slots



(f) Utilization of Overbook

Figure B.6: Simulation Results for WtW Case A, Arrival Case H2 ($\theta_{\max} = 5$)

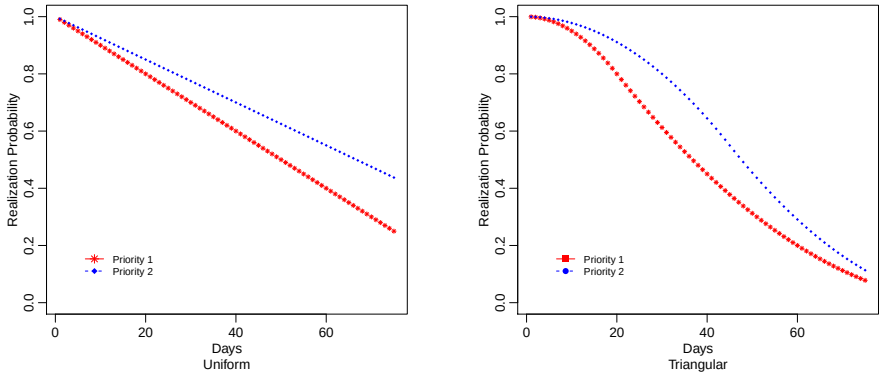
B.3 Additional Analysis on Compromise Prioritization

For our numerical experiments, we consider four alternative WtW distributions that are Uniform, Triangular, Exponential, and Weibull distributions. For each underlying WtW distribution, we assume patients from lower priority classes are less sensitive to access delays compared to patients from higher priority classes which is again creating the worse case for us since lower priority patients being less sensitive to appointment delays limits the improvement due to implementing TWP. Alternative WtW cases are illustrated in Figure B.7.

In addition to alternative WtW cases, we consider alternative arrival regimes where we only consider the cases with $\lambda_1 + \lambda_2 \geq C$ since for the cases that the total regular capacity can serve the total arriving demand, each priority class can be served with natural dilution. We consider three main regimes where in regime 1 (R1); $\lambda_1 \geq C$ and $\lambda_2 \leq C$ (λ_1=30, λ_2=5), in regime 2 (R2); $\lambda_1 \leq C$ and $\lambda_2 \geq C$ (λ_1=5, λ_2=30), and $\lambda_1 \leq C$ and $\lambda_2 \leq C$ (λ_1=15, λ_2=15). While these cases do not cover all possible arrivals, they serve our purpose of testing the performance of the model under different parameter settings.

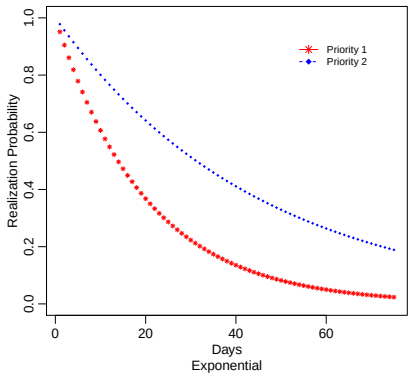
For our numerical experiments, we set $C = \ 20, \ , T = 75 \ , T_{\min} = \ 10$, and $T_{\max} = \ 20$ and a set of $\theta_{\max}$ values, $(0, 1, \ldots, 10)$, for each arrival regime and WtW distribution. Instead of $\theta_{\max} = 0$, we actually use 0.1 since $\theta_{\max} = 0$ leads no demand satisfied under the cases we consider.

The parameters that we pick are similar to the values that we observe from real life cases. Exploring the fill rates over a set of different $\theta_{\max}$ allows us to observe the trade-off between increasing the available capacity and serving more patients. Based on decision makers' preferences, an ideal level for the total capacity can be determined to serve patients with the targeted service levels.

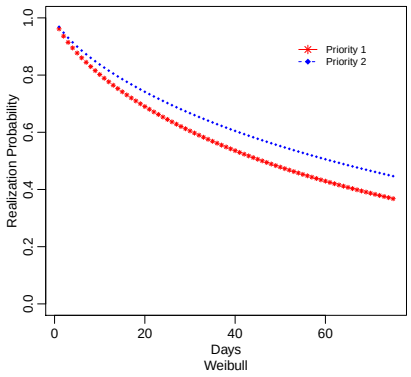


(a) WtW Case 1

(b) WtW Case 2



(c) WtW Case 3



(d) WtW Case 4

Figure B.7: WtW Cases for Numerical Analysis

We plot the trade-off curves as follows. We generate all possible time windows for priority 1 patients and with the service level provided to priority 1 patients with each time window, we fill the remaining effective capacity with priority 2 patients as we note in Section 3.5. In trade-off curves, we basically plot the inequality,

$$\sum_{n=1}^{2} \beta_n \lambda_n \leq \Lambda^*, \tag{B.1}$$

in terms of β_1 and β_2 where Λ^* is a function of the available capacity.

Notice that we can represent the relationship between the average TtA and fill

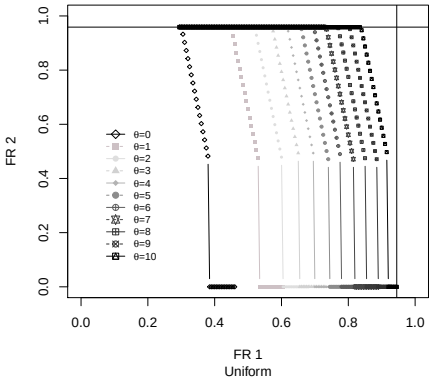
rate is many-to-many since the time windows with same average TtA can result in different fill rates. For instance, in WtW Case 1 where patient WtW is uniformly distributed, $P(\text{WtW} \leq x) = \frac{x-a}{b-a}$, we write the expression for β_n as

$$\begin{aligned} \frac{1}{E_n - B_n + 1} & \sum_{k=B_n}^{E_n} \left(1 - \frac{k-1-a}{b-a}\right) \\ &= \frac{1}{E_n - B_n + 1} \frac{(E_n - B_n + 1)(E_n + B_n - 2b - 2)}{2(b-a)} \end{aligned} \tag{B.2}$$

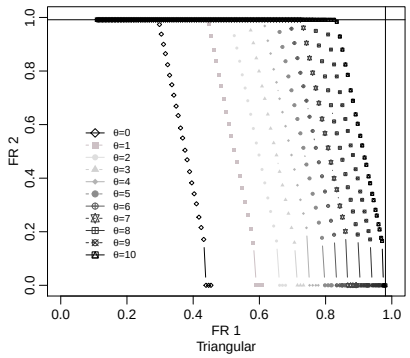
which indicates that for the $[B_i, E_i]$ pairs that results in the same average TtA (same $E_i + B_i$), there exists alternative time windows that results in same performance measures since there is one-to-one relationship between average TtA and fill rate under Uniform WtW. Therefore, under WtW case 1, for the average TtA target, time window pair $[B, E]$ is determined based on setting dependent preferences. Notice that for the time windows that result in same average TtA, some windows are wider with a lesser minimum TtA can be achieved while the narrower ones result in lesser maximum TtA possible. Some decision makers might prefer narrower intervals to reduce the variance of TtA among the patients while some choose to continue with time windows that start earlier in the booking horizon. When we are plotting the trade-off curves, we only pick the time windows with the highest fill rate among the ones that with the same average TtA value and refer them as dominating time windows.

We illustrate the trade-off curves under each WtW case in Figure B.8 and Figure B.9 under arrival regimes R1 and R2 separately. Maximum possible fill rates under different WtW case are shown in the figures with a vertical and a horizontal line for priority 1 and priority 2, respectively.

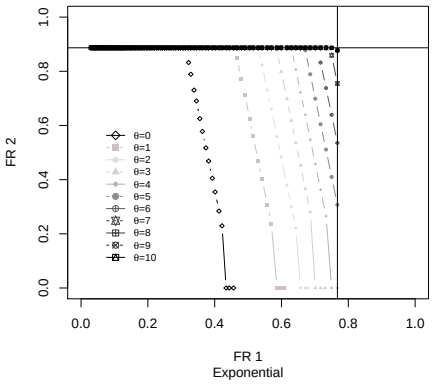
Trade-off curves that are presented in Figure B.8 and Figure B.9 give a complete picture of the performance measures and trade-off between these performance measures under different arrival regimes and WtW cases. Under any WtW case, the



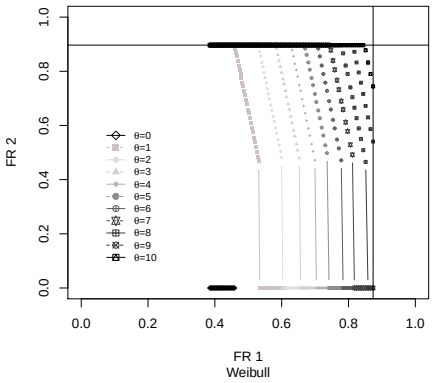
(a) WtW Case 1



(b) WtW Case 2



(c) WtW Case 3

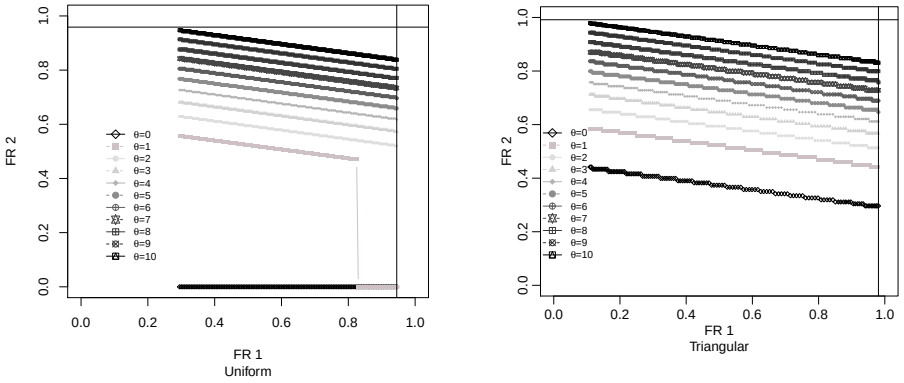


(d) WtW Case 4

Figure B.8: Trade-off Curves Under R1 and Different WtW Cases

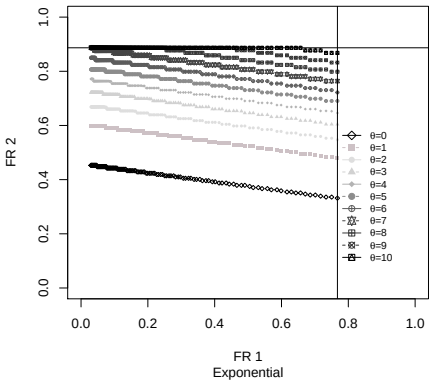
trade-off curves can be used to identify the possible service levels that can be reached with the available capacity or the required capacity for providing the targeted service levels for the patients.

The first observation that we can make from trade-off curves is that as we increase the available overbook slots, we observe relative improvement in performance measures diminish which is supporting our observation in Section 3.5. Considering that overbooks can result in additional direct waiting times for patients and reduction

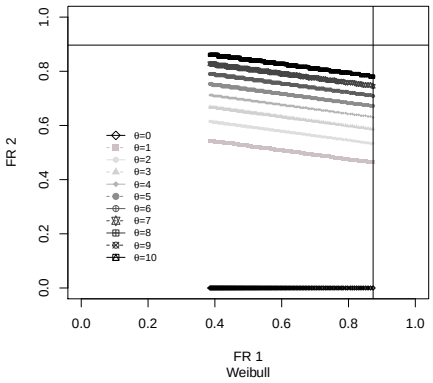


(a) WtW Case 1

(b) WtW Case 2



(c) WtW Case 3



(d) WtW Case 4

Figure B.9: Trade-off Curves Under R2 and Different WtW Cases

in provider satisfaction, it is expected that we observe a reduction in total system performance after a certain level of available overbooks. If cost parameters associated with overbooks and serving an additional patient is available, one can decide the optimal capacity to maximize the return. This capacity can be either in terms of overbook capacity or additional regular capacity that can be provided by hiring additional medical professionals.

In some settings, decision makers can focus on providing a service level in terms

of fill rates for all patient classes. This case can be considered as a setting where priorities exist but there is no distinct differences between the priority classes. For instance, for the case depicted in Figure B.8(a), instead of serving priority 1 patients with 90% fill rate and not serving priority 2 patients, one can choose to reduce the fill rate for priority 1 to 85% and serve priority 2 with 47.5% when $\theta_{\max}$ is set to 8. While this shift does not result any change in total number of patients served since the capacity is fixed, it allows us to provide care to an additional patient class with a minor reduction in fill rate for priority 1 patients.

In the cases in Figure B.8(a), it is not possible to serve both classes of patients while serving priority 1 with natural dilution without increasing the available capacity significantly and it is not even possible to serve priority 1 patients with natural dilution since the number of patients requesting appointments from priority 1 patients are significantly higher than the available capacity. In cases like these, the first task of the decision makers should be determining the ideal capacity that can respond to patients' needs. However, if increasing the capacity is costly or not possible due to physical limitations, an appropriate capacity allocation can be either rejecting all priority patients or use compromise prioritization, serving priority 1 patients given that the priority 2 patients should be served with maximum dilution.

Alternative set of time windows can be set by adjusting our three levers. We can analyze a specific setting to observe how the levers are associated. We examine the case that strict prioritization is used with the parameters where appointment requests arrive with R2, $\theta_{\max}$ is set to 4 slots, and patient behavior can be represented with WtW Case 3, and the goal of the decision maker is to improve the fill rate for priority 2 patients by 0.05. The results show that we can serve priority 1 patient with the earliest time window $[1, 10]$ with fill rate of 76.7% while serving priority 2 with $[11, 29]$ with fill rate 64.6% under $\theta_{\max} = 4$ case. To achieve the goal, we can serve priority 1

with time window $[6, 25]$ with fill rate 48% and serve priority 2 with the time window $[8, 25]$ with 69.8% fill rate or we can increase $\theta_{\max}$ to 5 and serve priority 2 with the time window $[8, 25]$ with 69.8% fill rate while serving priority 1 with 71.7% fill rate with time window $[1, 13]$. In this example, for the same $\theta_{\max}$ value, one needs to decrease the fill rate for priority 1 patients by 28% or increase the overbook capacity to be able to increase service level priority 2 patients by 0.05 with reducing fill rate for priority 1 patients by 0.05. Depending on cost parameters or any parameter that can capture the trade-off between serving priority 1 patients and increasing the expected overbooks, one can decide on the best action to increase the fill rate for priority 2 patients.

Depending on the underlying WtW distribution, the trade-off can be more or less. For instance, for the same objective and parameters under WtW Case 2, we observe that with strict prioritization, we serve priority 1 patients with 98.1% fill rate while serving priority 2 at 61.2% fill rate with time window $[32, 51]$. Again, to increase fill rate for priority 2 by by 0.05, we can decrease the fill rate for priority 1 patients to 68.6% with time window $[21, 31]$ or increase the overbook capacity to 5, and serve priority 1 patients with 72.4% fill rate with the time window $[15, 33]$.

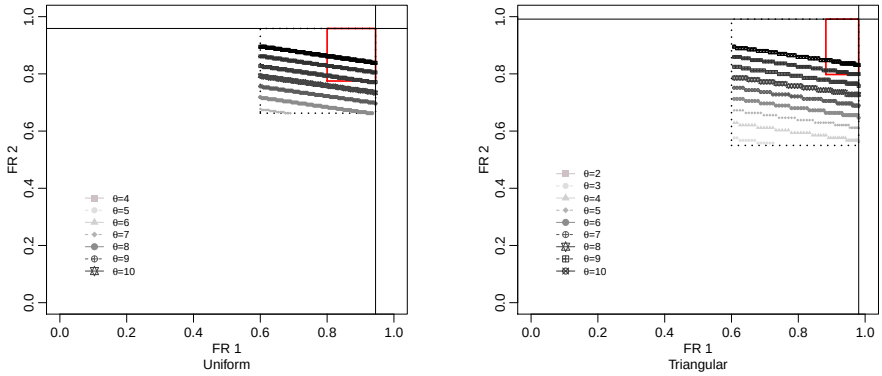
Compared to results from WtW Case 3, we observe that under WtW Case 2, we need to compromise the average TtA of priority 1 patient more to reach the desired target fill rate for priority 2 patient even with increasing the overbook capacity. The difference between the required compromise from the average TtA is due to the apparent differences in WtW distributions. From Figures B.7(b) and B.7(c), we can see that patients are more sensitive to delay values under WtW Case 3, therefore, desired reduction in fill rate is observed for lower delay values compared to WtW Case 2. Notice that for both cases, we need to reduce the fill rate for priority 1 patients approximately by 0.3 to increase fill rate for priority 2 patients by by 0.05. This is

due to the slope of the trade-off curve being the same for the same arrival regime which is $-\frac{1}{6}$ under R2.

Similar to what we did for the real life data, we show alternative results under average TtA targets where τ_1 is determined as 15 days and τ_2 is 30 days for priority class 1 and priority class 2, respectively. We mark the associated area on the trade-off curves that contains the alternative possible fill rate combinations for each WtW case with solid box in Figure B.10. One observation that we can make from the figure is that for each WtW case we cover a different level of fill rate combinations. Additionally, due to the differences in WtW distributions, the dominating time windows to hit the target TtA varies. For instance, for Uniform WtW, any $B_1 + E_1 = 30$, and any $B_2 + E_2 = 60$ combinations are possible time windows that can be used to hit average TtA targets where for Exponential and Weibull WtW it is $[6, 24]$ for priority 1 and $[21, 39]$ for priority 2 patients and for Triangular WtW, $[10, 20]$ and $[25, 35]$ for priority 1 and priority 2, respectively. These time windows listed are the ones to satisfy the targets at minimum.

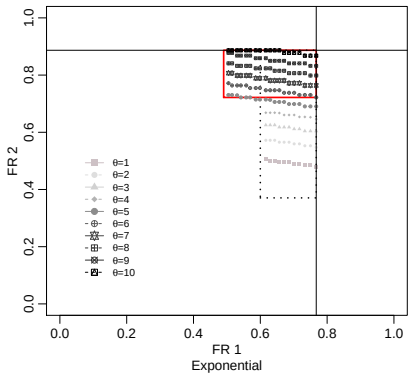
Another example that we examine is a fill rate target on priority 1 patients while keeping average TtA for priority 2 patients under a certain level. The dotted box marked on Figures B.10(a)-B.10(d) show the alternative fill rates possible that satisfy target on 60% service level for priority 1 patients and 45 days average TtA target for priority 2. The figures show that the required total capacity varies for each WtW case. For instance, in a WtW case where patients are highly sensitive to appointment delays such as Exponential WtW case, higher number of overbooks is not required compared to less sensitive WtW cases such as Uniform since for the same TtA value more patients are abandoning the system in Exponential WtW case.

It is possible include more details in the trade-off curves to observe both fill rates and certain ranges of average TtA values by marking the trade-off curves with different

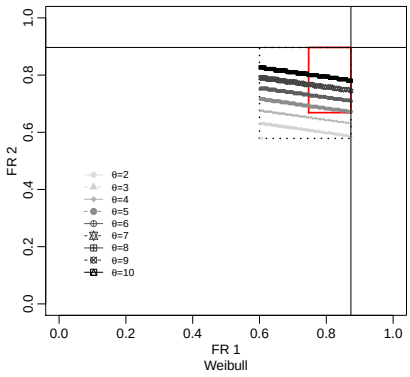


(a) WtW Case 1

(b) WtW Case 2



(c) WtW Case 3



(d) WtW Case 4

Figure B.10: Target Areas on Trade-off Curves

colors to express alternative combinations of average TtAs. Including average TtA on trade-off curves can help us to give decision makers a single tool that they let them observe three main components, average TtA, fill rate, expected number of overbooks, at the same time. Notice that fill rate and average TtA are similar performance measures since we assume that patient abandonments increase with the offered delay. In some cases, target fill rates can be more limiting than the average TtA target. For instance, in a highly competitive environment or in a setting where patients are

impatient like the WtW case 3 in FigureB.8(c), patients can be served with relatively shorter TtA metric with a moderate level of overbooks. However, the achievable fill rate is low since those patients tend to abandon the system even when the offered delay is low or they are mostly searching for same-day appointments.

APPENDIX C

APPENDICES OF CHAPTER 4

C.1 Definition of the Functional Operators

In this part, we define the operators $T^{\underline{u_{ij}}}$, which defines the cost function $J(\cdot)$ after a decision $\underline{u_{ij}}$ is taken at state $\tilde{X}$. Below, we explicitly define this operator for each state $\tilde{X}$.

When the system state is $\tilde{X} = (X_1 \geq 1, X_2 \geq 1, \underline{a_1} \neq 0, \underline{a_2} = 0)$:

$$\begin{aligned}
T^{(u_{i1}=,u_{i2}=0)}J(\tilde{X}) =& J(X_1, X_2, \underline{a_1} \neq 0, \underline{a_2} = 0) \\
T^{(u_{i1}=,u_{12}=1)}J(\tilde{X}) =& J(X_1 - 1, X_2, \underline{a_1} \neq 0, \underline{a_2} = e_1) + p_{12} \\
T^{(u_{i1}=,u_{22}=1)}J(\tilde{X}) =& J(X_1, X_2 - 1, \underline{a_1} \neq 0, \underline{a_2} = e_2)
\end{aligned}$$

When the state is $\tilde{X} = (X_1 \geq 2, X_2 \geq 2, \underline{a_1} = 0, \underline{a_2} \neq 0)$:

$$\begin{aligned}
T^{(u_{i1}=0,u_{i2}=0)}J(\tilde{X}) =& J(X_1, X_2, \underline{a_1} = 0, \underline{a_2} \neq 0) \\
T^{(u_{11}=1,u_{i2}=0)}J(\tilde{X}) =& J(X_1 - 1, X_2, \underline{a_1} = e_1, \underline{a_2} \neq 0) \\
T^{(u_{21}=1,u_{i2}=0)}J(\tilde{X}) =& J(X_1, X_2 - 1, \underline{a_1} = e_2, \underline{a_2} \neq 0) + p_{21}
\end{aligned}$$

When the system state is $\tilde{X} = (X_1 \geq 2, X_2 \geq 2, \underline{a_1} = 0, \underline{a_2} = 0)$:

$$\begin{aligned}
T^{\underline{u_{ij}}=0}J(\tilde{X}) =& J(X_1, X_2, \underline{a_1} = 0, \underline{a_2} = 0) \\
T^{(u_{11}=1,u_{22}=1)}J(\tilde{X}) =& J(X_1 - 1, X_2 - 1, \underline{a_1} = e_1, \underline{a_2} = e_2) \\
T^{(u_{11}=1,u_{12}=1)}J(\tilde{X}) =& J(X_1 - 2, X_2, \underline{a_1} = e_1, \underline{a_2} = e_1) + p_{12} \\
T^{(u_{11}=1,u_{i2}=0)}J(\tilde{X}) =& J(X_1 - 1, X_2, \underline{a_1} = e_1, \underline{a_2} = 0) \\
T^{(u_{i1}=0,u_{12}=1)}J(\tilde{X}) =& J(X_1 - 1, X_2, \underline{a_1} = 0, \underline{a_2} = e_1) + p_{12} \\
T^{(u_{i1}=0,u_{22}=1)}J(\tilde{X}) =& J(X_1, X_2 - 1, \underline{a_1} = 0, \underline{a_2} = e_2) \\
T^{(u_{21}=1,u_{i2}=0)}J(\tilde{X}) =& J(X_1, X_2 - 1, \underline{a_1} = e_2, \underline{a_2} = 0) + p_{21} \\
T^{(u_{21}=1,u_{22}=1)}J(\tilde{X}) =& J(X_1, X_2 - 2, \underline{a_1} = e_2, \underline{a_2} = e_2) + p_{21} \\
T^{(u_{12}=1,u_{21}=1)}J(\tilde{X}) =& J(X_1 - 1, X_2 - 1, \underline{a_1} = e_2, \underline{a_2} = e_1) + p_{12} + p_{21}
\end{aligned}$$

When the state is $\tilde{X} = (X_1 = 0, X_2 \geq 2, \underline{a_1} = 0, \underline{a_2} = 0)$:

$$
\begin{aligned}
T^{\underline{u_{ij}}=0} J(\tilde{X}) =& J(X_1, X_2, \underline{a_1} = 0, \underline{a_2} = 0) \\
T^{(u_{i1}=0, u_{22}=0)} J(\tilde{X}) =& J(X_1, X_2 - 1, \underline{a_1} = 0, \underline{a_2} = e_2) \\
T^{(u_{21}=1, u_{i2}=0)} J(\tilde{X}) =& J(X_1, X_2 - 1, \underline{a_1} = e_2, \underline{a_2} = 0) + p_{21} \\
T^{(u_{21}=1, u_{22}=1)} J(\tilde{X}) =& J(X_1, X_2 - 2, \underline{a_1} = e_2, \underline{a_2} = e_2) + p_{21}
\end{aligned}
$$

When the state is $\tilde{X} = (X_1 \geq 2, X_2 = 0, \underline{a_1} = 0, \underline{a_2} = 0)$:

$$
\begin{aligned}
T^{\underline{u_{ij}}=0} J(\tilde{X}) =& J(X_1, X_2, \underline{a_1} = 0, \underline{a_2} = 0) \\
T^{(u_{11}=1, u_{12}=1)} J(\tilde{X}) =& J(X_1 - 2, X_2, \underline{a_1} = e_1, \underline{a_2} = e_1) + p_{12} \\
T^{(u_{11}=1, u_{i2}=0)} J(\tilde{X}) =& J(X_1 - 1, X_2, \underline{a_1} = e_1, \underline{a_2} = 0) \\
T^{(u_{i1}=0, u_{12}=1)} J(\tilde{X}) =& J(X_1 - 1, X_2, \underline{a_1} = 0, \underline{a_2} = e_1) + p_{12}
\end{aligned}
$$

When the state is $\tilde{X} = (X_1 \geq 2, X_2 = 1, \underline{a_1} = 0, \underline{a_2} = 0)$:

$$
\begin{aligned}
T^{\underline{u_{ij}}=0} J(\tilde{X}) =& J(X_1, X_2, \underline{a_1} = 0, \underline{a_2} = 0) \\
T^{(u_{11}=1, u_{22}=1)} J(\tilde{X}) =& J(X_1 - 1, X_2 - 1, \underline{a_1} = e_1, \underline{a_2} = e_2) \\
T^{(u_{11}=1, u_{12}=1)} J(\tilde{X}) =& J(X_1 - 2, X_2, \underline{a_1} = e_1, \underline{a_2} = e_1) + p_{12} \\
T^{(u_{11}=1, u_{i2}=0)} J(\tilde{X}) =& J(X_1 - 1, X_2, \underline{a_1} = e_1, \underline{a_2} = 0) \\
T^{(u_{i1}=0, u_{12}=1)} J(\tilde{X}) =& J(X_1 - 1, X_2, \underline{a_1} = 0, \underline{a_2} = e_1) + p_{12} \\
T^{(u_{i1}=0, u_{22}=1)} J(\tilde{X}) =& J(X_1, X_2 - 1, \underline{a_1} = 0, \underline{a_2} = e_2) \\
T^{(u_{21}=1, u_{i2}=0)} J(\tilde{X}) =& J(X_1, X_2 - 1, \underline{a_1} = e_2, \underline{a_2} = 0) + p_{21} \\
T^{(u_{12}=1, u_{21}=1)} J(\tilde{X}) =& J(X_1 - 1, X_2 - 1, \underline{a_1} = e_2, \underline{a_2} = e_1) + p_{12} + p_{21}
\end{aligned}
$$

When the state is $\tilde{X} = (X_1 = 1, X_2 \geq 2, \underline{a_1} = 0, \underline{a_2} = 0)$:

$$
\begin{aligned}
T^{\underline{u_{ij}}=0} J(\tilde{X}) =& J(X_1, X_2, \underline{a_1} = 0, \underline{a_2} = 0) \\
T^{(u_{11}=1, u_{22}=1)} J(\tilde{X}) =& J(X_1 - 1, X_2 - 1, \underline{a_1} = e_1, \underline{a_2} = e_2) \\
T^{(u_{11}=1, u_{i2}=0)} J(\tilde{X}) =& J(X_1 - 1, X_2, \underline{a_1} = e_1, \underline{a_2} = 0) \\
T^{(u_{i1}=0, u_{12}=1)} J(\tilde{X}) =& J(X_1 - 1, X_2, \underline{a_1} = 0, \underline{a_2} = e_1) + p_{12} \\
T^{(u_{i1}=0, u_{22}=1)} J(\tilde{X}) =& J(X_1, X_2 - 1, \underline{a_1} = 0, \underline{a_2} = e_2) \\
T^{(u_{21}=1, u_{i2}=0)} J(\tilde{X}) =& J(X_1, X_2 - 1, \underline{a_1} = e_2, \underline{a_2} = 0) + p_{21} \\
T^{(u_{21}=1, u_{22}=1)} J(\tilde{X}) =& J(X_1, X_2 - 2, \underline{a_1} = e_2, \underline{a_2} = e_2) + p_{21} \\
T^{(u_{12}=1, u_{21}=1)} J(\tilde{X}) =& J(X_1 - 1, X_2 - 1, \underline{a_1} = e_2, \underline{a_2} = e_1) + p_{12} + p_{21}
\end{aligned}
$$

When the state is $\tilde{X} = (X_1 \geq 2, X_2 = 2, \underline{a_1} \neq 0, \underline{a_2} = 0)$:

$$
\begin{aligned}
T^{\underline{u_{ij}}=0} J(\tilde{X}) =& J(X_1, X_2, \underline{a_1} \neq 0, \underline{a_2} = 0) \\
T^{(u_{i1}, u_{12}=1)} J(\tilde{X}) =& J(X_1 - 1, X_2, \underline{a_1} \neq 0, \underline{a_2} = e_1) + p_{12}
\end{aligned}
$$

When the state is $\tilde{X} = (X_1 = 0, X_2 \geq 2, \underline{a_1} \neq 0, \underline{a_2} = 0)$:

$$
\begin{aligned}
T^{\underline{u_{ij}}=0} J(\tilde{X}) =& J(X_1, X_2, \underline{a_1} \neq 0, \underline{a_2} = 0) \\
T^{(u_{i1}, u_{22}=1)} J(\tilde{X}) =& J(X_1, X_2 - 1, \underline{a_1} \neq 0, \underline{a_2} = e_2)
\end{aligned}
$$

When the state is $\tilde{X} = (X_1 \geq 2, X_2 = 0, \underline{a_1} = 0, \underline{a_2} \neq 0)$:

$$
\begin{aligned}
T^{\underline{u_{ij}}=0} J(\tilde{X}) =& J(X_1, X_2, \underline{a_1} = 0, \underline{a_2} \neq 0) \\
T^{(u_{11}=1, u_{i2}=0)} J(\tilde{X}) =& J(X_1 - 1, X_2, \underline{a_1} = e_1, \underline{a_2} \neq 0)
\end{aligned}
$$

When the state is $\tilde{X} = (X_1 = 0, X_2 \geq 2, \underline{a_1} = 0, \underline{a_2} \neq 0)$:

$$
\begin{aligned}
T^{\underline{u_{ij}}=0} J(\tilde{X}) =& J(X_1, X_2, \underline{a_1} = 0, \underline{a_2} \neq 0) \\
T^{(u_{21}=1, u_{i2}=0)} J(\tilde{X}) =& J(X_1, X_2 - 1, \underline{a_1} = e_2, \underline{a_2} \neq 0) + p_{12}
\end{aligned}
$$

C.2 Proofs

C.2.1 Non-Idling Policy

In this section, we show that IW j should not be idled when a patient of class j (i.e., a patient whose primary IW is j) is boarded in the ED. We first establish the following monotonicity result.

Lemma 2 (Monotonicity) *For any $\tilde{X} \in \mathcal{S}$, $n \in \mathbb{Z}^+$, $\beta \in [0,1)$, and $k \in N_p$: $V_{n,\beta}(\underline{X}+e_k, \underline{a_1}, \underline{a_2}) \geq V_{n,\beta}(\underline{X}, \underline{a_1}, \underline{a_2})$, where $V_{n,\beta}(\cdot)$ represents the n-period discounted cost when the discount factor is β.*

Proof of Lemma 2

Similar to (4.4), the finite-horizon discounted cost optimality equation can be written as:

$$V_{n+1,\beta}(\tilde{X}) = \frac{1}{\psi}\Bigg[\underline{\theta}\ \underline{X}^T + \beta \min_{u=\underline{u_{ij}}\in\mathcal{U}(\tilde{X})} \Bigg\{ \sum_{i\in N_p}\sum_{j\in N_s} \lambda_i T^{\underline{u_{ij}}} V_{n,\beta}(\underline{X}+e_i, \underline{a_j}) + \sum_{i\in N_p}\sum_{j\in N_s}\sum_{l\in N_p} a_{lj}\mu_l T^{\underline{u_{ij}}} V_{n,\beta}(\underline{X}, \underline{a_j}-e_k) + \left(\psi - \sum_{i\in N_p}\lambda_i - \sum_{k\in N_p} a_{kj}\mu_k\right)\sum_{j\in N_s} V_{n,\beta}(\tilde{X})\Bigg\}\Bigg], \tag{C.1}$$

where $V_{n,\beta}(\tilde{X})$ is the optimal cost of the n-period problem starting at state $\tilde{X}$, along with terminal condition $V_{0,\beta}(\tilde{X}) = 0$ for every $\tilde{X} \in \mathcal{S}$. We prove this lemma by induction on n. For $n = 0$, we have $V_{0,\beta}(\tilde{X})$=0. Hence, $V_{0,\beta}(\underline{X}+e_k, \underline{a_1}, \underline{a_2}) = V_{0,\beta}(\underline{X}, \underline{a_1}, \underline{a_2}) = 0$ $(\forall \tilde{X} \in \mathcal{S}, \forall \beta \in [0,1), \forall k \in N_p)$. Assume that, for some $n \in \mathbb{Z}^+$, the required condition holds: $V_{n,\beta}(\underline{X}+e_k, \underline{a_1}, \underline{a_2}) \geq V_{n,\beta}(\underline{X}, \underline{a_1}, \underline{a_2})$ for any $\tilde{X} \in \mathcal{S}$, $\beta \in [0,1)$ and $k \in N_p$. We now show that the same condition holds for $n+1$. From

(C.1), we have:

$$
\begin{aligned}
V_{n+1,\beta}(\underline{X}+e_k,\underline{a}_1,\underline{a}_2) = \\
&\frac{1}{\psi}\Bigg[\underline{\theta}(\underline{X}+e_k)^T+\beta\Bigg\{\sum_{j\in N_s}\Big(\lambda_1 T^{\underline{u_{1j}}}V_n(\underline{X}+e_1+e_k,\underline{a}_1,\underline{a}_2)\\
&+\lambda_2 T^{\underline{u_{2j}}}V_n(\underline{X}+e_2+e_k,\underline{a}_1,\underline{a}_2)\Big)\\
&+\sum_{i\in N_p}\sum_{j\in N_s}\sum_{l\in N_p}a_{jl}\Big(p_{ij}+\mu_l V_n(\underline{X}+e_k-e_i,\underline{a_j}-e_l+e_i)\Big)\\
&+\left(\psi-\lambda_1-\lambda_2-\sum_{j\in N_s}\sum_{l\in N_p}a_{jl}\mu_l\right)V_n(\underline{X}+e_k,\underline{a}_1,\underline{a}_2)\Bigg\}\Bigg].
\end{aligned}
\tag{C.2}
$$

If the set of admissible actions are the same for both of the states $(\underline{X}+e_k,\underline{a}_1,\underline{a}_2)$ and $(\underline{X},\underline{a}_1,\underline{a}_2)$, the proof is straightforward, and follows directly from (C.2). However, since $\mathcal{U}(\underline{X}+e_k,\underline{a}_1,\underline{a}_2)$ can be a larger admissible set than $\mathcal{U}(\underline{X},\underline{a}_1,\underline{a}_2)$, the optimal action $u^*\in\mathcal{U}(\underline{X}+e_k,\underline{a}_1,\underline{a}_2)$ may not belong to $\mathcal{U}(\underline{X},\underline{a}_1,\underline{a}_2)$. If $u^*\notin\mathcal{U}(\underline{X},\underline{a}_1,\underline{a}_2)$, WLOG assume that $k=1$, and observe that the only possibility for $u^*\notin\mathcal{U}(\underline{X},\underline{a}_1,\underline{a}_2)$ is that queue 1 is empty at state $(\underline{X},\underline{a}_1,\underline{a}_2)$. We show that, if the same allocation policy u^* is used at this state but the IW that is assigned to Class 1 patients under u^* (say IW 1) is idled, and $\underline{X}+e_1$ is swapped with $\underline{X}$, a lower (or equal) value than $V_{n+1,\beta}(\underline{X}+e_k,\underline{a}_1,\underline{a}_2)$ can be obtained. That is, following a suboptimal policy at at state $(\underline{X},\underline{a}_1,\underline{a}_2)$ yields a cost that is not higher than $V_{n+1,\beta}(\underline{X}+e_k,\underline{a}_1,\underline{a}_2)$. The proof is then established, because $V_{n+1,\beta}(\underline{X},\underline{a}_1,\underline{a}_2)$ is the optimal cost at state $(\underline{X},\underline{a}_1,\underline{a}_2)$.

First, rewrite the formulation by separating the action related to Class 1.

$$
\begin{aligned}
V_{n+1,\beta}(\underline{X}+e_1,\underline{a_1},\underline{a_2}) =& \frac{1}{\psi}\Bigg[\underline{\theta}(\underline{X}+e_1)^T \\
&+\beta\Bigg\{\sum_{j\in N_s}\Big(\lambda_1 T^{\underline{u_{1j}}}V_n(\underline{X}+e_1+e_1,\underline{a_j}) \\
&+\lambda_2 T^{\underline{u_{2j}}}V_n(\underline{X}+e_2+e_1,\underline{a_j})\Big) \\
&+\sum_{i\in N_p}\sum_{k\in N_p} a_{1k}\left(p_{i1}+\mu_k(V_n\right. \\
&\left.-V_n(\underline{X}+e_1,\underline{a_1},\underline{a_2}))\right) \\
&+\sum_{i\in N_p} a_{2l}\left(p_{i2}+\mu_l V_n(\underline{X}+e_1-e_i,\underline{a_1},\underline{a_2}-e_l+e_i)\right) \\
&+\left(\psi-\lambda_1-\lambda_2-\sum_{l\in N_p} a_{jl}\mu_l\right)V_n(\underline{X}+e_1,\underline{a_1},\underline{a_2})\Bigg\}\Bigg] \qquad \text{(C.3)}
\end{aligned}
$$

From the induction assumption, we have: $\left(\psi-\lambda_1-\lambda_2-\sum_{l\in N_p} a_{jl}\mu_l\right)\times[V_n(\underline{X}+e_1,\underline{a_1},\underline{a_2})-V_n(\underline{X},\underline{a_1},\underline{a_2})]\geq 0$. Now, subtracting this positive term from (C.3), we have:

$$
\begin{aligned}
V_{n+1,\beta}(\underline{X}+e_1,\underline{a_1},\underline{a_2}) \geq& \frac{1}{\psi}\Bigg[\underline{\theta}(\underline{X}+e_1)^T \\
&+\beta\Bigg\{\sum_{j\in N_s}\Big(\lambda_1 T^{\underline{u_{1j}}}V_n(\underline{X}+e_1+e_1,\underline{a_j}) \\
&+\lambda_2 T^{\underline{u_{2j}}}V_n(\underline{X}+e_2+e_1,\underline{a_j})\Big) \\
&+\sum_{i\in N_p}\sum_{k\in N_p} a_{1k}\left(p_{i1}+\mu_k(V_n(\underline{X}+e_1-e_i,\underline{a_1}-e_k+e_i,\underline{a_2})\right. \\
&\left.-V_n(\underline{X}+e_1,\underline{a_1},\underline{a_2}))\right) \\
&+\sum_{i\in N_p}\sum_{l\in N_p} a_{2l}\left(p_{i2}+\mu_l V_n(\underline{X}+e_1-e_i,\underline{a_1},\underline{a_2}-e_l+e_i)\right) \\
&+\left(\psi-\lambda_1-\lambda_2-\sum_{l\in N_p} a_{jl}\mu_l\right)V_n(\underline{X},\underline{a_1},\underline{a_2})\Bigg\}\Bigg] \qquad \text{(C.4)}
\end{aligned}
$$

From the induction assumption, we can write:

$$
\begin{aligned}
V_{n+1,\beta}(\underline{X}+e_1,\underline{a_1},\underline{a_2}) \geq \frac{1}{\psi}\Bigg[&\underline{\theta X}^T+\beta\Bigg\{\sum_{j\in N_s}\Big(\lambda_1 T^{\underline{u_{1j}}}V_n(\underline{X}+e_1,\underline{a_j})\\
&+\lambda_2 T^{\underline{u_{2j}}}V_n(\underline{X}+e_2,\underline{a_j})\Big)\\
&+\sum_{i\in N_p}\sum_{k\in N_p}\mathbb{1}(a_{1k}=1)\left(p_{i1}+\mu_k(V_n(\underline{X}-e_i,\underline{a_1}-e_k+e_i,\underline{a_2})\right.\\
&\left.-V_n(\underline{X}+e_1,\underline{a_1},\underline{a_2})\right)\\
&+\sum_{i\in N_p}\sum_{l\in N_p}a_{2l}\left(p_{i2}+\mu_l V_n(\underline{X}-e_i,\underline{a_1},\underline{a_2}-e_l+e_i)\right)\Big)\\
&+\left(\psi-\lambda_1-\lambda_2-\sum_{l\in N_p}a_{jl}=1\mu_l\right)V_n(\underline{X},\underline{a_1},\underline{a_2})\Bigg\}\Bigg].
\end{aligned}
\tag{C.5}
$$

Next, we show that (C.5) provides an upper bound for $V_{n+1,\beta}(\underline{X},\underline{a_1},\underline{a_2})$. To observe this, consider an admissible (but not necessarily optimal) policy that idles the server allocated to Class 1, and use the same allocation as u^* for Class 2. This yields:

$$
\begin{aligned}
&V_{n+1,\beta}(\underline{X}+e_1,\underline{a_1},\underline{a_2}) \geq\\
&\frac{1}{\psi}\Bigg[\underline{\theta X}^T+\beta\Bigg\{\sum_{j\in N_s}\Big(\lambda_1 T^{\underline{u_{1j}}}V_n(\underline{X}+e_1,\underline{a_j})+\lambda_2 T^{\underline{u_{2j}}}V_n(\underline{X}+e_2,\underline{a_j})\Big)\\
&+\sum_{i\in N_p}\sum_{k\in N_p}a_{1k}\mu_k\left(V_n(\underline{X},0,\underline{a_2})-V_n(\underline{X}+e_1,\underline{a_1},\underline{a_2})\right)\\
&+\sum_{i\in N_p}\sum_{l\in N_p}a_{2l}\left(p_{i2}+\mu_l V_n(\underline{X}-e_i,\underline{a_1},\underline{a_2}-e_l+e_i)\right)\\
&+\left(\psi-\lambda_1-\lambda_2-\sum_{l\in N_p}a_{jl}\mu_l\right)V_n(\underline{X},\underline{a_1},\underline{a_2})\Bigg\}\Bigg].
\end{aligned}
\tag{C.6}
$$

Since this policy is an admissible (but not necessarily optimal) policy, it provides an upper bound for $V_{n+1,\beta}(\underline{X},\underline{a_1},\underline{a_2})$, which completes the proof.

Proposition 2 (Non-idling) *There exists an optimal policy which does not allow idling any IW j when there is a patient of class j boarded in the ED.*

Proof of Proposition 2

Let $\pi^{'}$ be a policy that allows idling IW j when IW j is available, and the queue of Class j patients is not empty. Construct another policy π^* that follows the same allocation strategy as $\pi^{'}$, but assigns patients of class j to IW j whenever IW j is available and the queue of Class j patients is not empty. We need to show that cost of policy π^* is higher than π^*. This requires us to show that the following property holds for every n and every state:

$$V_{n,\beta}^{\pi^*}(X - e_2, \underline{a_1}, \underline{a_2} = e_2) \leq V_{n,\beta}^{\pi^{'}}(X, \underline{a_1}, \underline{0}), \tag{C.7}$$

or

$$V_{n,\beta}^{\pi^*}(X - e_1, \underline{a_1} = e_1, \underline{a_2}) \leq V_{n,\beta}^{\pi^{'}}(X, \underline{0}, \underline{a_2}). \tag{C.8}$$

WLOG assume that $j = 1$. Since for $n = 0$ we have $V_{0,\beta}^{\pi^*}(\tilde{X})$=$V_{0,\beta}^{\pi^{'}}(\tilde{X})$=0, $V_{0,\beta}^{\pi^*}(\underline{X} - e_1, \underline{a_1} = e_1, \underline{a_2}) = V_{0,\beta}^{\pi^{'}}(\underline{X}, 0, \underline{a_2}) = 0\ \forall \tilde{X} \in \mathcal{S}$. Assume that, for some $n \in \mathbb{Z}^+$, property (C.8) holds for all $\tilde{X} \in \mathcal{S}$, $\beta \in [0, 1)$ and $k \in N_p$. We now show that the same condition holds for n+1. From (C.1) we have the following equations:

$$\begin{aligned}
&V_{n+1,\beta}^{\pi^*}(\underline{X} - e_1, \underline{a_1} = e_1, \underline{a_2}) = \\
&\frac{1}{\psi}\Bigg[\underline{\theta}(\underline{X} - e_1)^T + \beta\Bigg\{\sum_{j \in N_s}\Big(\lambda_1 T^{\underline{u_{1j}}} V_n^{\pi^*}(\underline{X}, \underline{a_1} = e_1, \underline{a_2}) \\
&+ \lambda_2 T^{\underline{u_{2j}}} V_n^{\pi^*}(\underline{X} - e_1 + e_2, \underline{a_1} = e_1, \underline{a_2})\Big) \\
&+ \mu_1 V_n^{\pi^*}(\underline{X} - 2e_1, \underline{a_1} = e_1, \underline{a_2}) \\
&+ \sum_{i \in N_p}\sum_{l \in N_p} a_{2l}\left(p_{i2} + \mu_l V_n^{\pi^*}(\underline{X} - e_1 - e_i, \underline{a_1} = e_1, \underline{a_2} - e_l + e_i)\right) \\
&+ \left(\psi - \lambda_1 - \lambda_2 - \mu_1 - \sum_{l \in N_p} a_{2l}\mu_l\right) V_n^{\pi^*}(\underline{X} - e_1, \underline{a_1} = e_1, \underline{a_2})\Bigg\}\Bigg],
\end{aligned} \tag{C.9}$$

$$V^{\pi'}_{n+1,\beta}(\underline{X},0,\underline{a_2}) = \frac{1}{\psi}\Bigg[\underline{\theta X}^T + \beta\Bigg\{ \sum_{j\in N_s}\Big(\lambda_1 T^{\underline{u_{1j}}} V^{\pi'}_n(\underline{X},0,\underline{a_2})$$
$$+ \lambda_2 T^{\underline{u_{2j}}} V^{\pi'}_n(\underline{X}+e_2,0,\underline{a_2})\Big)$$
$$+ \sum_{i\in N_p}\sum_{l\in N_p} a_{2l}\Big(p_{i2} + \mu_l V^{\pi'}_n(\underline{X}-e_i,0,\underline{a_2}-e_l+e_i)\Big)$$
$$+ \left(\psi - \lambda_1 - \lambda_2 - \sum_{l\in N_p} a_{2l}\mu_l\right) V^{\pi'}_n(\underline{X},0,\underline{a_2})\Bigg\}\Bigg]. \tag{C.10}$$

Rewriting (C.9) by separating the actions related to Class 1 departure, and subtracting the result from (C.10), we have:

$$V^{\pi'}_{n+1,\beta}(X,\underline{0},\underline{a_2}) - V^{\pi^*}_{n+1,\beta}(X-e_1,\underline{a_1}=e_1,\underline{a_2}) =$$
$$\frac{1}{\psi}\Bigg[\theta_1 + \beta\Bigg\{\mu_l\Big(V^{\pi'}_n(X_1,X_2-1,0,a_2)$$
$$-V^{\pi^*}_n(X_1-1,X_2-1,\underline{a_1}=e_1,a_2)\Big)$$
$$+ \mu_1\Big(V^{\pi^*}_n(X_1-1,X_2,\underline{a_1}=e_1,a_2) - V^{\pi^*}_n(X_1-2,X_2,\underline{a_1}=e_1,a_2)\Big) \tag{C.11}$$
$$+ \left(\psi - \lambda_1 - \lambda_2 - \sum_{l\in N_p} a_{jl}\mu_l\right)\Big(V^{\pi'}_n(X_1,X_2,0,a_2)$$
$$-V^{\pi^*}_n(X_1-1,X_2,\underline{a_1}=e_1,a_2)\Big)\Bigg\}\Bigg] \geq 0 \tag{C.12}$$

The inequality follows from the induction assumption as well as the monotonicity of the value function, and shows that (C.8) holds for $n+1$.

Proof of Proposition 1

Consider a primary-secondary patient and IW pair say IW 1 and IW 2. WLOG assume that $\mu_1 \geq \mu_2$. We prove the claim for time/period T+1 assuming that it holds for all t≤T. For T=0, the claim trivially holds. Consider the case that the expended service time of all patient types are equal to 0 at time t=1. Suppose that

there are patients of both classes waiting in the ED at t=1. Let π be an optimal policy. Then, by the induction hypothesis, we assume that π follows the $\theta\mu$ rule from t=2 on. Now suppose that π selects a patient of Class 2 at t=1. Since the service discipline is non-preemptive, π may select a of patient Class 1 only after the service completion of the patient of Class 2. Let $\bar{\pi}$ be the policy that is identical to π except that it interchanges the first time the Class 1 and Class 2 patients are served. The rest of the decisions of $\bar{\pi}$ are the same as those of π. From those defined above, when $p_{ij} = 0 \ \ \forall i \in N_p, j \in N_s$:

$$J_{\pi}(\tilde{X}, T) - J_{\bar{\pi}}(\tilde{X}, T) = \theta_1\mu_1 - \theta_2\mu_2. \tag{C.13}$$

If $\theta_1\mu_1 \geq \theta_2\mu_2$, the equation contradicts the assumption that π is an optimal policy.

Proof of Theorem 1

For the case where $X_1 > 0$, the result is straightforward since assigning primary type patients to IW 1 does not incur any penalty cost. From Proposition 1, we know that under no penalty cost, it is optimal to follow the $c\mu$ priority policy. Hence, when we include the penalty costs only assignment of Class 2 patients becomes more costly. Therefore, it is optimal to assign Class 1 patients to IW 1 whenever they are boarded in ED.

To prove the optimality of a threshold policy for IW 1 when $X_1 = 0$, we need to show that the difference $V_n(0, X_2 - 1, \underline{a_1} = e_2, \underline{a_2}) - V_n(0, X_2, \underline{a_1} = 0, \underline{a_2})$ is decreasing in X_2, so that assigning Class 2 patients to IW 1 becomes desirable at some level of Class 2 queue length, in spite of the associated penalty cost. Notice that whenever

$X_1 > 0$, IW 1 serves Class 1 patients. Observe that

$$\begin{aligned} &V_n(0, X_2 - 1, \underline{a_1} = e_2, \underline{a_2}) - V_n(0, X_2, \underline{a_1} = 0, \underline{a_2}) \\ &\geq V_n(0, X_2, \underline{a_1} = e_2, \underline{a_2}) - V_n(0, X_2 + 1, \underline{a_1} = 0, \underline{a_2}). \end{aligned} \tag{C.14}$$

We can rewrite the system state by dropping $\underline{a_2}$, since our focus is on times when there is no Class 1 patient boarded in the ED and IW 1 is available. Thus, we rewrite the above inequality as:

$$V_{n,\beta}(\underline{X}, 0) - V_{n,\beta}(\underline{X} - e_2, e_2) \leq V_{n,\beta}(\underline{X} + e_2, 0) - V_{n,\beta}(\underline{X}, e_2), \tag{C.15}$$

which is the same structure that is introduced in Koole G (1995) [1] . Following the proof in Koole G (1995), we define a set of functions $\mathcal{F}$ that satisfy:

$$f(\underline{X}, 0) + f(\underline{X}, e_2) \leq f(\underline{X} + e_2, 0) + f(\underline{X} - e_2, e_2),$$

where $f \in \mathcal{F}$ for all $\underline{X} > \underline{0}$. Now, we assume that $V_{n,\beta} \in \mathcal{F}$, and show that $V_{n+1,\beta} \in \mathcal{F}$ (note that trivially $V_0 \in \mathcal{F}$). Define $min_{u \in \mathcal{U}(\tilde{X})}(T^{u_{i1}} V_{n,\beta}(\tilde{X}))$ as $W_{n,\beta}(\tilde{X})$ and observe that $W_{n,\beta} \in \mathcal{F}$. Assume that the optimal action at state $(\underline{X} + e_2, 0)$ is assigning Class 2 to IW 1. We have:

$$\begin{aligned} &W_{n,\beta}(\underline{X}, 0) + W_{n,\beta}(\underline{X}, e_2) \leq V_{n,\beta}(\underline{X} - e_2, e_2) + V_{n,\beta}(\underline{X}, e_2) \\ &= W_{n,\beta}(\underline{X} - e_2, 0) + W_{n,\beta}(\underline{X} + e_2, e_2), \end{aligned}$$

since action of assigning Class 2 to IW 1 is suboptimal at state $(\underline{X}, e_2)$. Now assume that the optimal action at state $(\underline{X} + e_2, 0)$ is keeping IW 1 idle. We have:

$$\begin{aligned} W_{n,\beta}(\underline{X}, 0) + W_{n,\beta}(\underline{X}, e_2) &\leq V_{n,\beta}(\underline{X}, 0) + V_{n,\beta}(\underline{X}, e_2) \\ &\leq V_{n,\beta}(\underline{X} + e_2, 0) + V_{n,\beta}(\underline{X} - e_2, e_2), \end{aligned}$$

[1] Koole G (1995). A simple proof of the optimality of a threshold policy in a two-server queueing system. *S*ystems & Control Letters 26(5):301–303.

since $V_n \in \mathcal{F}$. Notice that since idleness is the optimal action at state$(\underline{X} + e_2, 0)$, we can rewrite the last part of the inequality as $W_n(\underline{X} + e_2, 0) + W_n(\underline{X} - e_2, e_2)$, which in turn shows that $W_n \in \mathcal{F}$. If we rewrite V_{n+1} as:

$$V_{n+1,\beta}(\tilde{X}) = \frac{1}{\psi}\Big[\underline{\theta X}^T + \lambda_1 W_{n,\beta}(X + e_1, \underline{a_1}) + \lambda_2 W_{n,\beta}(X + e_2, \underline{a_1}) \tag{C.16}$$
$$+ \sum_{k\in N_p}\sum_{i\in N_p} a_{k1}\mu_k W_{n,\beta}(X, \underline{a_1} - e_k) + (\psi - \sum_{i\in N_p}\lambda_i - \sum_{k\in N_p} a_{k1}\mu_k)V_{n,\beta}(\tilde{X})\Big],$$

we can conclude that $V_{n+1,\beta} \in \mathcal{F}$ from the induction assumption and the fact that $W_{n,\beta} \in \mathcal{F}$.

Proof of Lemma 1

We need to show that if $J \in \mathcal{F}$ then $TJ \in \mathcal{F}$ where $TJ(\underline{Y}) = T_\theta J(\underline{Y}) + \beta\left(T_a J(\underline{Y}) + T_* J(\underline{Y})\right)$. Note that T_θ and T_a trivially satisfy properties (4.14) and (4.15). Thus, it is sufficient to show that operator T_* preserves properties (4.14) and (4.15). Assume $J \in \mathcal{F}$, $\theta_1\mu_1 \geq \theta_2\mu_2$ and $\mu_2 \geq \mu_1$ hold. To show the preservation of property (i), we need to examine all possible actions at states (Y), $(Y - e_1)$, $(Y - e_2)$, $(Y + e_1)$, and $(Y + e_1 - e_2)$ by using the induction assumption. Notice that there are 2^5 possible cases; however, properties (4.14) and (4.15) restrict several cases, which leave us with the cases shown in Table C.1. This table also shows the patient class that is assigned to IW 2. We next consider each of the case shown in Table C.1 separately.

Case 1 Note that the set of actions that are defined in Case 1 are feasible when $Y_1 \geq 2$. Consider the state $Y_1 = 1, Y_2 \geq 1$, and $u_{22} = 1$ as a feasible (not necessarily optimal) action for state $Y - e_1$. See Case 7 for the action where patient Class 2 is

Table C.1: Possible Actions

Cases	$Y+e_1$	Y	$Y-e_2$	$Y-e_1$	$Y+e_1-e_2$
Case 1	1	1	1	1	1
Case 2	2	2	2	2	2
Case 3	1	2	1	2	1
Case 4	1	2	2	2	1
Case 5	2	2	2	2	1
Case 6	2	2	1	2	1
Case 7	1	1	1	2	1

assigned to IW 2 at state $Y-e_1$. We have

$$
\begin{aligned}
&[\tilde{\mu_1}\Delta_1 T_* J(Y) - \tilde{\mu_2}\Delta_2 T_* J(Y+e_1-e_2)] - [\tilde{\mu_1}\Delta_1 T_* J(Y-e_1) - \tilde{\mu_2}\Delta_2 T_* J(Y-e_2)] \\
&= (1-\Lambda-\tilde{\mu_1})\left([\tilde{\mu_1}\Delta_1 T_* J(Y) - \tilde{\mu_2}\Delta_2 T_* J(Y+e_1-e_2)]\right. \\
&\left.- [\tilde{\mu_1}\Delta_1 T_* J(Y-e_1) - \tilde{\mu_2}\Delta_2 T_* J(Y-e_2)]\right) \\
&+ \tilde{\mu_1}\left[\tilde{\mu_1}\Delta_1 T_* J(Y-2e_1) - \tilde{\mu_2}\Delta_2 T_* J(Y-e_1-e_2)\right] \geq 0.
\end{aligned}
$$

The inequality in the first line follows from nonnegativity of the term $(1-\Lambda-\tilde{\mu_1})$ and the induction assumptions. The second line follows from the optimality of $c\mu$ rule when $p_{ij}=0$.

Case 2 The proof of Case 2 can be established similar to that of Case 1 by replacing $\tilde{\mu_1}$ by $\tilde{\mu_2}$. Again, similar to Case 1, assigning Class 2 patients to IW 2 is feasible when $Y_2 \geq 2$. If the state is $Y_1 \geq 1, Y_2 = 1$, the action of assigning Class 2 patients to IW 2 is not feasible for the states $Y-e_2$ and $Y+e_1-e_2$. However, the action of assigning Class 1 patients to IW 2 is a feasible (not necessarily optimal) at these states (see Case 6 for case of assigning Class 1 patients at states $Y-e_2$ and $Y+e_1-e_2$).

Case 3 Note that

$$\begin{aligned}
\tilde{\mu_1}\Delta_1 T_* J(Y) - \tilde{\mu_2}\Delta_2 T_* J(Y+e_1-e_2) = & \\
& (1-\Lambda)\left[\tilde{\mu_1}\Delta_1 J(Y) - \tilde{\mu_2}\Delta_2 J(Y+e_1-e_2)\right] \\
& -\mu_1\left[\tilde{\mu_1}\Delta_1 J(Y) - \tilde{\mu_2}\Delta_2 J(Y+e_1-e_2) - p_{12}\right] \\
& \geq (1-\Lambda)\left[\tilde{\mu_1}\Delta_1 J(Y) - \tilde{\mu_2}\Delta_2 J(Y+e_1-e_2)\right] \\
& -\tilde{\mu_2}\left[\tilde{\mu_1}\Delta_1 J(Y) - \tilde{\mu_2}\Delta_2 J(Y+e_1-e_2) - p_{12}\right] \\
& \left[\tilde{\mu_1}\Delta_1 T_* J(Y-e_1) - \tilde{\mu_2}\Delta_2 T_* J(Y-e_2)\right] \\
& = (1-\Lambda)\left[\tilde{\mu_1}\Delta_1 J(Y-e_1) - \tilde{\mu_2}\Delta_2 J(Y-e_2)\right] \\
& -\tilde{\mu_2}\left[\tilde{\mu_1}\Delta_1 J(Y-e_1) - \tilde{\mu_2}\Delta_2 J(Y-e_2) - p_{12}\right].
\end{aligned}$$

Subtract the last term from the second term, we observe that

$$\begin{aligned}
(1-\Lambda-\tilde{\mu_2})\left[(\tilde{\mu_1}\Delta_1 J(Y) - \tilde{\mu_2}\Delta_2 J(Y+e_1-e_2))\right. & \\
\left. -(\tilde{\mu_1}\Delta_1 J(Y-e_1) - \tilde{\mu_2}\Delta_2 J(Y-e_2))\right] \geq 0, &
\end{aligned}$$

since $J \in \mathcal{F}$ and the fact that $(1-\Lambda-\tilde{\mu_2})$ is nonnegative.

Case 4 Assigning Class 2 patients to IW 2 is feasible when $Y_2 \geq 2$. If the state is $Y_1 \geq 1, Y_2 = 1$, the action of assigning Class 2 patients to IW 2 is not feasible for the state $Y - e_2$. However, the action of assigning Class 1 patients to IW 2 is feasible (not necessarily optimal) at this state (see Case 3 for case of assigning Class 1 patients at

state $Y - e_2$). We have

$$\begin{aligned}
&\tilde{\mu_1}\Delta_1 T_* J(Y) - \tilde{\mu_2}\Delta_2 T_* J(Y + e_1 - e_2) \\
&\qquad \leq (1 - \Lambda - \tilde{\mu_2}) \left[\tilde{\mu_1}\Delta_1 J(Y) - \tilde{\mu_2}\Delta_2 J(Y + e_1 - e_2)\right] \\
&\qquad + \tilde{\mu_1} p_{12} \tilde{\mu_1}\Delta_1 T_* J(Y - e_1) - \tilde{\mu_2}\Delta_2 T_* J(Y - e_2) \\
&\qquad = (1 - \Lambda - \tilde{\mu_2}) \left[\mu_1\Delta_1 J(Y - e_1) - \tilde{\mu_2}\Delta_2 J(Y - e_2)\right] \\
&\qquad + \tilde{\mu_2} \left[\tilde{\mu_1}\Delta_1 J(Y - e_1 - e_2) - \tilde{\mu_2}\Delta_2 J(Y - 2e_2) - p_{12}\right] \\
&\qquad = (1 - \Lambda - \tilde{\mu_2}) \left[\left[\tilde{\mu_1}\Delta_1 J(Y) - \tilde{\mu_2}\Delta_2 J(Y + e_1 - e_2)\right]\right. \\
&\qquad \left. - \left[\tilde{\mu_1}\Delta_1 J(Y - e_1) - \tilde{\mu_2}\Delta_2 J(Y - e_2)\right]\right] + p_{12}\tilde{\mu_1} \\
&\qquad \geq \mu_2 \left[\tilde{\mu_1}\Delta_1 J(Y - e_1 - e_2) - \tilde{\mu_2}\Delta_2 J(Y - 2e_2) - p_{12}\right],
\end{aligned}$$

where the inequality follows from $J \in \mathcal{F}$, nonnegativity of term $p_{12}\mu_1$, and optimality of assigning patient Class 2 at state $(Y - e_2)$.

Case 5 Assigning Class 2 patients to IW 2 is feasible when $Y_2 \geq 2$. If the state is $Y_1 \geq 1, Y_2 = 1$, the action of assigning Class 2 patients to IW 2 is not feasible at state $Y - e_2$. However, the action of assigning Class 1 patients to IW 2 is a feasible (not necessarily optimal) at this state (see Case 6 for case of assigning Class 1 patients for

the state $Y - e_2$). We have

$$\begin{aligned}
\tilde{\mu_1}\Delta_1 T_* J(Y) - \tilde{\mu_2}\Delta_2 T_* J(Y + e_1 - e_2) = \\
&(1 - \Lambda - \tilde{\mu_2})\left[\tilde{\mu_1}\Delta_1 J(Y) - \tilde{\mu_2}\Delta_2 J(Y + e_1 - e_2)\right] \\
&+ \tilde{\mu_2} p_{12}\tilde{\mu_1}\Delta_1 T_* J(Y - e_1) - \tilde{\mu_2}\Delta_2 T_* J(Y - e_2) = \\
&(1 - \Lambda - \tilde{\mu_2})\left[\tilde{\mu_1}\Delta_1 J(Y - e_1) - \tilde{\mu_2}\Delta_2 J(Y - e_2)\right] \\
&+ \tilde{\mu_2}(\tilde{\mu_1}\Delta_1 J(Y - e_1 - e_2) - \tilde{\mu_2}\Delta_2 J(Y - 2e_2)) \\
&[\tilde{\mu_1}\Delta_1 T_* J(Y) - \tilde{\mu_2}\Delta_2 T_* J(Y + e_1 - e_2)] \\
&- [\tilde{\mu_1}\Delta_1 T_* J(Y - e_1) - \tilde{\mu_2}\Delta_2 T_* J(Y - e_2)] \\
&= (1 - \Lambda - \tilde{\mu_2})\left[[\tilde{\mu_1}\Delta_1 J(Y) - \tilde{\mu_2}\Delta_2 J(Y + e_1 - e_2)]\right. \\
&\left.- [\tilde{\mu_1}\Delta_1 J(Y - e_1) - \tilde{\mu_2}\Delta_2 J(Y - e_2)]\right] \\
&- \tilde{\mu_2}\left[\tilde{\mu_1}\Delta_1 J(Y - e_1 - e_2) - \tilde{\mu_2}\Delta_2 J(Y - 2e_2) - p_{12}\right] \geq 0,
\end{aligned}$$

where the last inequality follows from the optimality of assignment of patient Class 2 at state $(Y - e_2)$.

Case 6 In this case, we have

$$\begin{aligned}
\tilde{\mu_1}\Delta_1 T_* J(Y) - \tilde{\mu_2}\Delta_2 T_* J(Y + e_1 - e_2) = \\
&(1 - \Lambda - \tilde{\mu_2})\left[\tilde{\mu_1}\Delta_1 J(Y) - \tilde{\mu_2}\Delta_2 J(Y + e_1 - e_2)\right] \\
&+ \tilde{\mu_2} p_{12}\tilde{\mu_1}\Delta_1 T_* J(Y - e_1) - \tilde{\mu_2}\Delta_2 T_* J(Y - e_2) \\
&= (1 - \Lambda - \tilde{\mu_2})\left[\tilde{\mu_1}\Delta_1 J(Y - e_1) - \tilde{\mu_2}\Delta_2 J(Y - e_2)\right] \\
&+ \tilde{\mu_2} p_{12}\left[\tilde{\mu_1}\Delta_1 T_* J(Y) - \tilde{\mu_2}\Delta_2 T_* J(Y + e_1 - e_2)\right] \\
&- [\tilde{\mu_1}\Delta_1 T_* J(Y - e_1) - \tilde{\mu_2}\Delta_2 T_* J(Y - e_2)] \\
&= (1 - \Lambda - \tilde{\mu_2})\left[[\tilde{\mu_1}\Delta_1 J(Y) - \tilde{\mu_2}\Delta_2 J(Y + e_1 - e_2)]\right. \\
&\left.- [\tilde{\mu_1}\Delta_1 T_* J(Y - e_1) - \tilde{\mu_2}\Delta_2 T_* J(Y - e_2)]\right] \geq 0,
\end{aligned}$$

since $J \in \mathcal{F}$.

Case 7 In this case, we have

$$[\tilde{\mu_1}\Delta_1 T_* J(Y) - \tilde{\mu_2}\Delta_2 T_* J(Y + e_1 - e_2)] - [\tilde{\mu_1}\Delta_1 T_* J(Y - e_1)$$
$$-\tilde{\mu_2}\Delta_2 T_* J(Y - e_2)] = (1 - \Lambda - \tilde{\mu_1}) [[\tilde{\mu_1}\Delta_1 J(Y) - \tilde{\mu_2}\Delta_2 J(Y + e_1 - e_2)]$$
$$- [\tilde{\mu_1}\Delta_1 T_* J(Y - e_1) - \tilde{\mu_2}\Delta_2 T_* J(Y - e_2)]] + \tilde{\mu_1} [\tilde{\mu_1}\Delta_1 J(Y - e_1)$$
$$-\tilde{\mu_2}\Delta_2 J(Y - e_2) - p_{12}] \geq 0,$$

where the inequality follows from $J \in \mathcal{F}$ and optimality of assigning the IW to Class 1 at state (Y).

To show the preservation of the second property, we need to consider all possible actions at the states $(Y), (Y - e_1), (Y + e_2)$. Again properties (4.14) and (4.15) restrict several cases, which leave us with the cases presented in Table C.2:

Table C.2: Possible Actions

Cases	$Y + e_2$	Y	$Y - e_1$	$Y - e_1 + e_2$	$Y - e_2$
Case 1	1	1	1	1	1
Case 2	2	2	2	2	2
Case 3	1	1	1	2	1
Case 4	1	1	2	2	1
Case 5	2	1	1	2	1
Case 6	2	1	2	2	1
Case 7	2	2	2	2	1

Case 1

Assigning Class 1 patients to IW 2 is feasible when $Y_1 \geq 2$. If the state is $Y_{1=1}, Y_2 \geq 1$,

the action of assigning Class 1 patients to IW 2 is not feasible at states $Y - e_1$ and $Y - e_1 + e_2$. However, the action of assigning Class 2 patients to IW 2 is a feasible (not necessarily optimal) at these states (see Case 4 for case of assigning Class 2 patients for the states $Y - e_1$ and $Y - e_1 + e_2$). We have

$$
\begin{aligned}
&[\tilde{\mu_1}\Delta_1 T_* J(Y - e_1) - \tilde{\mu_2}\Delta_2 T_* J(Y - e_2)] - [\tilde{\mu_1}\Delta_1 T_* J(Y - e_1 + e_2) \\
&-\tilde{\mu_2}\Delta_2 T_* J(Y)] = (1 - \Lambda - \tilde{\mu_1})\left[(\tilde{\mu_1}\Delta_1 J(Y - e_1) - \tilde{\mu_2}\Delta_2 J(Y - e_2))\right. \\
&\left. - (\tilde{\mu_1}\Delta_1 J(Y - e_1 + e_2) - \tilde{\mu_2}\Delta_2 J(Y))\right] \\
&+ \tilde{\mu_1}\left[(\tilde{\mu_1}\Delta_1 J(Y - 2e_2) - \tilde{\mu_2}\Delta_2 J(Y - e_1 - e_2))\right. \\
&\left. - (\tilde{\mu_1}\Delta_1 J(Y - 2e_1 + e_2) - \tilde{\mu_2}\Delta_2 J(Y - e_1))\right] \geq 0.
\end{aligned}
$$

The inequality follows from nonnegativity of the term $(1 - \Lambda - \tilde{\mu_1})$ and the fact that $J \in \mathcal{F}$.

Case 2

The proof of Case 2 can be shown similar to that of Case 1 by replacing $\tilde{\mu_1}$ by $\tilde{\mu_2}$, and by assigning Class 2 to the IW. Similar to Case 1, assigning Class 2 patients to IW 2 is feasible when $Y_2 \geq 2$. If the state is $Y_1 \geq 1, Y_2 = 1$, the action of assigning Class 2 patients to IW 2 is not feasible at state $Y - e_2$. However, the action of assigning Class 1 patients to IW 2 is feasible (not necessarily optimal) at this state (see Case 7 for case of assigning Class 1 patients for the state $Y - e_2$).

Case 3

Assigning Class 1 patients to IW 2 is feasible when $Y_1 \geq 2$. If the state is $Y_{1=1}, Y_2 \geq 1$, the action of assigning Class 1 patients to IW 2 is not feasible at state $Y - e_1$. However, the action of assigning Class 2 patients to IW 2 is feasible (not necessarily optimal) in this state (see Case 4 for case of assigning Class 2 patients for the state $Y - e_1$).

We have

$$[\tilde{\mu_1}\Delta_1 T_* J(Y-e_1) - \tilde{\mu_2}\Delta_2 T_* J(Y-e_2)] - [\tilde{\mu_1}\Delta_1 T_* J(Y-e_1+e_2)$$
$$-\tilde{\mu_2}\Delta_2 T_* J(Y)] \geq (1-\Lambda-\tilde{\mu_1})\left[(\tilde{\mu_1}\Delta_1 J(Y-e_1) - \tilde{\mu_2}\Delta_2 J(Y-e_2))\right.$$
$$\left.-(\tilde{\mu_1}\Delta_1 J(Y-e_1+e_2) - \tilde{\mu_2}\Delta_2 J(Y))\right]$$
$$+\tilde{\mu_1}\left[\tilde{\mu_1}\Delta_1 J(Y-2e_2) - \tilde{\mu_2}\Delta_2 J(Y-e_1-e_2) - p_{12}\right] \geq 0,$$

where he inequality follows from the fact that $J \in \mathcal{F}$ and the optimality of assigning Class 1 at state $(Y-e_1)$.

Case 4

We have

$$[\mu_1\Delta_1 T_* J(Y-e_1) - \mu_2\Delta_2 T_* J(Y-e_2)] - [\mu_1\Delta_1 T_* J(Y-e_1+e_2)$$
$$-\mu_2\Delta_2 T_* J(Y)] \geq (\bar{\psi}-\Lambda-\mu_1)\left[(\mu_1\Delta_1 J(Y-e_1) - \mu_2\Delta_2 J(Y-e_2))\right.$$
$$\left.-(\mu_1\Delta_1 J(Y-e_1+e_2) - \mu_2\Delta_2 J(Y))\right] \geq 0.$$

The inequality follows nonnegativity of the term $(1-\Lambda-\tilde{\mu_1})$ and the fact that $J \in \mathcal{F}$.

Case 5

Assigning Class 1 patients to IW 2 is feasible when $Y_1 \geq 2$. If the state is $Y_{1=1}, Y_2 \geq 1$, the action of assigning Class 1 patients to IW 2 is not feasible at state $Y-e_1$. However, the action of assigning Class 2 patients to IW 2 is feasible (not necessarily optimal) at this state (see Case 6 for case of assigning Class 2 patients for the state $Y-e_1$). We have

$$[\tilde{\mu_1}\Delta_1 T_* J(Y-e_1) - \tilde{\mu_2}\Delta_2 T_* J(Y-e_2)] - [\tilde{\mu_1}\Delta_1 T_* J(Y-e_1+e_2)$$
$$-\tilde{\mu_2}\Delta_2 T_* J(Y)] \geq (1-\Lambda)\left[(\tilde{\mu_1}\Delta_1 J(Y-e_1) - \tilde{\mu_2}\Delta_2 J(Y-e_2))\right.$$
$$\left.-(\tilde{\mu_1}\Delta_1 J(Y-e_1+e_2) - \tilde{\mu_2}\Delta_2 J(Y))\right]$$
$$-\tilde{\mu_2}\left[\tilde{\mu_1}\Delta_1 J(Y-e_1+e_2) - \tilde{\mu_2}\Delta_2 J(Y) - p_{12}\right] \geq 0,$$

where he inequality follows from the fact that $J \in \mathcal{F}$ and from the optimality of assigning the IW to Class 2 at state $(Y + e_2)$.

Case 6

We have

$$[\tilde{\mu_1}\Delta_1 T_* J(Y - e_1) - \tilde{\mu_2}\Delta_2 T_* J(Y - e_2)] - [\tilde{\mu_1}\Delta_1 T_* J(Y - e_1 + e_2)$$
$$-\tilde{\mu_2}\Delta_2 T_* J(Y)] \geq (1 - \Lambda - \tilde{\mu_2}) \left[(\tilde{\mu_1}\Delta_1 J(Y - e_1) - \tilde{\mu_2}\Delta_2 J(Y - e_2))\right.$$
$$\left. - (\tilde{\mu_1}\Delta_1 J(Y - e_1 + e_2) - \tilde{\mu_2}\Delta_2 J(Y))\right] \geq 0,$$

where the inequality follows from nonnegativity of the term $(1 - \Lambda - \tilde{\mu_2})$ and the fact that $J \in \mathcal{F}$.

Case 7

We have:

$$[\tilde{\mu_1}\Delta_1 T_* J(Y - e_1) - \tilde{\mu_2}\Delta_2 T_* J(Y - e_2)] - [\tilde{\mu_1}\Delta_1 T_* J(Y - e_1 + e_2)$$
$$-\tilde{\mu_2}\Delta_2 T_* J(Y)] = (1 - \Lambda - \tilde{\mu_2}) \left[(\tilde{\mu_1}\Delta_1 J(Y - e_1) - \tilde{\mu_2}\Delta_2 J(Y - e_2))\right.$$
$$\left. - (\tilde{\mu_1}\Delta_1 J(Y - e_1 + e_2) - \tilde{\mu_2}\Delta_2 J(Y))\right]$$
$$- \tilde{\mu_2} \left[\tilde{\mu_1}\Delta_1 J(Y - e_1) - \tilde{\mu_2}\Delta_1 J(Y - e_2) - p_{12}\right] \geq 0,$$

where the inequality follows from the fact that $J \in \mathcal{F}$ and from the optimality of assigning the IW to Class 2 at state (Y).

Additionally, we can gain further insights by using Lemma 1. The results above show that the threshold level depends on the model parameters, since the threshold can be defined as $\min\{Y_1 : [\tilde{\mu_1}\Delta_1 T_* J(Y - e_1) - \tilde{\mu_2}\Delta_2 T_* J(Y - e_2)] \geq p_{12}\}$. Using this, we can identify how the threshold level changes as the model parameters change. From Lemma 1, we observe that the difference $[\tilde{\mu_1}\Delta_1 T_* J(Y - e_1) - \tilde{\mu_2}\Delta_2 T_* J(Y - e_2)]$ is nondecreasing in the number of Class 1 patients in the queue. Also, consider $\hat{p_{12}}$

where $\hat{p_{12}} \geq p_{12}$. We can conclude that the threshold level increases as p_{12} increases, since

$$\min\{Y_1 : [\tilde{\mu_1}\Delta_1 T_* J(Y - e_1) - \tilde{\mu_2}\Delta_2 T_* J(Y - e_2)] \geq \hat{p_{12}}\} \geq$$
$$\min\{Y_1 : [\tilde{\mu_1}\Delta_1 T_* J(Y - e_1) - \tilde{\mu_2}\Delta_2 T_* J(Y - e_2)] \geq p_{12}\}.$$

Similar to the above results, the threshold level also depends the service rates μ_1 and μ_2. It can be observed that the difference $[\mu_1\Delta_1 T_* J(Y - e_1) - \mu_2\Delta_2 T_* J(Y - e_2)]$ is nondecreasing in μ_1 (nonincreasing in μ_2), which means that the threshold level increases (decrease) as these parameter increase. Proof of Theorem 2 directly follows from Lemma 1.

Computational Results

To gain insights into the structure of the optimal policy, we first generate a set of test cases. Table C.3 presents these cases. For each case, we solve our MDP numerically using a convergence criteria of 10^{-4}, and truncate the queue lengths at $X_1 = X_2 = 70$. To avoid the "boundary effects," i.e., when the number in each queue gets close to the boundary of the state space under consideration, we only present the optimal policy for states in which the number in queues are no more than 30.

Table C.3: Numerical Test Cases

Cases	λ_1	λ_2	μ_1	μ_2	θ_1	θ_2	p_1	p_2
1	1	1	2	1	1	1	1	1
2	1	1	2	1	1	1	1000	1000
3	1	1	2	1	1000	1000	1	1
4	1	1	2	1	1	1	10	1
5	1	1	2	1	1	1	1	10

In the Figures C.1-C.5 green (dark gray) color represents serving Class 1 patients, yellow (light gray) color represent serving Class 2 patients, and dark blue (black) represents idling the server. In Case 1, the effect of differences in the service rates is analyzed (Figure C.1). In Case 2, the effect of high penalty costs, and in Case 3, the effect of high holding costs is analyzed. In cases 4 and 5 (Figures C.4-C.5), the effect of the penalty cost on the optimal policy structure is analyzed. Our MDP-based

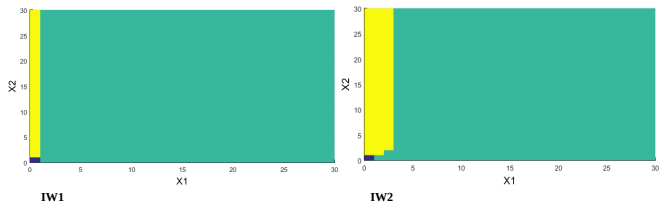


Figure C.1: Case 1

numerical results show that when $\theta_1\mu_1 \geq \theta_2\mu_2$, the structure of the optimal control policy is a state-dependent threshold-type policy, where IW 1 serves Class 1 when $X_1 > 0$, and IW 2 performs as a dedicated server, and switches to the $c\mu$ rule after the threshold. When $p_{ij} \gg \theta_i \quad (\forall i \in N_p, j \in N_s)$ (see, e.g., Case 2), both of the

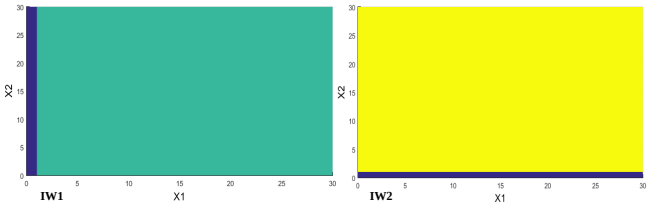


Figure C.2: Case 2

units start to work as dedicated units. Moreover, when their primary queue is empty, they idle, even if there is a patient in the non-primary queue waiting for assignment. Under $\theta_i \gg p_{ij} \quad (\forall i \in N_p, j \in N_s)$ (see, e.g., Case 3) the well-known $c\mu$ rule becomes the optimal policy. This policy gives strict priority to Class 1 patients whenever there

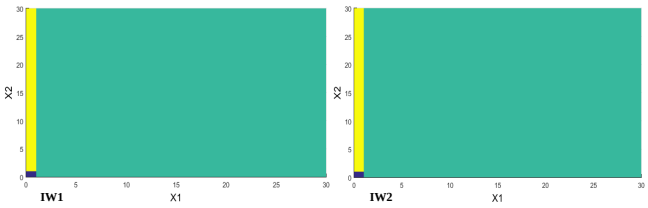


Figure C.3: Case 3

is a Class 1 patient waiting in the ED. We also observe that, when p_{12} increases, (a)

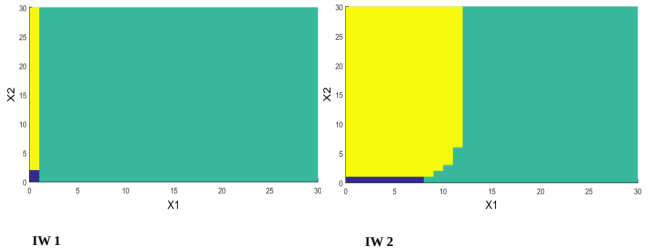


Figure C.4: Case 4

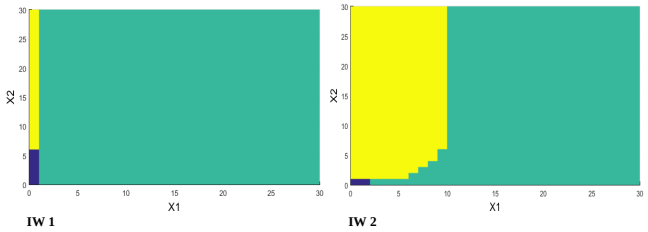


Figure C.5: Case 5

IW 2 delays serving Class 1 patients, (b) the threshold level increases, and (c) IW 1 still serves as a dedicated unit when $X_1 > 0$, and switches to serves Class 2 patients when $X_1 = 0$. When p_{21} increases IW 1 serves Class 2 patients when $X_1 = 0$ and $X_2 > T_2$, where T_2 is some threshold level on number of Class 2 patients boarded in the ED. In addition, as the threshold on Class 2 patients increase, we observe that the threshold for Class 1 patients in IW 2 increases.

C.3 Birth-and-Death Processes

In this appendix we present the illustrations of the birth-and-death processes used to construct the BDT heuristic policy discussed in the main body.

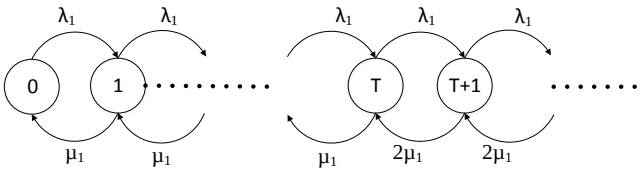


Figure C.6: Birth-and-Death Process Approximation for Class 1 Patients

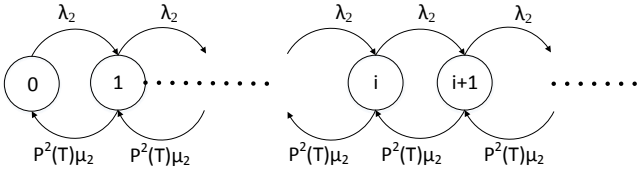


Figure C.7: Birth-and-Death Process Approximation for Class 2 Patients

C.4 Numerical Cases

We generate 216 problem instances. These problem instances cover various cost and arrival rate combinations. Tables C.4-C.6 provide a summary of the related information. More details are available upon request.

C.5 Data Analyses

In Tables C.7-C.8, we present a summary of some of the main results from our data analyses. More details are available upon request.

Table C.4: ROAE Cost (Per Time Unit) Combinations

θ_1	θ_2
1	1
2	1
1	2

Table C.5: Penalty Cost Combinations

p_{12}	p_{21}
0	0
1	1
10	10
100	100

Table C.6: Arrival Rate Combinations in the Test Suite

λ_1/μ_1	λ_2	μ_2
0.1-0.9	0.4	1
0.1-0.9	0.8	1

Table C.7: p-values for Comparison of Service Times for Primary and Secondary IWs

Patient Type	p-value
Type 1-CP	0.750
Type 2-CP	0.216
Type 1-CHF	0.601
Type 2-CHF	0.218

Table C.8: Average Service Time (in Days)

IW	ED admits	Direct admits	OR admits
4 West	3.57	7.41	4.05
5 West	3.60	4.38	2.93

C.6 Simulation Model

C.6.1 Cost Cases for Simulation

In our simulation model, we assume that the cost associated with the risk of adverse events that may occur while a patient is boarded in the ED is the same for both patient classes (patients requiring a bed from 4 West or 5 West). The reason behind this assumption is the similarity between the ESI distribution among 4 West and 5 West ED admit patients. Our data analyses show that 30% of ED patients that require a bed from 4 West are ESI 2 patients, and 69% of them are ESI 3 patients, while these proportions for 5 West patients are 28% and 70%, respectively. Since patients with similar severity are subject to similar levels of adverse events, we assume that the cost associated with the risk of adverse events are the same for 4 West and 5 West patients admitted through the ED.

Table C.9: ROAE Cost (per Hour) Cases

Cases	θ
Case 1	1
Case 2	5
Case 3	10

Table C.10: Penalty Cost Parameters

Cases	Type1	Type2
Case 1	1	0.5
Case 2	5	2.5
Case 3	10	5

C.7 An Extended Simulation Model

In the simulation model that is described in Section 4.6, we use the data that is obtained from our partner hospital. Due to limitations in data, we cannot have the exact patient flow that is described in Figure 4.3, and instead model the patient flow as described in Figure 4.6. In this section, we generate a simulation model that does not use the exact hospital data we have collected but models the patient flow that is

described in Figure 4.3 with additional characteristics that is obtained from the data analyses. Additionally, we model 5 p.m. discharge rounds to include a well-known concept used in some hospitals.

In the extended model, we model 8 IWs and 8 patient classes, and model the patient flow as it is described in Figure 4.3. Since we do not have exact data on the primary-secondary unit pairs, we randomly group the IWs as 2W-3W, 3E-7E, 4W-5W (which is the case for CP and CHF patients), and 4E-7W. Notice that IW 2 West is the ICU which is not an eligible unit for accepting patients from other IWs or assigning its primary patient to other IWs. Therefore, we drop 2W-3E pair from our analyses to fairly compare alternative policies with LEWC-p, and continue the simulation model with 6 IWs (3 primary-secondary pairs). As it is depicted in Figure 4.3, patients can only be assigned to their primary or secondary IW and there is no patient flow between different pairs. We include 5 p.m. discharge rounds by assuming that the patients who have discharge times earlier during the same day have to wait until 5 p.m. to leave the hospital.

We compare the performance of the proposed LEWC-p policy with alternative flow policies such as dedicated service policy, $c\mu$, and overflow trigger times. The results of the simulation model are shown in Table C.11, where we report the results for each policy as the proportion of that under LEWC-p (i.e., performance measure of the policy divided by the performance measure of LEWC-p) based on various cost combinations. LEWC-p policy reflects the trade-off between ROAE and quality of care: it is not as conservative as the dedicated policy in secondary IW assignments, and yet not as aggressive as $c\mu$ in making use of such assignments. When the cost associated to ROAE is relatively high compared to the cost of overflows, we observe that LEWC-p performs closer to $c\mu$. Another policy that we simulate is the overflow trigger times. Under this policy, patients can be overflowed to a secondary IW only

Table C.11: Performance Measures as a Proportion of Those Under LEWC-p

Cost	Performance Measure	Dedicated	$c\mu$	Ovrflw. Trig.(2hr)	Ovrflw. Trig.(4hr)
L. θ-H. p	Avg. no. of patients boarded	1.03	0.62	0.80	0.99
	Overflow prop.	0.00	2.07	1.71	1.02
	Avg. boarding time	1.17	0.59	0.86	0.99
	2-hour boarding rate	1.10	0.76	0.91	1.01
M.θ-M.p	Avg. no. of patients boarded	2.65	0.78	1.20	1.35
	Overflow prop.	0.00	1.81	0.97	0.73
	Avg. boarding time	1.81	0.79	1.14	1.26
	2-hour boarding rate	1.29	0.60	1.02	1.13
H. θ-L. p	Avg. no. of patients boarded	5.14	0.98	2.08	2.63
	Overflow prop.	0.00	1.03	0.82	0.68
	Avg. boarding time	3.21	0.99	1.23	1.39
	2-hour boarding rate	2.75	1.01	1.17	1.28

when they wait longer than a predetermined value. This policy prevents excessive boarding times, but is insensitive to system state since it is a static policy. The nature of our system requires dynamic decisions such as LEWC-p instead of static ones.