Stochastic and Deterministic Methods for Patient Flow Optimization in Care Service Networks

by

Jonathan Eugene Helm

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Industrial and Operations Engineering)
in The University of Michigan
2012

Doctoral Committee:

Associate Professor Mark Peter Van Oyen, Chair
Professor Mark Stephen Daskin
Emeritus Professor Walton M Hancock
Professor Wallace J Hopp
Assistant Professor Christopher Ryan Friese

This dissertation is dedicated to my wife, Sayuri Katrina Davis Helm, without whom none of this would have been possible. Her support during this process has been absolutely invaluable and will never be forgotten.

ACKNOWLEDGEMENTS

This dissertation is dedicated to my wife, Sayuri Katrina Davis Helm, without whom none of this would have been possible. Her support during this process has been absolutely invaluable. I would also like to thank my parents, Thomas E. Helm and Virginia M. Helm, as well as my brothers Amul D. Tevar and Jivan R. Deglise-Favre-Hawkinson for all their support. Thanks to my daughter, Lillian Yuki Helm, for being my best friend and for listening to my talks. The key player in my academic success was my advisor, Mark P. Van Oyen, whose tireless efforts and valued counsel have truly shaped me both personally and professionally. I would also like to thank two other colleagues and friends without whom this work would never have happened; Walton M. Hancock, whose pioneering work in the field and valuable mentorship have driven this new line of research, and Olof H. Minto whose guidance and support brought me to this path and sustained me while I traveled it. In that vein, I must also mention two other important mentors and contributors to my success, Mariel S. Lavieri and Thomas R. Rohleder. And to all my good friends at Michigan who made my time here enjoyable. I would like to thank Shervin AhmadBeygi, Hoda Parvin, and Fang Dong for being colleagues and friends from start to finish. My neighbors Brendan (the deal getter) See, Gregory (extreme sportsman) King, and Marcial (raise the roof) Lapp-Lappman. Thank you to Julian Pan, a great friend and colleague. Thanks to Coach Hoke for making my final year a Michigan football success. And finally, thanks to everyone at Michigan for all the memories.

TABLE OF CONTENTS

LIST OF FIGURES

Figure

LIST OF TABLES

Table

CHAPTER I

Introduction

The classical *Hospital Admission Scheduling and Control* (HASC) problem identified the late 1970s addresses one of the major systemic failures in hospital care delivery, *census variability*, through better management of inpatient admissions. Solution methodologies to reduce this variability are now known as *census smoothing*. This work solves the both the scheduling and the control portion of the HASC problem to optimality. The patient flow modeling and optimization approaches are further extended to fast-track priority patients through networks of specialist services. The research is validated and the impact is demonstrated through collaborations with multiple hospitals across three continents.

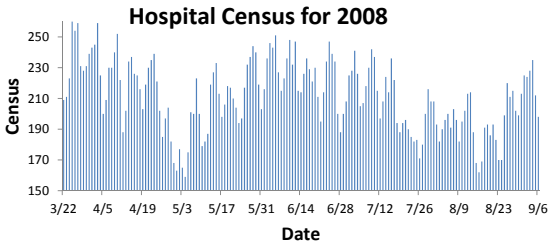
1.1 From Practice to Theory: A Scientific Approach to the Patient Flow Optimization.

This work was developed over four years of collaborative research with hospitals and healthcare providers around the world. We have worked with both large and medium sized hospitals and teaching and non-teaching hospitals in 4 different countries. The causes and consequences of census variability detailed below, along with the classic workload patterns that lead to systemic hospital congestion were observed in every case. From this research, we conclude that the problem we address

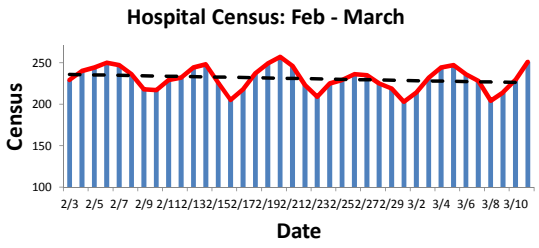
is a global one which, despite the difference among hospitals and healthcare systems, occurs with remarkable consistency. Our partner hospitals have agreed that the approaches developed herin merit development and implementation as a path toward sustainable care delivery that has high quality, excellent access, and low cost.

1.2 Consequences of Census Variability.

Hospital census variability is problematic throughout the world and impacts cost, access, quality and safety in healthcare delivery. Studies show that census variability leads to Emergency Department (ED) overcrowding, radiology backlogs, nurse burnout, and overcrowding in the Intensive Care Unit (ICU) and Post Acute Care Unit (PACU). This results in compromised quality of care, emergency patient diversions, excessive inpatient Length of Stay (LOS), and significant excess cost (see [54, 83, 14, 36, 69, 77, 79]). Figure 1.1(a) is a census time series from a partner hospital that illustrates typical census variability. Furthermore, most hospitals also exhibit a pattern of a mid-week census spike followed by a sharp drop in census (see 1.1(b)). This weekly census "hump" contributes to hospital overcrowding despite a modest average census (the dotted line in Figure 1.1(b)).



(a) Census variation over time

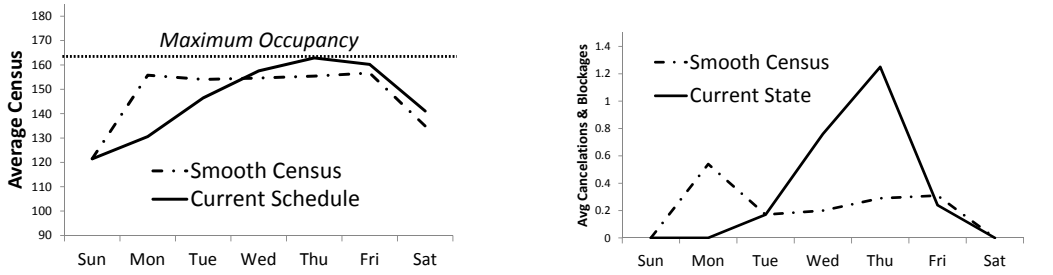


(b) Weekly occupancy pattern

Figure 1.1: Census variability in hospitals.

Census variability in combination with the mid-week hump contributes significantly to one of the major systemic problems in hospitals: "bed block." When bed

block occurs, emergency patients are forced to remain in the emergency department or in the hallway until a bed becomes available. This contributes to emergency department overcrowding (see [18, 47, 20]). When emergency departments become overcrowded, patient wait times increase dramatically along with the rate of accidents and mortalities (see [83]). Based on data from a partner hospital, we found that the cyclic census hump (see Figure 1.2(a)) contributes significantly to avoidable mid-week patient blockages and cancelations (see Figure 1.2(b)). The dotted line in these figures demonstrates the potential benefit of smoothing census across the week based on a high fidelity simulation incorporating the mechanisms developed here. The two census plots exhibit the same average daily census; however, the smoothed census benefits from significantly reduced blockages.



(a) Current vs smoothed avg. census

(b) Current vs smoothed blockages & cancelations

Figure 1.2: Controlling census variability in hospitals.

Quality of care also suffers from lack of census smoothing. Census variability leads to highly variable workloads for the nursing staff, labs, radiology, pharmacy and others. Overloaded nursing staff is linked to mortality, nurse burnout and job dissatisfaction (see [2]). While float nursing pools and other "chase" staffing strategies can be used, quality of care suffers and staff dissatisfaction increases (see [8, 53]). Additionally variable workloads create large backlogs in radiology and ancillary services that delay diagnosis, treatment and patient discharge (see [72]). Census variability

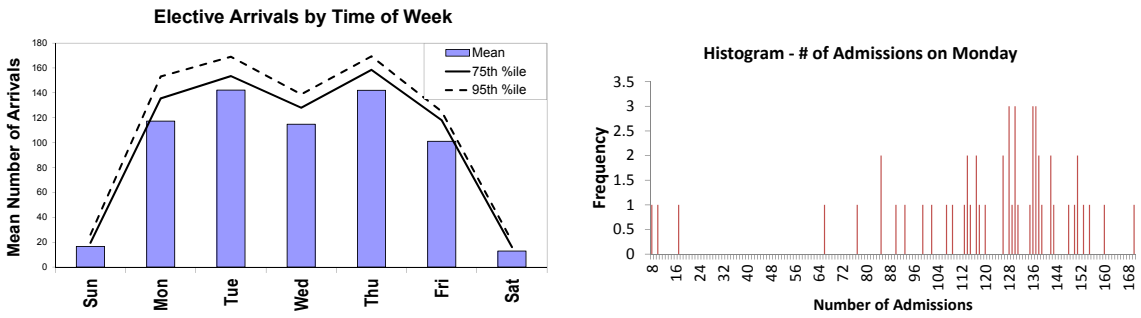
has also been linked directly to increased LOS, worsening patient disposition (see [20]), and even increased mortality rate (see [83, 79]).

To cope with high levels of congestion and overcrowding, hospitals have developed complex routing policies to take advantage of the flexibility of most bed wards: i.e., diverting overflow surgical patients into medicine ward beds. Patient safety, however, is not well served by placing patients "off-unit" (see [3]). Another consequence of off-unit placement is that elective surgical admissions can contribute to ED overcrowding, as surgical patients often overflow into medicine wards, blocking ED patients, whose primary destination is the medicine ward (see [45]).

1.3 Causes of Census Variability.

It is well known that both a weekly pattern in elective admissions (see Figure 1.3(a)) and the week-to-week variation in number of elective admissions on a given day (see Figure 1.3(b)) significantly contribute to both the weekly census hump and the week-to-week variation in census causing hospital congestion and patient blockages (see [4]). While these figures represent one hospital, our work with hospitals on three continents reveals the same pattern in every case and confirms what other researchers have found (see [33, 36]). The weekly census "hump" is generated by a weekly elective admission pattern that is heavily front-loaded and tapers off later in the week (see Figure 1.3(a)). The variability around the daily mean census can be attributed to the significant variation in the number of elective admissions scheduled on a given day from week to week. Figure 1.3(b) illustrates that the number of Monday elective admissions ranges widely from 8 to 170.

Table 1.1 demonstrates the magnitude of variability in the number of elective admissions by day of week (DOW). It may be surprising to note that elective admis-



(a) Weekly elective admission pattern

(b) Elective adm. histogram for Mondays (52 wks)

Figure 1.3: Variability in elective admissions over the course of one year.

sions actually exhibit *higher* coefficient of variation (i.e. standard deviation divided by mean) on many days than emergency admissions. Because elective admissions are more variable than emergency arrivals, despite the fact that they are controllable, we emphasize elective admission stabilization in our census smoothing approach.

Category	DOW	Std. Dev		Mean		CV	
		Emergency	Elective	Emergency	Elective	Emergency	Elective
Hospital	Sun	8.44	5.51	48.23	16.57	0.18	0.33
Hospital	Mon	13.23	32.98	64.79	117.32	0.20	0.28
Hospital	Tue	11.64	17.98	62.17	142.26	0.19	0.13
Hospital	Wed	10.59	23.88	57.53	114.79	0.18	0.21
Hospital	Thu	13.89	28.93	58.02	142.04	0.24	0.20
Hospital	Fri	10.96	20.49	64.79	101.09	0.17	0.20
Hospital	Sat	8.94	4.76	52.69	12.83	0.17	0.37

Table 1.1: Variation in numbers of total elective and emergency admissions by day of week (DOW).

The high level of workload variability, both seasonality and week to week variability, demonstrated above for one partner hospital is rife throughout the healthcare system, whether one considers hospitals or networks of outpatient specialist services. Because much of the variability is caused by the healthcare providers themselves, there is the possibility for significant, impactful change through improved patient scheduling and control.

1.4 A Path Forward: The Optimization of Patient Flow

This research seeks to *stabilize workloads* in healthcare networks by optimizing *patient admissions*. The three chapters discuss different mechanisms for managing admissions in different healthcare situations. Chapter II explores queueing network optimization methods for designing workload stabilizing schedules for hospitals – the scheduling portion of HASC. Chapter III addresses the control portion of the HASC by proposing a heuristic policy based on insights from a Markov Decision Process for dynamically managing admissions in hospitals. Chapter IV extends the queueing network optimization approach from Chapter II to a network of outpatient specialist services, developing a two stage optimization to address the needs of priority patients in healthcare systems. Finally, Chapter V reviews the important contributions of the work from both the theoretical and the application perspective.

CHAPTER II

Design and Optimization Methods for Elective Hospital Admissions

This chapter focuses on the scheduling portion of the HASC problem. That is, the approach seeks to optimize the scheduling of elective admission to smooth workloads across the network of services delivered by a hospital. This chapter develops new analytical models of controlled hospital census that can, for the first time, be incorporated into a Mixed Integer Programming model to optimally solve the *scheduling* portion of the HASC. This new solution method stabilizes elective admissions and coordinates admissions with other hospital subsystems to reduce system congestion. We formulate a new Poisson-arrival-location model (PALM) based on an innovative stochastic location process that we developed and call the Patient Temporal Resource Requirements (PATTERN) model. The PALM approach is then extended to the class of deterministic controlled-arrival-location models (d-CALM). This work provides the theoretical foundations for an efficient admissions management system as well as a practical decision support methodology to stabilize hospital census.

2.1 Smoothing Hospital Census

Hospitals with high throughput (achieved through better resource usage) can provide better access to their community at a lower cost. Thus the key to efficient

hospital management is high throughput with limited blockages. To smooth the hospital census the hospital must address both (1) the census mid-week "hump" and (2) weekly variability in admissions. The key to smoothing the hospital census lies in modeling the downstream time-phased patient resource requirements to inform hospital admission and scheduling decisions. In particular it is important to consider the ward/bed requirements for any mix of admitted patients over the course of their hospital stay. This translates into developing a model-based forecast of ward census levels over time for any particular mix and volume of patient admissions. We develop a computationally tractable model by statistically characterizing patient pathways through wards based on historical data.

The importance of census levels and census variability to admission decision making has been studied in several contexts. Connors uses stochastic patient flow models to link admissions decisions with hospital census [12]. Harrison uses simulation to show that census variability in combination with high census levels increases the risk of hospital overcrowding [36]. Jun argues that effective patient flow management can benefit the hospital through high patient throughput, low patient wait times, short LOS, and low clinic overtime [51] .

To effectively solve the *Hospital Admission Scheduling and Control (HASC) Problem*, models must incorporate control/scheduling decisions into census forecast models. Early work in this area began in the late 1970s with [32, 33, 26]. These early approaches took a comprehensive simulation modeling approach to the entire patient care pathways through the network of wards that comprise the hospital. Schedule improvement relied on a simulation-based heuristic approach to modeling the impact of admissions on census levels. Using simulation, the landmark work [33] designed and implemented an inpatient admissions scheduling and control system to achieve

high average census subject to constraints on the number of cancelations and emergency patient blockages. [21, 22, 11, 1, 4] have all studied the impact of elective admissions on census levels in various wards, optimizing schedules with Mixed Integer Programming (MIP) models. Recently, [35, 40] used simulation frameworks to improve scheduling decisions for better hospital resource usage.

[41] presented a Markov Decision Process (MDP) approach that focuses on the *control* side of the HASC problem to dynamically manage an inpatient call-in queue and elective surgery cancelation. It also showed via simulation that it can be effective to manage the *scheduling* side of the HASC problem. Given the high impact elective scheduling has on system performance, this chapter makes a contribution by (1) developing *analytical* census modeling methods, rather than simulation-based methods and (2) embedding them in a non-heuristic optimization to solve the *scheduling* side of the HASC problem and to yield significant managerial insight.

Past work has either been simulation based, or has not considered the full HASC system dynamics. For example, the MIP papers focus on a single ward or isolated feedforward subset of hospital resources. The scope of our work includes modeling the entire hospital, full patient care trajectories, and census levels by ward; moreover it includes the far more realistic generalized network dynamics of the hospital wards and the use of flexible wards to serve patients off-unit. In short, we are able to solve the *scheduling* side of the complete HASC problem exactly using non-heuristic optimization methods. To our knowledge, this has not been done before and thus our work represents a significant advance in basic science and understanding of the important HASC problem. To better capture the hospital dynamics, we model the hospital as a general network of interacting wards/units, incorporating the two primary types of interaction between wards that were not previously considered: (1)

transfers between different wards within the hospital as a result of a change in the patient's condition and (2) the use of off-unit capacity when a patient's primary ward is full. Consider a patient who arrives for surgery and is placed in an ICU bed for recovery. When the patient's condition improves, they transfer to a surgery bed for the remainder of their stay. Alternately, consider a surgical patient who leaves the operating room and must be placed "off-unit" (e.g. in a medicine ward), because all surgical beds are occupied.

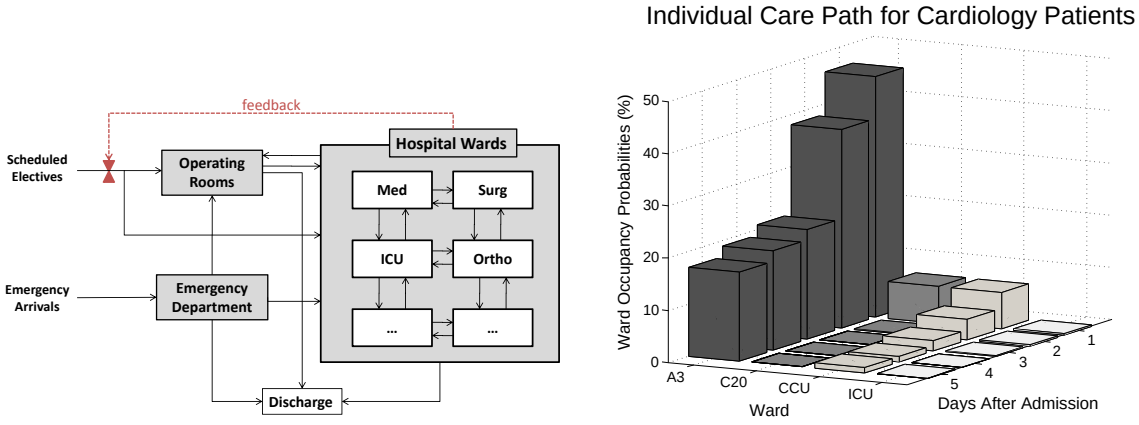
Ignoring the off-unit and inter-ward transfer mechanisms omits critical dynamics of hospital system functioning. In one of our partner hospitals 56% of patients transfer wards (after being admitted to an inpatient ward) at least once during their hospital stay and among patients who transfer, the average is 1.6 transfers per visit. Considering only the first ward, or a feed-forward subset of wards, ignores a significant load that patients place on other hospital resources. Off-unit interactions must be considered because, while placing patients off-unit is feasible, it is not desirable for patients or hospitals. Nursing skills differ from ward to ward, so the mismatch presented by off-unit patients detracts from quality of care (including safety) as well as nurse satisfaction (see [8]). Research shows that patients placed off their preferred ward experience more bottlenecks to discharge, increasing length of stay (see [3]). Additionally, the percent of off-unit patients is often quite significant; even in one of the better managed hospitals we worked with around 17% of patients were located off-unit.

In all previous research on HASC, the models that have considered total system flow lack a non-heuristic optimization component. On the other hand, the existing optimization models fail to consider complete patient trajectories through a general network of hospital wards as well as off-unit interactions. A primary contribution

of this chapter is in linking models that optimize *system-level* objectives to stochastic models of patient flow using *complete* patient trajectories through a *network* of hospital wards and the modeling of *ward interaction* mechanisms.

2.2 Characterization of the Stochastic Census Process

Figure 2.1(a) illustrates our methodological approach. We model the hospital as a network of interacting wards. The primary resource modeled is the hospital beds, differentiated by ward. The model uses the detailed temporal resource requirements via a data-driven network patient flow model to inform elective admission decisions while accounting for the resource requirements of the emergency patients. By optimizing elective admissions, accounting for the ED and ward beds, we show that it is possible to determine the volume and mix of elective patients to generate a consistent, stable workload and minimize blockages and cancelations while maintaining or increasing patient throughput.



(a) Conceptual patient flow model

(b) Cardiology patient flow path

Figure 2.1: Variability in elective admissions over the course of one year.

We begin by developing a stochastic model of patient flow through the network of hospital wards. As a basic building block, we first characterize the patient care pathway for each patient type. While the model is adaptable to many different

definitions of patient type (e.g. diagnosis/Diagnosis Related Group, preferred ward upon admission, etc.), we consider patient type to be the patient's admitting service (e.g. cardiac, gastrointestinal, neurology, etc.). We generate a probabilistic flow model of the resources (beds) used by a patient of a given type over their entire stay in the hospital. Figure 2.1(b) shows the expected load (which is also a probability) a cardiology patient places on hospital wards over the course of their treatment, where the y dimension indicates days after admission.

After we characterize the elective census level in each ward for a given elective schedule using these care pathways, we combine the elective census process with the emergency census process to characterize the total census levels in each ward for a given elective admission schedule by day of week. Finally, this census process is linked to elective admission decision variables in an optimization model to determine the optimal mix and volume of patients over time subject to system performance constraints. In our case study we consider constraints on bed block, for example.

2.2.1 System Design and Assumptions

We begin by clarifying the modeling assumptions and the perspective of our approach. In our admission plan design, we allow for a specific planning horizon (e.g. a week) and, when the system goes beyond the planning horizon, the admission plan is repeated exactly as before. This means that we are working with a cyclostationary system in equilibrium. From a practical standpoint a weekly cycle is appealing to our hospital partners. Since doctors usually have fixed clinic times, OR time, research time, etc. each week, a repeating elective admission schedule fits well within the practical constraints of the hospital environment.

Another design element of our system is that the number of elective patients to be scheduled is deterministic. Elective admissions are usually scheduled with

sufficient lead times that allow the control of the number of patients of a particular type (e.g. knee replacement) to be performed on a specific day of the week. As noted in Chapter I, a primary cause of census variability is the extreme variability in the number of elective admissions from week to week. Figure 1.3 underscores the extent to which elective admissions deviate from the mean volumes from one week to the next. Overcoming schedule instability requires organizational change to enforce operating discipline on the admissions decision makers. This requires leadership, but the approach is welcomed once it is realized that the changes benefit patients, physicians, nurses and administrators in important ways. This chapter focuses on a centralized control approach, but other approaches could consider pricing schemes on operating room time that vary by day of week. If such an approach were preferable this chapter will still provide the goal that such an incentive scheme should drive toward. While deterministic elective arrivals will only be approximated in practice, the necessity of stabilizing the volume of elective admissions from week to week has been highlighted in several papers in the literature (see [4, 32, 21]).

A further assumption is that the care path of each patient is independent of other hospital inpatients. This is a mild assumption considering our approach to census modeling. Specifically, we initially develop models of demand for hospital services without regard for hospital capacity. In essence, we assume infinite capacity when developing our census models and therefore one patient does not block or otherwise alter the care path of another patient. Capacity requirements are then superimposed on the raw demand model for calculating patient service metrics.

Finally, we anticipate our elective admissions system being used most often at a daily granularity because this gives more flexibility and locates decision making power with admission planners, increasing likelihood of acceptance of our management

system. The examples that follow present the elective admission system in terms of days, though the modeling approach is significantly more general.

2.2.2 Development of the PATTERN Stochastic Location Process Model

To understand the effects of scheduling decisions and emergency arrivals on census levels across the network of hospital wards, consider as a building block the resource (bed) requirements of a single patient over the course of their treatment, which we call Patient Temporal Resource Requirements (PATTERN). To describe the flow of patients through hospital wards, we develop a stochastic location process model in the spirit of [65, 66]. Some applications and extensions of this approach include [59, 62]. Let the state space be $\mathcal{S} = \{1, \ldots, M \cdot n, \Delta^*, \Delta_*\}$, where state i indicates that the patient is currently in ward i, state Δ^* represents the state where the patient has left the hospital (i.e. discharged) and Δ_* represents the state where the patient has not yet arrived at the hospital. Patients move through the state space according to the $\mathcal{S}$-valued stochastic location process $\{L_s(t) : s \in \mathbb{R}\}$, where s is the arrival time and $t > s$ is the time of interest. For notational convenience we let $\mathcal{S} = \mathcal{S}^0 \cup \{\Delta^*, \Delta_*\}$, so that $\mathcal{S}^0$ represents the locations within the hospital. Thus $L_s(t)$ denotes the location of a patient at time t given that the patient was admitted at time s.

Remark II.1. The fact that $L_s(t)$ can depend on s enables the modeling of a key hospital feature that the length of stay and care path can depend on the time of admission

As an example of the remark, discharge policies often differ on weekends. Also, high mid-week congestion forces off-unit placement that can increase length of stay. To characterize the stochastic location process, let Σ_s be the set of right-continuous functions with left limits that first enter the hospital $\mathcal{S}^0$ at time s. Thus, Σ_s represents the set of all possible sample paths of the stochastic location process $L_s(t)$. An

element $\sigma_s \in \Sigma_s$ is a (deterministic) mapping $\sigma_s : \mathbb{R} \to \mathcal{S}$ such that $\sigma_s(t)$ represents the location of the patient at time t. Implicitly, $\sigma_s \in \Sigma_s$ has the property that $\sigma_s(s) \in \mathcal{S}^0$ and $\sigma_s(t) = \Delta_*$ for all $t < s$. Figure 2.2 represents three different sample path functions. The solid line represents path $\sigma_{s_1}(t)$, a sample path of the process $L_{s_1}(t)$, the dashed line represents the path $\sigma_{s_2}(t)$, a sample path of the process $L_{s_2}(t)$, and the dotted line represents the path $\sigma_{s_3}(t)$, a sample path of the process $L_{s_3}(t)$. Path $\sigma_{s_2}(t)$, for example, represents a patient who arrives at time s_2 at ward 1, transfers to ward 2 for a brief stay, and then returns to ward 1 before being discharged slightly before time t. Note that a location function $\sigma \in \Sigma_s$ is a right-continuous step function that takes values in $\mathcal{S}^0$ over a continuous interval $[s, T_s)$ for some finite T_s and that $\sigma(t) = \Delta_*$ for $t < s$ and $\sigma(t) = \Delta^*$ for $t \geq T_s$.

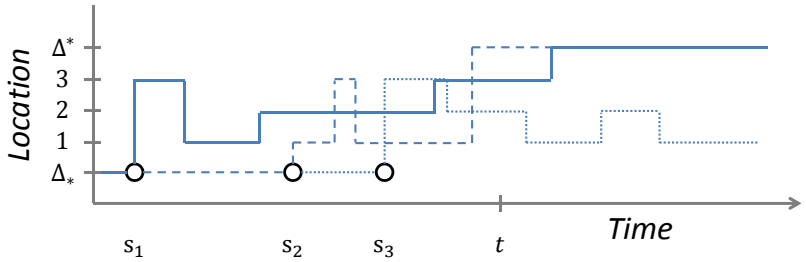


Figure 2.2: Patient sample care paths.

We let the entire function space Σ be the collection of all Σ_s. For any subset $\Gamma \subseteq \Sigma$, there is an associated probability measure, $P_s(\Gamma)$, that represents the probability associated with a set of location functions. The s subscript in the probability measure denotes the time of the patient's arrival so $P_s(\cdot)$ characterizes the dynamics of the stochastic location process, $L_s(t)$. Note that $P_s(\Sigma_s) = 1$ and $P_s(\Sigma_t) = 0$ for $t \neq s$.

We will demonstrate how this measure is used to find the probability that a patient is in ward u at time t, given that they arrived at the hospital at time s. We first define a set of location functions and then a measure on that set that will yield the desired location probability. Consider the set of location functions that describe

whether the patient is in ward u at some time t. The measure of this set is the probability of being in ward u at time t. To this end, define the set of sample paths in Σ,

$$\Gamma_{t,u} = \big\{\sigma_s \in \Sigma_s \ : \ s < t \text{ and } \sigma_s(t) = u\big\}, \tag{2.1}$$

to capture the set of all location functions that place a patient in ward u at time t. Of course to be in the hospital at time t, the patient must have arrived before time t, which is implied by the conditions on the elements of the set. Moreover, we require that the patient not remain in the hospital forever (consistent with [65]). As mentioned, the specific measure of this set is defined by the dynamics of the location stochastic process $L_s(t)$. One common location stochastic process in queueing is a semi-Markov process. The solution to such processes for general distributions and general transition functions, however, is often intractable, requiring approximations for solutions. Rather than rely on further approximation methods, we use the approach detailed in [38]. Each patient type will have their own stochastic location model (in our case study we consider 20 different patient types). Thus for patient type k

$$P_{s,k}\big(\Gamma_{t,u}\big) \ = p_{s,k,u}(t-s) \tag{2.2}$$

$$P_{s,k}\big(\Gamma_{t,\Delta^*}\big) = p_{s,k,\Delta^*}(t-s) = 1 - \sum_{j=1}^{W} p_{s,k,j}(t-s), \tag{2.3}$$

where $p_{s,k,u}(t)$ is the probability that a patient of type k who arrives at time s is in ward u, t time units after admission to the hospital. These probabilities can be obtained from historical data (see [38]). An example of these probabilities for discrete time points for cardiology patients is shown in Table 2.1. Entry (j,t) of the matrix represents the probability that the patient will require a bed in ward j, t time periods (e.g. days) after admission. In this table, ward A3 is a cardiology ward, CCU is the

critical care unit, and ICU is the intensive care unit, and C20 is a ward for short stay patients (usually less than 2 days). The discrete version of the probabilities can be described by a PATTERN Matrix. Note that the probabilities need not sum to 1 because implicitly the remaining probability mass not assigned to a ward is the probability of the patient not requiring a hospital ward bed at time t.

	Time (Days)					
Ward	1	2	3	4	5	$\cdots$
$A3$	45.92%	37.54%	21.22%	19.16%	17.03%	$\cdots$
$C20$	6.17%	0.14%	0.00%	0.00%	0.00%	$\cdots$
CCU	7.10%	3.48%	2.34 %	1.14%	0.92%	$\cdots$
ICU	0.14%	0.14%	0.14%	0.07%	0.07%	$\cdots$

Table 2.1: Patient Temporal Resource Requirements (PATTERN) matrix for a cardiology patient.

2.2.3 The Emergency Census Process

There are several ways to characterize the emergency census process. [21] characterizes the process for a single ward in terms of means and variances. For our purposes, we prefer a complete characterization of the emergency census process over a network of wards. [48] developed an interpolation method based on historical flow data to characterize census quantiles, but this approach is incapable of incorporating changes in the underlying dynamics of emergency patient flow. Further, a complete approach characterizes a probability distribution on the number of emergency patients by ward.

To approximate this probability distribution, we develop an approximation of the emergency patient flow using the following scenario. For the time being, ignore the elective patients and consider only the flow of emergency patients through an otherwise empty hospital. Since we are first modeling just the demand for services, we consider an open network of infinite server queues. Because of the infinite server approach, we can model each ward as a cluster of infinite server queues with one queue for each emergency patient type, each with its own non-homogeneous arrival

rate, its own service distribution, and its own routing probabilities. In queueing, the network for each patient type is denoted by [65] as $\left(M_t/G_t/\infty\right)^N/G_t$.

It has been shown that the non-stationary Poisson Process is a good model for emergency patient arrivals (see [36]), and we allow for general, non-stationary service time distributions as well as non-stationary routing probabilities that may also depend on the length of stay in a given ward. The feature of interest in this model is, of course, the number of patients demanding a bed in each ward as well as the total number of patients demanding a bed in the hospital. This requires obtaining distributions on subsets of the vector state space. If there are M wards and n emergency patient types, then we are considering a network of $M \cdot n$ queues. Let $\mathfrak{W}_i(t) = Q_i^1(t) + Q_i^2(t) + \cdots + Q_i^n(t)$ represent the amount of emergency patient demand for ward i at time t, where $Q_i^j(t)$ is the demand of type j patients for ward i. Analogously, $\mathbf{Q}(t) = \sum_{i=1}^{M} \sum_{j=1}^{n} Q_i^j(t)$ represents the total emergency patient load placed on the hospital at time t. These two quantities are sufficient for our later analysis in which we overlay capacity constraints on the demand model to calculate blockages and off-unit census.

2.2.4 PATTERN Poisson-arrival-location Model (PALM) of Emergency Census

To specify the PATTERN PALM model for the emergency census process we rely on the Poisson random measure approach proposed by [65]. In this model patients arrive according to a non-homogeneous Poisson process and then flow through the hospital according to our PATTERN stochastic location process $L_s(t)$ described in Section 2.2.2. Details of the standard Poisson random measure and its extension to a doubly stochastic Poisson process can be found in the Appendix 2.5. We define our PATTERN PALM random measure in terms of the composition of the standard Poisson random measure $\mathbf{M}$, and the PATTERN intensity measure, μ. In this section

we refer to $\mathcal{M}$, the set of measures μ on $\mathbb{R}^+$, and $\mathcal{N} = \{\mu \in \mathcal{M} : \mu(t) \in \mathbb{Z}^+\}$, the set of measures $\mu \in \mathcal{M}$ that yield integer values. Additionally, $\mathcal{B}(\cdot)$ represents the Borel sigma algebra. Full definitions and descriptions of these concepts are detailed in the Online Appendix.

Here we provide an alternative definition of the intensity of the Poisson random measure for the PALM model to enable the extension of the arrival-location modeling approach to deterministic controlled arrivals in Section 2.2.5. We begin by specifying the location random measure in a similar manner to the definition of the standard Poisson random measure. That is, we define a mapping from the probability space $(\Sigma, \mathcal{B}, \mathbb{P})$ into the measure space $(\mathcal{M}, \mathcal{B}(\mathcal{M}))$. Let the probability that a patient of type k arriving at time s is in ward j at time t be defined as

$$\mathbb{P}_k(\sigma_s \in \Sigma_s : \sigma_s(t) = j) \equiv P_{s,k}(\sigma \in \Sigma : \sigma(t) = j) = \begin{cases} 0 & \text{if } t < s \\ p_{s,k,j}(t-s) & \text{if } t \geq s. \end{cases} \tag{2.4}$$

The random location measure of the stochastic process, $L_s(t)$, for the subset of wards $\mathcal{J} \subseteq \mathcal{S}$ is then specified by

$$\Lambda_{k,s}(t, \mathcal{J}, \sigma) = \begin{cases} 1 & \text{if } \sigma(t) \in \mathcal{J},\ \sigma \in \Sigma_s \\ 0 & \text{otherwise.} \end{cases} \tag{2.5}$$

Now we can specify the random intensity measure, $\mathbf{N}_k$, for patients of type k by combining the non-homogeneous Poisson arrival process having nonnegative deterministic integrable external-arrival-rate function $\alpha_k(t) \in \mathbb{R}^+$ with the location random measure from Equations 2.4 and 2.5. The arrival rate function, $\alpha_k(t)$, drives the number of type k emergency patient arrivals. Once a patient has arrived at time s, the patient then flows through the wards according to the PATTERN stochastic location process, $L_{s,k}(t)$ with dynamics driven by the probability measure $P_{s,k}(\cdot)$.

The rate of flow into the group of wards $\mathcal{J}$ at time t of type k arrivals entering the hospital at time s follows by multiplying the non-stationary arrival rate by the stochastic location random measure: $\alpha_k(s)\Lambda_{k,s}(t,\mathcal{J})$. Random measure $\mathbf{N}_k$ gives the random arrival-transition intensity to wards $\mathcal{J}$ at time t of type k arrivals entering the hospital over the interval $(a,b]$.

$$\mathbf{N}_k((a,b],t,\mathcal{J}) = \int_a^b \alpha_k(s)\Lambda_{k,s}(t,\mathcal{J})ds. \tag{2.6}$$

Intuitively, this can be related to Poisson splitting of a non-homogenous Poisson process. The external arrival intensity drives the number of arrivals over a period of time; however, each arrival will be in a particular location depending on the location stochastic process $L_s(t)$. Therefore the external arrival intensity is distributed across the wards (or "departed") over time. Because $\mathbf{N}_k$ is a random intensity, $\mathbf{M} \circ \mathbf{N}_k$ is a random measure that represents a doubly stochastic Poisson process. For our purposes, the mean arrival-transition intensity in combination with the Poisson random measure is sufficiently precise and computationally efficient. The mean (deterministic) transition intensity measure, μ_k, and it's properties are defined in the following lemma (proved in the Online Appendix):

Lemma II.2. *For the deterministic average arrival intensity measure, μ_k, the following hold*

(i) $\mu_k((a,b],t,\mathcal{J}) \equiv \mathbf{E}[\mathbf{N}_k((a,b],t,\mathcal{J})] = \int_a^b \alpha_k(s) \sum_{j\in\mathcal{J}} p_{s,k,j}(t-s)ds$,

(ii) μ_k *is a measure on* $\mathbb{R}\times\mathbb{R}\times\mathcal{S}$.

We combine the mean arrival-transition intensity measure with the standard Poisson random measure to obtain the PATTERN Poisson random measure for type k patients, $\mathbf{M}_k = \mathbf{M}\circ\mu_k$. Let the $B_i = (a_i,b_i]\times t_i \times \mathcal{J}_i$ represent the event that patients

arrive at the hospital on interval $[a_i, b_i)$ and those patients are in the set of wards $\mathcal{J}_i \subseteq \mathcal{S}^0$ at some time in the future, t_i. Then $\mathbf{M}_k$ can be shown to have a product form Poisson distribution with rate γ_i:

$$P\big(\mathbf{M}_k(B_1) = m_1, \mathbf{M}_k(B_2) = m_2), \ldots, \mathbf{M}_k(B_n) = m_n\big) = \prod_{i=1}^{n} \frac{e^{-\gamma_i}\gamma_i^{m_i}}{m_i!} \tag{2.7}$$

$$\gamma_i \equiv \mathbf{E}[\mathbf{M}_k(B_i)] = \mu_k((a_i, b_i], t_i, \mathcal{J}_i) = \int_{a_i}^{b_i} \alpha_k(s) \sum_{j \in \mathcal{J}_i} p_{s,k,j}(t_i - s)ds. \tag{2.8}$$

Eq. 2.8 follows from Lemma II.2. We now quantify the distribution on the number of emergency patients in the cyclostationary system (mentioned in Section 2.2.1) in steady state, where the arrival pattern is repeated on a weekly basis. If we let τ_k be the maximum length of stay for a patient of type k then we have the following result, which is proved in the Appendix 2.5.

Theorem II.3. *The number of emergency patients in each ward, denoted by $Q_1(t)$, $Q_2(t), \ldots, Q_n(t)$, are independent Poisson random variables for each time $t \in \mathbb{R}^+$ with finite mean given by*

$$m_j(t) = \sum_{k=1}^{n} \int_{t-\tau_k}^{t} \alpha_k(s) p_{s,k,j}(t - s)ds. \tag{2.9}$$

2.2.5 PATTERN Deterministic controlled-arrival-location Model (d-CALM) of Elective Census

To build a model of the elective census over time we begin in Section 2.2.2 by characterizing the care paths for each type of patient that is admitted to the hospital. In Section 2.2.5 we use these care paths in combination with the elective admission schedule to model the census process for a given admission plan. In Section 2.2.5 we calculate the first and second moments of the census process. The approach

for the elective census model represents an extension of the PALM methodology to processes with deterministic arrivals, or what we call the deterministic controlled-arrival-location model (d-CALM). In this approach, arrivals occur at specific times (possibly in batches), rather than according to a Poisson distribution. Once a patient of type k has arrived at time s, they flow through the hospital according to their PATTERN stochastic location process $L_{s,k}$ as in Section 2.2.2, with subscript k denoting patient type k. This makes explicit our condition that each patient type follows their own location process determined by the characteristics of their medical condition.

Defining the Elective Census Stochastic Process.

Combining the PATTERN model for individual patients with the elective admission schedule, Θ, it is possible to model the total elective census in the hospital over time. One approach is to formulate a point process as in Section 2.2.4. For patients of type k, let $\big((t_{k,1}, \Theta_{k,t_{k,1}}), (t_{k,2}, \Theta_{k,t_{k,2}}), \dots\big)$ represent the sequence of deterministic arrivals with $t_{k,i}$ being the time of arrival of the i^{th} batch of patients of type k and $\Theta_{k,t_{k,i}}$ being the number of type k patients scheduled for time $t_{k,i}$. Let $\Omega = \Sigma^{\infty}$ so that $\omega_k = \{\sigma_{k,(t_{k,1}),1}, \sigma_{k,(t_{k,1}),2}, \dots, \sigma_{k,(t_{k,1}),\Theta_{k,t_{k,1}}}, \sigma_{k,(t_{k,2}),1}, \sigma_{k,(t_{k,2}),2}, \dots, \sigma_{k,(t_{k,2}),\Theta_{k,t_{k,2}}}, \dots\} \in \Omega$ represents the set of location functions for the scheduled arrivals. We define the d-CALM probability measure for patients of type k being in ward j as

$$\mathbb{P}_k(\{\omega \in \Sigma^{\infty} : \sigma_{k,(t_{k,n}),n}(t) = j\}) = \begin{cases} 0 & \text{if } t < t_{k,n} \\ p_{(t_{k,n}),k,j}(t - t_{k,n}) & \text{if } t \geq t_{k,n} \end{cases}, \tag{2.10}$$

where $p_{(t_{k,n}),k,j}(t - t_{k,n})$ is as before in Eq. 2.3. Then we can define the d-CALM point process, for a realization vector ω as

$$\mathbf{N}_{k,j,\Theta}(t, \omega) = \begin{cases} \sum_{s \in \{t_{k,i} : t_{k,i} < t\}} \sum_{n=1}^{\Theta_{k,s}} \Lambda_{k,s}(t, j, \sigma_{k,s,n}) & \text{if } t_1 < t \\ 0 & \text{if } t_1 > t, \end{cases} \tag{2.11}$$

where $\Lambda_{k,s}(\cdot)$ is the patient type k random measure defined for the stochastic location process in Eq. 2.4 and 2.5 of Section 2.2.4. It can be seen that this point process can be written instead as the process

$$\mathbf{N}_{k,u,\Theta}(t) = \sum_{s\in\{t_{k,i}:t_{k,i}<t\}} \sum_{j=1}^{\Theta_{k,s}} \mathbb{1}\{L^{j}_{s,k}(t) = u\}, \tag{2.12}$$

where $\mathbf{N}_{k,u,\Theta}(t)$ is the number of elective patients of type k in ward u at time t under schedule Θ. We will work with this more convenient form to analyze the d-CALM process, which is equivalent to the point process defined by Eq. 2.10 and 2.11. The ward level census can be calculated by summing over patient types and the hospital census can be calculated by summing over all the wards. Now we also include the system design assumption of a cyclically repeating elective admission schedule. We present the case where the hospital is concerned with daily measures of admissions and census as an example. We analyze this case (though the approach will work more generally) as it is particularly useful for managerial insight and operational planning.

Using Eq. 2.12 the census in ward u, C_{u,d_1}, can be calculated on a given day d_1 of the planning horizon. After presenting the computations, we present an illustrative example. First define $\mathfrak{W}$ as the set of hospital wards and $\mathfrak{D}$ as the set of patient types (e.g. diagnoses). If we take the planning horizon to be one week, $N = 7$ for example, the total hospital census on a given day d_1 can be calculated for a finite horizon of length t weeks (C^{t}_{u,d_1} from Eq. 4.2) or an infinite horizon (C^{∞}_{u,d_1} from Eq. 4.3),

$$C^{t}_{u,d_1} = \sum_{d_2=1}^{7} \sum_{k\in\mathfrak{D}} \sum_{j=0}^{\Theta_{k,d_2}} \sum_{n=0}^{t} \mathbb{1}\left\{L^{j,n}_{d_2+7n,k}(d_1+7t) = u\right\} \tag{2.13}$$

$$C^{\infty}_{u,d_1} = \lim_{t\to\infty} C^{t}_{u,d_1}, \tag{2.14}$$

where $L^{j,n}_{s,k}(\cdot)$ represents the $(j,n)^{th}$ i.i.d instance of the location process $L_{s,k}(\cdot)$, one process for each admitted patient, j, on a given week, n, and $\mathbb{1}\{\cdot\}$ is the indicator function. In Eq. 4.2 and 4.3, the first sum refers to the day of the week that the patient was admitted. The second sum refers to the diagnosis of the patient and the third sum represents the number of patients of that diagnosis that are to be scheduled on day d_2 of the planning horizon. The final sum over n iterates through weeks (or through cycles of the planning horizon). To obtain total hospital census it is sufficient to sum Eq. 4.2 and 4.3 over the set of hospital wards.

These equations are best understood through a simple example. Consider a plan that admits 2 cardiology patients (patient type = CAR) every Monday. What is the load that this plan places on the cardiology ward, ward c, on Tuesdays? Let $\mathbb{1}\{L^{j,n}_{s,CAR}(t) = c\}$ represent whether the (j,n) indexed cardiology patient is in the cardiology ward c on day t given they were admitted on day s. On the first Monday, the system admits two cardiology patients (call them patient (1,0) and (2,0)). This leads to a census for Tuesday of the first week ($n = 0$) of $\mathbb{1}\{L^{1,0}_{1,CAR}(2) = c\} + \mathbb{1}\{L^{2,0}_{1,CAR}(2) = c\}$. Note that $\mathbb{1}\{L^{1,0}_{1,CAR}(2) = c\}$ and $\mathbb{1}\{L^{2,0}_{1,CAR}(2) = c\}$ are i.i.d. because they represent two different patients. In the second week we admit two more cardiology patients (call them patient (1,1) and (2,1)). Since the first two cardiology patients admitted previously may still be in the hospital (and thus on day 8 of their length of stay) the census for the Tuesday of the second week ($n = 1$) is $\mathbb{1}\{L^{1,0}_{1,CAR}(9) = c\} + \mathbb{1}\{L^{2,0}_{1,CAR}(9) = c\} + \mathbb{1}\{L^{1,1}_{8,CAR}(9) = c\} + \mathbb{1}\{L^{2,1}_{8,CAR}(9) = c\}$. If we let the system run for t weeks, then the census on the Tuesday of week t is given by $\sum_{n=0}^{t} \mathbb{1}\{L^{1,n}_{7n+1,CAR}(7t+2) = c\} + \mathbb{1}\{L^{2,n}_{7n+1,CAR}(7t+2) = c\}$.

This shows how we construct the census profile for Eq. 4.2 and 4.3. We are primarily interested in the steady state behavior of the system, and thus rely mostly

on the infinite horizon formulation of Eq. 4.3 in the analysis that follows.

To this end we define:

Definition II.4. $\hat{X}_d(\Theta) \equiv L_d^{\infty}$ is the steady state elective census vector for all hospital wards, for day d of the planning horizon under admission plan Θ.

Moments of the PATTERN d-CALM Elective Census Process.

The above formulation of the elective census process allows us to calculate the first and second moments of the process analytically, which facilitates the use of the census process in an optimization formulation. Take the planning horizon to be one week, $N = 7$ for example, the mean of census for ward u of the hospital a given day d_1 can be calculated from Eq. 4.2 and 4.3 by the monotone convergence theorem as

$$\begin{aligned}\mu_{d_1,u}(\Theta) &= \mathbf{E}\Big[\sum_{d_2=1}^{7}\sum_{i\in\mathfrak{D}}\sum_{j=0}^{\Theta_{i,d_2}}\lim_{t\to\infty}\sum_{n=0}^{t}\mathbb{1}\big\{L^{j,n}_{d_2+7n,i}(d_1+7t))=u\big\}\Big] \\ &= \sum_{d_2=1}^{7}\sum_{i\in\mathfrak{D}}\Theta_{i,d_2}\cdot\sum_{n=0}^{\infty}p_{d_2+7n,i,u}(d_1-d_2+7(t-n)). \qquad (2.15)\end{aligned}$$

The equality follows from the fact that $\mathbb{1}\{\mathbf{X} = x_k\}$ follows a Bernoulli distribution and thus $\mathbf{E}\big[\mathbb{1}\{\mathbf{X} = x_k\}\big] = p_k$. The mean census level in the hospital can be calculated by summing Eq. 4.4 over the set of all wards, $\mathfrak{W}$, as $\sum_{u\in\mathfrak{W}}\mu_{d_1,u}(\Theta)$.

We compute the variance of the elective census process for two types of variance: (1) the variance in ward census and (2) the variance in total hospital census. The variance and covariance of (1) and (2) is given, with proof in the Appendix 2.5, by

Lemma II.5. *The covariance of the cyclostationary ward and total census processes is*

(i) $Cov\big(\mathbb{1}\{L^{j_1,n_1}_{s_1,k_1}(t)=u_1\},\mathbb{1}\{L^{j_2,n_2}_{s_2,k_2}(t)=u_2\}\big)=0$ *for all* $(j_1,n_1,k_1,s_1)\neq(j_2,n_2,k_2,s_2)$,

(ii) $Cov\big(\mathbb{1}\{L^{j,n}_{s,k}(t) = u_1\}, \mathbb{1}\{L^{j,n}_{s,k}(t) = u_2\}\big) = -p_{s,k,u_1}(t-s)p_{s,k,u_2}(t-s)$ *for* $u_1 \neq u_2$.

Theorem II.6. *Letting* $d(n) = d_1 - d_2 + 7(t-n)$*, the variance of the cyclostationary ward and total census processes is*

(i) $\sigma^2_{d_1,u}(\Theta) = \sum_{d_2=1}^{7} \sum_{k\in\mathfrak{D}} \Theta_{k,d_2} \sum_{n=0}^{\infty} p_{d_2+7n,k,u}(d(n))(1 - p_{d_2+7n,k,u}(d(n)))$,

(ii) $\sigma^2_{d_1}(\Theta) = \sum_{u\in\mathfrak{W}} \sigma^2_{d_1,u}(\Theta)$
$$- \sum_{d_2=1}^{7} \sum_{k\in\mathfrak{D}} \Theta_{k,d_2} \sum_{n=0}^{\infty} \sum_{u_1\neq u_2} p_{d_2+7n,k,u_1}(d(n))p_{d_2+7n,k,u_2}(d(n)).$$

We see that $\sigma^2_{d_1}(\Theta)$ can be written as a linear function of the admission plan (decision) Θ, and thus included in an integer programming framework for determining optimal schedules. The variance and covariance terms can be calculated offline and enter the optimization as data. Since, from Theorem II.6, the variance is still linear in terms of our decision variables $\Theta_{i,d}$, the model remains solvable by standard MIP solution approaches.

2.2.6 Validating the Hospital Census Model

The total census process (for wards and for the hospital) is approximated by the sum of the elective census process and the emergency census process (Sections 2.2.5 and 2.2.3). In this section we show that our approximation of the census process closely matches the actual census process using a year of historical data from a partner hospital. The partner hospital in this validation had 23 wards and 20 different patient types (one for each major admitting service). The arrival rates to each ward are given in Appendix 2.5.3, Fig. 2.8.

Figure 2.3 and Table 2.2 show the mean census levels by day of week for the entire hospital for both the approximation and for the historical census levels.

The deviations are seen to be relatively small and have little effect on accurately approximating system metrics such as cancelations and blockages as shown in Section

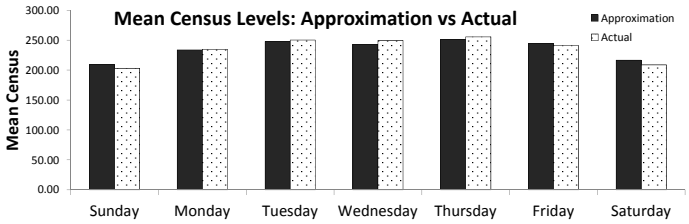


Figure 2.3: Comparison of the mean census approximation vs historical mean census

	Sunday	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday
Approximation	209	234	248	243	251	244	217
Actual	203	235	250	250	256	241	209
% Diff	2.8%	-0.6%	-1.0%	-2.8%	-1.8%	1.3%	3.7%

Table 2.2: Comparison of the mean census approximation vs. historical mean census

2.3.3. Having such an accurate analytical approximation is extremely important to enable the optimization of the elective admission schedule. Prior efforts at solving this problem for the entire hospital have relied on simulations to achieve accurate census approximations, making optimization difficult (see [41, 32, 35]). The approximation developed in this and previous sections can be easily incorporated into an integer programming optimization model as demonstrated in Section 2.3.

2.3 Optimization of Elective Admissions Mix and Volume

In Section 2.2 we developed a modeling and analysis method for quantifying census under a given admission plan. In this section we design an integer programming model to determine the optimal schedule given a set of metrics. For our metrics, we trade off two conflicting objectives in hospital management: (1) the desire to admit as many elective patients as possible (alternatively to keep bed utilization high) and (2) the desire to limit the number of blockages and off-unit census for both emergency and elective patients. The stochastic process from Section 2.2 characterizes the raw demand for beds, so to quantify the blockages we need to superimpose the hospital capacity on this model. Section 2.3.1 presents a method for calculating various

blockage metrics in a manner that can be incorporated into an integer programming formulation. Section 2.3.2 presents two different formulations for the elective admissions mix and volume that could be useful to hospitals. Section 2.3.3 validates the method by comparing the forecasted census from the optimization model with a high fidelity simulation of hospital operations.

2.3.1 Computation of System Effectiveness Metrics

In hospitals, there are two significant types of bed block: (1) ward-level bed block and (2) hospital-level bed block. Type (1) prevents a patient from entering a particular ward, forcing the patient into an off-unit ward. Type (2) prevents any access to the hospital (e.g. cancelation, diversion). Limiting both types of bed block is critical to operating a high performing hospital as detailed in Chapter I and Section 2.1.

To calculate these metrics, one must consider the stochastic dynamics of hospital and ward census. To this end, it is possible to obtain the distribution on the number of elective patients in each ward, but *not* in the entire hospital, which is a feature of controlled arrivals that makes d-CALM more complex than the PALM. The ward u census on day d_1, C^t_{u,d_1}, is a sum of independent Bernoulli random variables, which is known to be Poisson-Binomial (see [10]). If $Z_1, \ldots, Z_{N(\Theta,t)}$ represent the different indicators in Eq. 4.2, each having probability of success p_i, then $C^t_{u,d_1} = Z_1 + \cdots + Z_{N(\Theta,t)}$ is a Poisson-Binomial random variable with distribution:

$$P(C^t_{u,d_1} = n) = \left\{ \prod_{i=1}^{N(\Theta,t)} (1 - p_i) \right\} \sum_{i_1 < \ldots < i_n} w_{i_1} \cdots w_{i_n}, \tag{2.16}$$

where $w_i = p_i/(1 - p_i)$ and the sum is over all non-double counting choices of n of $N(\Theta, t)$ (see [10]).

Unfortunately, not only is this distribution difficult to compute, it also introduces

significant non-linearities in the decision variables because $N(\Theta, t)$ depends on the admission schedule Θ. This prohibits the use of MIP or other tractable optimization methods. Additionally, the Poisson-Binomial only models sums of independent indicators so, as will be seen, the total hospital census does not follow the Poisson-Binomial distribution.

Several approximations we investigated also failed to provide solutions that could be incorporated into our MIP. To begin, recall that the contribution of each patient to ward (hospital) census is modeled using a PATTERN stochastic location process combined with a deterministic arrival rate function. The census is then calculated by summing indicators of the PATTERN process. In particular, the census on a given day, d_1, can be rewritten as:

$$C^t_{d_1,u} = \sum_{d_2=1}^{7} \sum_{k \in \mathfrak{D}} \sum_{n=0}^{t} \sum_{j=0}^{\Theta_{k,d_2}} \mathbb{1}\left\{ L^{j,n}_{d_2+7n,k}(d_1 + 7t)) = u \right\}. \tag{2.17}$$

Consider that for fixed k, d_2, u, and n, $\sum_{j=0}^{\Theta_{k,d_2}} \mathbb{1}\left\{ L^{j,n}_{d_2+7n,k}(d_1 + 7t)) = u \right\}$ is a sum of Θ_{k,d_2} i.i.d. Bernoulli random variables, which is Binomial$(\Theta_{k,d_2}, p_{d_2+7n,k,u}(d_1 - d_2 + 7(t-n)))$. Worse still, if we want to consider the joint distribution of the different wards of the hospital then the sum becomes the sum of i.i.d. categorical random variables, which is distributed as Multinomial$(\Theta_{k,d_2}, \mathbf{p}_{d_2+7n,k}(d_1 - d_2 + 7(t-n)))$ where $\mathbf{p}_{d_2+7n,k}(d_1 - d_2 + 7(t-n)) = \left[p_{d_2+7n,k,1}(d_1 - d_2 + 7(t-n)), \ldots, p_{d_2+7n,k,W}(d_1 - d_2 + 7(t-n)) \right]$ is the probability vector of the categorical random variables defining whether or not the patient will reside in the various hospital wards at a given time (see [50]). So the census random variable $C^t_{d_1}$, defined in Eq. 2.17 is the sum of many different Binomial (or Multinomial if we consider the joint process) random variables, which is not tractable. It is clear that, if we want to incorporate the p.m.f. of the elective census in the same manner as the emergency census, this will lead to

an intractable optimization model because the decision variable Θ involves factorials.

Another approach considers the fact that, under certain conditions, the sum of independent Binomial random variables can be approximated by a Poisson or Normal distribution among other methods. Unfortunately, the Normal probabilities and quantiles are defined with respect to the standard deviation, which is non-linear in our decision variables as shown in Section 2.2.5. The Poisson approach also suffers from the fact that its probabilities are non-linear in Θ.

Due to the complications involved in working with distributions or variances of the elective census in an optimization framework, we propose the following approximation to obtain estimates of the expected blockages and off-unit census illustrated in Figure 2.4. The approach begins by calculating the mean elective census by day of week, indicated by the solid bar in Figure 2.4 (which we justify below). The number of beds remaining (i.e. capacity minus mean elective demand) is referred to as the *reserved capacity* (for emergency patients). Starting with the mean census as a baseline, we add the emergency patients, indicated by the individual bars on top of the solid bar, and account for the probability of each level of emergency patients using the PATTERN PALM model of Section 2.2.3. Blockages are tallied when the number of emergency patients plus the mean number of elective patients exceeds the hospital capacity, $B = \sum_{i \in \mathfrak{W}} B_i$ where B_i is the capacity of ward i. Thus the blockages are calculated with respect to the emergency patient distribution obtained from the PATTERN PALM model, while accounting for the mean number of electives in the hospital.

This approximation preserves the linearity required for efficient solutions to a mixed integer program as will be shown in Section 2.3.2. Since the controlled cyclostationary system will optimize a deterministic number of elective admissions by day

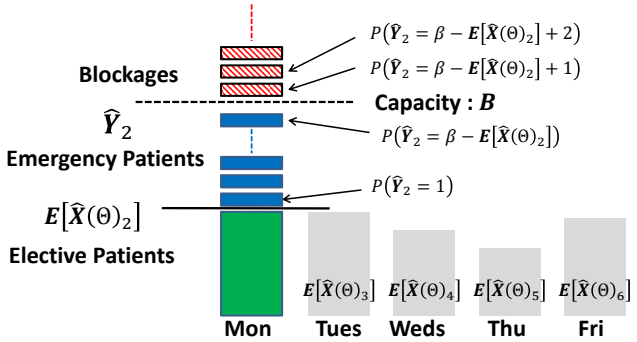


Figure 2.4: Illustration of expected blockage constraint for the entire hospital.

of week, the majority of the census variability will now come from the emergency patients, which we capture with the emergency census distribution. In Section 2.3.3, we demonstrate the accuracy of this approximation on data from a partner hospital. The high level of accuracy suggested by the testing described in Section 2.3.3 indicates that more complicated approaches incorporating variance in elective census may not be necessary in light of the need for tractability/solution speed.

The off-unit census levels can be calculated in a similar way to the total hospital blockages by considering the census in each ward and comparing it to the ward capacity. Any amount of demand by a patient in the hospital (i.e. one not blocked from entering) in excess of the ward's capacity must necessarily be considered off-unit.

2.3.2 Mixed Integer Programming Formulation

We begin this section with notation and then proceed to a formulation of the elective admission mix and volume optimization model. The planning horizon we consider is days $1, \ldots, 7$ to correspond to a weekly schedule.

Sets

$\mathfrak{D}$ set of all patient diagnosis types

$\mathfrak{W}$ set of hospital wards

Parameters

B_i ward i capacity in terms of beds

β limit on the average number of blockages per week

α_u percent of total cancelations that are attributed to ward u

$\hat{\beta}_u$ limit on the average number of off-unit patients allowed for ward u

$p_{d_1}^{k,u}$ probability that an elective patient of type k is in ward u d_1 days after admission

$\hat{p}_{n,d}^{u}$ probability there are n emergency patients in ward u on day d from the PATTERN PALM model

$\tilde{p}_{n,d}$ probability there are n emergency patients in the hospital on day d from the PATTERN PALM model

$\theta_{k,d}$ current elective admission volume of type k patients on day d.

$\hat{\theta}_{k,d}$ maximum number of elective admissions of type k allowed on day d.

$\mathbf{R}$ reward vector where R_k is the reward for admitting patient of type k

Decision Variables

$\Theta_{k,d}$ number of type $k \in \mathfrak{D}$ patients scheduled on day d

$\delta_{n,d}$ number of blockages if there are n emergency patients in the hospital on day d

$\hat{\delta}_{n,d}^{u}$ number of ward u off-unit patients on day d if there are n emergency patients in ward u

It is important to note here that the probabilities $p_{d_1}^{k,u}$, $\hat{p}_{n,d}^{u}$, and $\tilde{p}_{n,d}$ are all calculated offline per the analysis in Sections 2.2.3 and 2.2.5 and then become data inputs to the two mixed integer programs that follow.

Maximum Elective Admissions Formulation.

First we present a formulation that maximizes the number of elective admissions subject to constraints on bed blockage. $\mathbb{1}$ denotes a column vector of all ones. Merely for the sake of generality we include the "reward" row vector $\mathbf{R}$ providing a relative value for a patient of type k served. In practice we let $\mathbf{R}$ be a row of all 1's (every patient type has the same value) and then manipulate the constraints if management's goal is to increase the volume of one particular service.

(2.18)

$$\max_{\Theta,\delta,\hat{\delta}} \quad \mathbf{R}\cdot\Theta\cdot\mathbb{1}$$

$s.t.$

(2.19)

$$\delta_{j,d_1} \geq j - \sum_{u\in\mathfrak{W}}\left(B_u - \sum_{d_2=1}^{7}\sum_{k\in\mathfrak{D}}\Theta_{k,d_2}\cdot\sum_{n=0}^{\infty}p^{k,u}_{(7n+d_1-d_2)}\right) \qquad d_1 = 1,\ldots,7, j = 1,2,\ldots$$

(2.20)

$$\sum_{d=0}^{7}\sum_{n=0}^{\infty}\tilde{p}_{n,d}\delta_{n,d} \leq \beta$$

(2.21)

$$\delta_{n+1,d} \geq \delta_{n,d} \qquad d = 1,\ldots,7, n = 1,2,\ldots$$

(2.22)

$$\hat{\delta}^u_{j,d_1} \geq j + \sum_{d_2=1}^{7}\sum_{k\in\mathfrak{D}}\Theta_{k,d_2}\cdot\sum_{n=0}^{\infty}p^{k,u}_{(7n+d_1-d_2)} - B_u - \alpha_u\sum_{d=0}^{7}\sum_{n=0}^{\infty}\delta_{n,d}\cdot\tilde{p}_{n,d}$$

$$\forall u\in\mathfrak{W}, d_1 = 1,\ldots,7, j = 1,2,\ldots$$

(2.23)

$$\sum_{n=0}^{\infty} \hat{p}^u_{n,d} \hat{\delta}^u_{n,d} \leq \hat{\beta}_u \qquad \forall u \in \mathfrak{W}, d = 1, \ldots, 7$$

(2.24)

$$\hat{\delta}^u_{n+1,d} \geq \hat{\delta}^u_{n,d} \qquad d = 1, \ldots, 7, n = 1, 2, \ldots$$

(2.25)

$$\sum_{d=1}^{7} \Theta_{k,d} \geq \sum_{d=1}^{7} \theta_{k,d} \qquad \forall k \in \mathfrak{D}$$

(2.26)

$$\Theta_{k,d} \leq \hat{\theta}_{k,d} \qquad \forall k \in \mathfrak{D}, d = 1, \ldots, 7$$

(2.27)

$$\Theta_{k,d}, \delta_{k,d}, \hat{\delta}^u_{k,d} \in \mathbb{Z}^+$$

The objective function, Eq. 2.18, maximizes the weighted throughput of elective patients. Constraint 2.19 manages the integer helper decision variable $\delta_{n,d}$ that enables the program to measure expected blockages. This constraint consists of several terms that will be explained individually. On the right hand side of the equation we have $\sum_{u \in \mathfrak{W}} B_u$, representing the capacity of the hospital. Subtracted from the total hospital capacity is the expected elective bedload on day d taken from Eq. 4.4. Call the quantity resulting from the subtraction $E[RC]$. Thus the right hand side of the equation, $j - E[RC]$, represents the amount by which the number of emergency patients in the hospital exceeds the expected empty beds remaining after elective admissions are accounted for. If this quantity is non-negative, it represents the number of blockages if j emergency patients are in the hospital. If the RHS is positive, then δ_{j,d_1} is forced to be at least as great as the number of blockages

that would occur in the scenario where there are j emergency patients and the mean number of elective patients in the hospital. If the RHS is negative, then the model will trigger no blockages and δ_{j,d_1} can be set to 0.

Constraint 2.20 is the constraint that approximates the expected number of weekly blockages for a given schedule Θ and limits it to at most β. The method to obtain this approximation is detailed in Section 2.3.1. Constraint 2.21 is a cut that was added to the model to increase the speed of the CPLEX implementation of a branch and bound algorithm. Because of the large number of δ decision variables, this cut greatly reduces the number of combinations that must be considered by branch and bound. Without this constraint, a model with three wards and three patient types failed to solve in under 24 hours; while solving in under 30 seconds with the constraint.

Constraints 2.22 - 2.24 serve the same function for measuring and limiting expected off-unit census as Constraints 2.19 - 2.21 do for expected blockages. The one piece that is different is the subtraction of the term $\alpha_u \sum_{d=0}^{7} \sum_{n=0}^{\infty} \delta_{n,d} \cdot \tilde{p_{n,d}}$ relative to Eq. 2.19. This term accounts for the fact that, if patients are canceled or otherwise not admitted to the hospital, they will not contribute to off-unit census in the wards they would have been admitted to. The parameter α_u refers to the historical trend and/or hospital protocols for what types of patients get canceled when a cancelation decision must be made.

The final two constraints, Equations 2.25 and 2.26, represent the reality that the model should not change the elective admission schedule in ways incommensurate with historical hospital practice. Specifically Eq. 2.25 ensures that, under the improved schedule, each service can at least maintain historical volumes. This means that the model will not take away business from any specialty or practice.

Eq. 2.26 ensures that the model respects capacity constraints beyond hospital beds. These could be limits on the amount of Operating Room time, or the fact that most hospitals choose to admit few or no elective patients on the weekends (e.g. $\Theta_{k,Sunday} \leq 0 \quad \forall k$)

Minimum Blockages Formulation.

Another useful formulation is to keep the weekly volume of elective admissions fixed and attempt to minimize the number of blockages. This model reshuffles the mix of elective admissions across the days of the week to eliminate unnecessary blockages caused by an unstable, unbalanced schedule. The main difference in this formulation is that the objective function becomes the expected number of blockages

$$\min_{\Theta,\delta,\hat{\delta}} \quad \sum_{d=0}^{7} \sum_{n=0}^{\infty} \tilde{p}_{n,d} \delta_{n,d}, \tag{2.28}$$

and the weekly volume is strictly equal to the current weekly volume, i.e. Eq. 2.25 becomes

$$\sum_{d=1}^{7} \Theta_{k,d} = \sum_{d=1}^{7} \theta_{k,d} \; \forall k \in \mathfrak{D}. \tag{2.29}$$

2.3.3 Validating the Hospital Census Optimization Model

As in Section 2.2.6, it is important to quantify the accuracy of the hospital census and blockage approximations for the optimal elective schedule. Because there is no historical record of hospital census and blockages for the optimal schedule, we compare the census approximation with a high-fidelity simulation model that has already been validated against historical hospital data (see [41, 42, 40]).

A year's worth of historical hospital data was used to calibrate both the optimization and simulation models for a core subset of nine hospital wards (out of 22 total), including medicine, surgical and ICU/CCU wards. This reduction limited the

significant amount of data analysis and cleaning without degrading the value of the study.

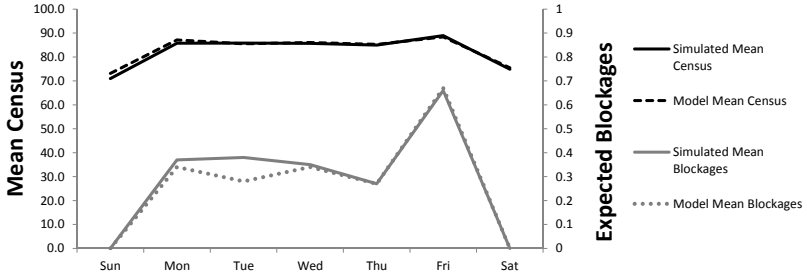


Figure 2.5: Simulation output vs stochastic model output for characteristic hospital measures.

	Sun	Mon	Tue	Wed	Thu	Fri	Sat	Tot
Sim Census	71.0	85.8	85.8	85.7	85.0	88.9	74.9	N/A
Approx Census	73.2	87.1	85.5	86.1	85.3	88.4	75.5	N/A
% Diff Census	*3.1%*	*1.6%*	*-0.4%*	*0.4%*	*0.4%*	*-0.6%*	*0.8%*	*N/A*
Sim Blockages	0.00	0.37	0.38	0.35	0.27	0.66	0.00	2.03
Approx Blockages	0.00	0.34	0.28	0.34	0.27	0.67	0.00	1.9
% Diff Blockages	*0.0%*	*-8.1%*	*-26%*	*-2.9%*	*0.0%*	*-1.5%*	*0.0%*	*6.4%*

Table 2.3: Simulation output vs stochastic model output for characteristic hospital measures

Figure 2.5 and Table 2.3 confirm that the stochastic census model is a good approximation of actual census levels and blockages. The small bias toward higher census levels can be explained by the manner in which the simulation treats cancelations and blockages. In the simulation, the demand from cancelations and blockages is considered lost (an approximation of reality), whereas the census approximation models the overall *demand* for beds without loss. Although blockages are calculated, the blocked patients are not removed from the demand calculations which yields the depression of census in the simulation versus the analytical model. The reality is likely somewhere in between, as some demand is lost and some is rescheduled. Regardless, the estimated values are very close; weekly blockages only differ by 6% in absolute value and the census differs on average by only 1%.

Because the stochastic census model is an accurate approximation, the detailed

(and therefore slow) simulation is *no longer needed* to express the tradeoffs between census and blockages to design effective admission schedules.

2.3.4 Case Study, Proof of Concept, and Managerial Insights

To demonstrate the effectiveness and potential uses of our approach to elective admissions scheduling, we validate our method using historical hospital data. The hospital is a medium sized, non-teaching hospital, and as in Section 2.3.3, we model the nine medicine, surgical, and ICU/CCU wards of the hospital and compare optimized schedules with the current schedule.

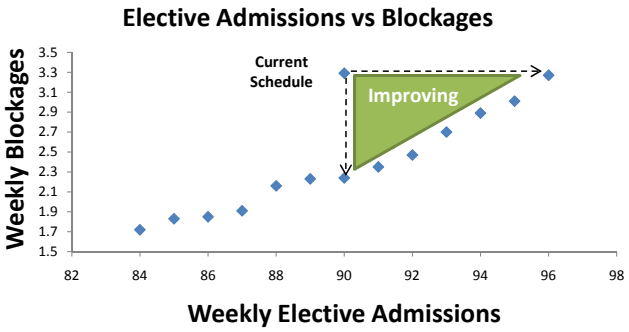
A year's worth of data is used to model daily census; therefore we consider only patients that stayed in the hospital for at least one night. In 2008, 14,827 patients stayed at least one night. Out of these overnight patients, 7,016 were emergency patients while the remaining 7,811 were scheduled patients. Patients transferred within the hospital 20,462 times, for an average of around 1.4 transfers per patient. This transfer ratio serves to underscore the importance of modeling the ward network effects in hospitals. The nine wards we model comprise about 60% of the total patient volume with similar characteristics to the total patient population.

One of the primary goals of this modeling approach is to address patient blockage, both elective cancelations and emergency patient bed block, without reducing the number of patients served. The wards modeled admitted 90 elective inpatients per week on average. The *minimum blockage formulation* was employed, constraining the weekly elective volume to equal 90 and also constraining the volumes on each admitting service to match the current level so that the mix remains constant. The optimization generated an optimal schedule matching these criteria, which we then simulated (for completeness) to compare with the current schedule. The result was an average *32% reduction in cancelations per week* as shown in "Minimum Blockages"

of Table 2.6(a).

Schedule	Elective Adm per Wk	Blockages per Wk
Current	90	3.29
Min Blockage	90	2.34
Max Adm	96	3.27

(a) Current vs optimized schedules



(b) Pareto curve – throughput vs blockage

Figure 2.6: Controlling census variability in hospitals.

Another goal a hospital might have is to increase the volume of patients served while maintaining the same level of service. This would lead to increased revenues while still delivering the same or enhanced access. To achieve this goal, the *maximum admissions formulation* is employed, constraining the blockages to be less than or equal to the current (3.29 per week) and maximizing the number of admissions. This included a constraint to ensure that each service is given at least as many elective admissions as in the current schedule (i.e. Constraint 2.25). The result, "Maximum Admissions" in Table 2.6(b), is an *additional 310 elective admissions per year* (6 per week) with slightly better access (3.27 blockages per week).

These prescriptive models are useful in exploring the boundaries of hospital efficiency, but hospitals may prefer a balance between volume and blockage. Our model can provide information, guidelines and a method for achieving the preferred balance. The Pareto curve in Figure 2.6 presents the tradeoff between elective admission volume and blockages. Notice that the current schedule is above the Pareto curve so it can be improved by increasing admissions, decreasing blockages, or both.

To generate this curve, we use the extreme points as boundaries and employ the

minimum blockage formulation by iterating the weekly number of elective admissions between 90 and 96 and determining the schedule with the fewest blockages at each admission level. This curve represents an important advance in decision support that enables hospital administrators to understand the key tradeoffs involved in scheduling their admissions and gives them the freedom to choose their desired operating point. Hospital personnel are likely to prefer an approach that provides them information and allows them to make strategic decisions rather than being prescribed a specific solution.

This Pareto curve also represents an advance in the basic science of admission scheduling. Past simulation-based approaches incorporating general network effects would likely struggle to produce an operating curve; the only other known attempt required 8 hours of computation per data point (see [42]). Our optimization models were able to generate the operating curve automatically in a matter of minutes, with each point taking about 30 seconds.

2.4 Conclusions and Future Work

We have developed new models for a longstanding unsolved problem in hospital operations. This methodology can efficiently generate optimal schedules to meet high-level hospital criteria while modeling the entire hospital as a coordinated system. The results have significant potential to inform hospital decision makers as to how to use admission scheduling as a tool to create a healthcare delivery system that is lest costly while providing better access, quality and service to patients.

Rather than mandating specific implementations of elective procedure scheduling, our approach provides decision support on case mix and volume by patient type by day of week. Thus, we mitigate barriers to adoption. Additionally, the discipline and predictability obtained by embracing this system of smoothed census will streamline

hospital procedures, stabilize the operating environment for hospital personnel, more efficiently utilize fixed hospital resources, and yield significant cost savings. For example, census variability reduction enables, among other things, cost savings in nurse staffing while better facilitating proper nurse to patient ratios.

The HASC problem has been approached in many ways; however, previous approaches have not been able to generate optimal schedules for the entire hospital, including ward network effects. The simulation approaches capture the critical general network effects, but they lack a clear schedule optimization method. The scheduling optimization models, on the other hand, have not included the general network effects, such as ward transfers and off-unit census, that are critical to accurately modeling the true census load on hospital wards. Our modeling approach has bridged this gap by accurately capturing the census and blockage dynamics analytically, eliminating the need for simulation and enabling the use of MIP methods. To do so we formulated a PATTERN PALM "arrival-location-model" to show that the emergency demand for beds by ward can be characterized as independent Poisson random variables. Secondly, we extended the PALM approach to a new deterministic controlled-arrival-location model (d-CALM) for elective admissions and analyzed its properties.

The proposed MIP models can identify schedules that reduce blockages and/or increase elective volumes, but also enable us to generate a Pareto operating curve that trades off blockages and admission volume. This curve represents an effective decision making tool for hospital administrators, as it enables flexibility and choice rather a prescribed fixed solution. This approach is likely to increase acceptance by administrators, enabling them to make important decisions based on deeper managerial insights.

Future work might include linking the ward census models to an operating room schedule and/or other critical hospital subsystems. Work that develops optimization methodology for the *control* portion of the HASC model in combination with the *scheduling* portion would solve the full HASC problem and likely add significant value to the field. Finally, the generality of the approach opens the possibility of application to the effective redesign of many other patient flow systems.

2.5 Appendix

2.5.1 Poisson Random Measure.

We begin this section with some general definitions for point processes that we will use in the construction of our emergency census process. These definitions and detailed analysis can be found in [24]. The Poisson random measure is defined by its *intensity measure*, so we start by defining the space in which the intensity measure lies as follows.

Definition II.7. Let $\mathcal{M}$ be the set of measures μ on $\mathbb{R}^+$ such that $\mu(0) = 0$ and $\mu(t) < \infty$ for all $t \in \mathbb{R}^+$.

It is important for the analysis that a metric can be defined in $\mathcal{M}$ that makes $\mathcal{M}$ separable and complete (see [24]). Let $\mathcal{B}(\mathcal{M})$ be the Borel σ-algebra on $\mathcal{M}$. To describe point processes, it is necessary to also define the following space, $\mathcal{N}$

Definition II.8. Let $\mathcal{N} = \{\mu \in \mathcal{M} : \mu(t) \in \mathbb{Z}^+ \text{ for all } t \in [0, \infty)\}$ be the set of measures $\mu \in \mathcal{M}$ that yield integer values.

The space $\mathcal{N}$ is important for the Poisson random measure we wish to define because the Poisson distribution takes only integer values. Now we can define a random measure as

Definition II.9. A *random measure* is a measurable mapping from the probability space $(\Omega, \mathcal{B}, \mathbb{P})$ into the measure space $(\mathcal{M}, \mathcal{B}(\mathcal{M}))$ or $(\mathcal{N}, \mathcal{B}(\mathcal{N}))$ as the situation requires.

We characterize our PATTERN PALM Poisson random measure by extension of the *standard Poisson random measure*, $\mathbf{M}$, which we now define as in [24] for clarity. Let $\Omega = \mathbb{R}^\infty$ and define an element $\omega \in \Omega$ by the process's i.i.d. interarrival times: $\omega = (s_0, s_1, s_2, \ldots)$. Let $\mathcal{B}$ be the natural σ-algebra on $\mathbb{R}^\infty$ and $\mathbb{P}$ be defined as

$$\mathbb{P}(\{\omega \in \mathbb{R}^\infty : s_k \leq x\}) = \begin{cases} 0 & \text{if } x < 0 \\ 1 - e^{-x} & \text{if } x \geq 0 \end{cases} . \tag{2.30}$$

That is, $\mathbb{P}$ assigns exponential probabilities with rate $\lambda = 1$ to each interarrival time, s_k. Now the standard Poisson random measure can be defined as

$$\mathbf{M}(t, \omega) = \begin{cases} k & \text{if } \sum_{j=0}^{k-1} s_j \leq t,\ \sum_{j=0}^{k} s_j > t \text{ and } \sum_{j=0}^{\infty} s_j = \infty \\ 0 & \text{if } s_0 > t \text{ or if } \sum_{j=0}^{\infty} s_j < \infty \end{cases} . \tag{2.31}$$

Let the distribution of the standard Poisson random measure be $\hat{\Pi}$, which is a probability measure on $(\mathcal{N}, \mathcal{B}(\mathcal{N}))$ such that for any set $A \in \mathcal{B}(\mathcal{N})$, $\hat{\Pi}\{A\} = \mathbb{P}\{\mathbf{M} \in A\}$. The standard Poisson random measure can be generalized by scaling time according to a deterministic intensity measure, μ. By letting $\mathbf{M} \circ \mu = \mathbf{M}(\mu) : \mathcal{N} \times \mathcal{M} \to \mathcal{N}$ be a Poisson product measure with distribution denoted Π_μ, we obtain a random measure model for a Poisson process with intensity μ. A special case is to let $\mu\big((0, t]\big) = \lambda t$, in which case we get the traditional Poisson process with rate λ. In our case, however, we are interested in allowing the intensity, which will be denoted by Λ, to be a realization of a random measure as well. In particular, we want the intensity to represent the non-stationary arrivals to the various wards of

the hospital over time under a total arrival process that is non-homogeneous Poisson, which yields a *doubly stochastic Poisson process* in the terminology of [24].

Let $\tilde{\mathbf{M}} = \mathbf{M} \circ \Lambda$ be a Poisson random measure on a product space $\mathcal{N} \times \mathcal{M}$, where the intensity random measure Λ has distribution which is denoted by Π. The standard Poisson random measure, $\mathbf{M}$, has distribution $\hat{\Pi}$, so the distribution of $\tilde{\mathbf{M}}$ is then specified for any set $B \in \mathcal{N}$ by

$$\begin{aligned} P_{\tilde{\Pi}}(B) &= \hat{\Pi} \times \Pi(\{(\nu, \mu) \in \mathcal{N} \times \mathcal{M} : \nu \circ \mu \in B\}) \\ &= \int_{\mathcal{M}} \hat{\Pi}(\{\nu \in \mathcal{N} : \nu \circ \mu \in B\}) \Pi(d\mu) \\ &= \int_{\mathcal{M}} \Pi_{\mu}(B) \Pi(d\mu), \ B \in \mathcal{B}(\mathcal{N}), \end{aligned} \tag{2.32}$$

where the first step follows from Fubini's theorem (see [19]) and the definition of $\mathbf{M} \circ \mu$. It is important to note that $\Pi_{\mu}(B)$ is a $\mathcal{B}(\mathcal{M})$-measurable function in μ, so the preceding integral is valid (see [24]).

2.5.2 Proofs

Lemma II.2

Proof. (i) First note that $\Gamma_{t,\mathcal{J}} = \bigcup_{j \in \mathcal{J}} \Gamma_{t,j}$, where $\Gamma_{t,j}$ is defined as in Eq. 2.1. Therefore

$$\begin{aligned} P_{s,k}(\Gamma_{t,\mathcal{J}}) = &\sum_{j \in \mathcal{J}} P_{s,k}(\Gamma_{t,j}) - \sum P_{s,k}(\bigcap \text{any two events}) + \\ &\sum P_{s,k}(\bigcap \text{any three events}) + \dots \\ &+ (-1)^{|\mathcal{J}|+1} P(\bigcap_{j \in \mathcal{J}} \Gamma_{t,j}) = \sum_{j \in \mathcal{J}} P_{s,k}(\Gamma_{t,j}) = \sum_{j \in \mathcal{J}} p_{s,k,j}(t-s)) \end{aligned} \tag{2.33}$$

The first equality follows from the fact that $P_{s,k}$ is a probability measure. The second equality follows from the reasonable assumption that a patient can only be in one ward at a time, and therefore the intersection of any two sets necessarily has

measure zero. Because $\alpha_k(\cdot)$ is assumed integrable and $\Lambda_{k,s}(\cdot,\cdot) \in [0,1]$, then for any $b \in \mathbb{R}^+ \bigcup \{\infty\}$:

$$\begin{aligned}\mu_k((a,b],t,\mathcal{J}) \equiv \mathbf{E}[\mathbf{N}_k((a,b],t,\mathcal{J})] &= \mathbf{E}\left[\int_a^b \alpha_k(s)\Lambda_{k,s}(t,\mathcal{J})\right] \\ &= \int_a^b \alpha_k(s)\mathbf{E}[\Lambda_k((s,t],\mathcal{J})] \\ &= \int_a^b \alpha_k(s)P_{s,k}(\Gamma_{t,\mathcal{J}}) \\ &= \int_a^b \alpha_k(s)\sum_{j\in\mathcal{J}} p_{s,k,j}(t-s)).\end{aligned}$$

The first equality follows by applying the definition of $\mathbf{N}_k((a,b],t,\mathcal{J})$. The second inequality follows from the dominated convergence theorem. The third equality follows by applying the definition of the random measure Λ_k from Eq. 2.4 and 2.5. The final equality follows from Eq. 2.33.

(ii) The location functions are easily seen to be right continuous, with limits from the left existing. If we endow the function space Σ with the Skorohod J1 topology (as in [65]) then Σ can be shown to be Polish, thereby removing measure theoretic complications. Because we are dealing with patients, our PATTERN process has only finitely many jumps and the total length of stay is finite. Additionally, the probability measure $P_{s,k}$ is a measure on Σ and is a measurable function of s, so the integral is valid.

Now we show that μ_k is a measure. First, it is clear that $\mu_k(\emptyset) = 0$ because integrating over the null set returns zero. Secondly, integrals over disjoint sets are countably additive resulting in the countable additivity of μ_k. That is, if we let $E_i = (a_i, b_i] \times t_i \times \mathcal{J}_i$, then

$$\mu_k(\lim_{n\to\infty}\bigcup_{i=1}^{n} E_i) = \sum_{i=1}^{\infty}\int_{a_i}^{b_i}\alpha_k(s)\sum_{j\in\mathcal{J}_i} p_{s,k,j}(t_i-s)ds = \sum_{i=1}^{\infty}\mu_k(E_i) \tag{2.34}$$

This directly results in countable additivity of the measure μ_k, therefore μ_k is a measure. □

Theorem II.3

Proof. From Appendix 2.5.1, we know that we can extend the standard Poisson random measure with an intensity measure. From Lemma II.2, we know that the deterministic intensity, μ_k, is in fact a measure. Therefore the product form result for each patient type then follows directly from the properties of the Poisson random measure and the fact that we are considering disjoint subsets of $\mathbb{R} \times \mathbb{R} \times \mathcal{S}$ (e.g. one for each pair of ward and patient type).

Finally, the number of patients in a given ward is just the sum over all patient types,

$$\mathbf{M}_1 + \mathbf{M}_2 + \ldots + \mathbf{M}_n, \tag{2.35}$$

which is Poisson with rate $\sum_{k=1}^{n} \gamma_k = \sum_{k=1}^{n} \int_{t-\tau_k}^{t} \alpha_k(s) p_{s,k,j}(t-s)$, because the sum of independent Poisson random variables is Poisson with a rate that is the sum of the individual rates.

□

Lemma II.5

Proof. (i) Each pair of indicators where $(j_1, n_1, k_1, s_1) \neq (j_2, n_2, k_2, s_2)$ represents two different patients. Therefore the two indicator random variables are independent, which follows from our assumption that the care paths of two different patients are independent. To see that each pairing does indeed represent two different patients, note that each patient's stochastic process is uniquely indexed by the patient type, k, the week in which they are admitted, n, the day of the week they were admitted, s,

and their admission number on the day they are admitted, j. Therefore the indicators $\mathbb{1}\{L^{j_1,n_1}_{s_1,k_1}(t) = u_1\}$ and $\mathbb{1}\{L^{j_2,n_2}_{s_2,k_2}(t) = u_2\}$ are independent and their covariance term is necessarily zero.

(ii) In this case the two indicators represent the same patient at the same point in time. In this case, the two indicator random variables are *not* independent, since if a patient is in ward u_1 at time t they are clearly not in ward $u_2 \neq u_1$. For such pairs the covariance becomes

$$Cov\big(\mathbb{1}\{L^{j,n}_{s,k}(t) = u_1\}, \mathbb{1}\{L^{j,n}_{s,k}(t) = u_2\}\big) =$$
$$\mathbf{E}\big[\mathbb{1}\{L^{j,n}_{s,k}(t) = u_1\}, \mathbb{1}\{L^{j,n}_{s,k}(t) = u_2\}\big] - \mathbf{E}\big[\mathbb{1}\{L^{j,n}_{s,k}(t) = u_1\}\big]\mathbf{E}\big[\mathbb{1}\{L^{j,n}_{s,k}(t) = u_2\}\big] =$$

(2.36)

$$P(L^{j,n}_{s,k}(t) = u_1, L^{j,n}_{s,k}(t) = u_2) - P(L^{j,n}_{s,k}(t) = u_1)P(L^{j,n}_{s,k}(t) = u_2) = -p_{s,k,u_1}(t)p_{s,k,u_2}(t).$$

The last equality follows because the probability that a patient is in ward u_1 and in ward u_2 simultaneously is assumed to be zero. □

Theorem II.6

Proof. (i) In calculating ward census (see Eq. 4.3) any two indicator variables in the sum have the property that $(j_1, n_1, k_1, s_1) \neq (j_2, n_2, k_2, s_2)$, because the sum in Eq. 4.3 contains each combination of (j, k, n, s) at most once. Therefore the covariance is zero by Lemma II.5 (i). Thus the variance of the ward census can be calculated by

$$\sigma^2_{d_1,u}(\Theta) = \mathbf{Var}\Big[\sum_{d_2=1}^{7}\sum_{k\in\mathfrak{D}}\sum_{j=0}^{\Theta_{k,d_2}}\lim_{t\to\infty}\sum_{n=0}^{t}\mathbb{1}\big\{L^{j,n}_{d_2+7n,k}(d_1+7t))=u\big\}\Big]$$

$$= \sum_{d_2=1}^{7}\sum_{k\in\mathfrak{D}}\Theta_{k,d_2}. \tag{2.37}$$

$$\lim_{t\to\infty}\sum_{n=0}^{t} p_{d_2+7n,k,u}(d_1-d_2+7(t-n))(1-p_{d_2+7n,k,u}(d_1-d_2+7(t-n))). \tag{2.38}$$

The second equality follows by taking the variance inside the sum and from the fact that $\mathbb{1}\{\cdot\}$ is a Bernoulli random variable with variance $p(1-p)$.

(ii) Now consider the variance of the total hospital census. For all indicators where $(j_1, n_1, k_1, s_1) \neq (j_2, n_2, k_2, s_2)$, the covariance is zero by Lemma II.5 (i). However, since we are now summing over all wards, there are pairs of indicators for which this condition does not hold (i.e. we are considering whether a given patient is contributing to ward u_1 or to ward u_2 on a given day).

$$\sigma^2_{d_1}(\Theta) = \sum_{i=1}^{2}\sum_{a_i=1}^{7}\sum_{k_i\in\mathfrak{D}}\sum_{j_i=1}^{\Theta_{k_i,a_i}}$$
$$\lim_{t\to\infty}\sum_{n_i=0}^{t}\sum_{u_i\in\mathfrak{W}} Cov(\mathbb{1}\{L^{j_1,n_1}_{a_1+7n_1,k_1}(d_1+7t))=u_1\},\mathbb{1}\{L^{j_2,n_2}_{a_2+7n_2,k_2}(d_1+7t))=u_2\})$$
$$= \sum_{u\in\mathfrak{W}}\sigma^2_{d_1,u}(\Theta) - \sum_{a=1}^{7}\sum_{k\in\mathfrak{D}}\Theta_{k,a}\sum_{n=0}^{\infty}\sum_{u_1\neq u_2} p_{d_2+7n,k,u_1}(d(n))p_{d_2+7n,k,u_2}(d(n)).$$

The first equality follows from the definition of variance. The second equality follows by applying Lemma II.5 (i) to all terms representing different patients and then applying Lemma II.5 (ii) to those terms representing the same patient visiting different wards at the same time.

□

2.5.3 Important Considerations for Practical Application of Admission Schedule Optimization

In this section of the appendix we discuss several important considerations that must be taken into account when attempting to apply the theoretical methodology developed in this chapter to a real-world hospital. This section begins with a discussion of the data needs for parameterizing the model. Next, we discuss how the hospital would use the methodology on an on-going basis to ensure continued success by dynamically monitoring outcomes and rerunning the model as necessary. Finally, we talk about the optimal schedule that was generated during the case study and the economic ramifications of implementing this schedule in the partner hospital.

Model Parameterization from Hospital Data

This section discuss the types of data needed to parameterize the hospital scheduling optimization model. An example of the data required to calculate the patient care pathways is given in Figure 2.7. The Patient ID and Admit No. column enables the data mining algorithm to track individual patient pathways throughout the patient's hospital care segment. Patient ID is a unique identifier assigned to each patient, whereas the Admit No. is the number assigned to a particular visit for that patient. The algorithm can be run only using Admit No. and not Patient ID if necessary since the goal is to capture the care path for a particular type of admission. The ward column identifies which services the patient uses. Combining the ward information with the ward start time and end time enables the algorithm to calculate in what order the services are used, and for how long. Admitting service is typically used to classify patients into buckets and is often used for the "patient type" variable in the theoretical model. Other patient type classifiers can be used, but care must be taken to ensure that the patient type is a classifier that is known

when the patient is admitted (e.g. discharge diagnosis would not work as a patient type) and should classify the patients into buckets that are not statistically "too small." Finally, the Admission Type column provides information about whether the patient was an elective or emergency patient.

In all the hospitals that we have worked with, this data has been available through each hospital's IT systems. However, due to lack of standardization across IT systems, a significant amount of time and effort is often required to transform the data from its raw form into the useable format shown in Fig. 2.7 (see for example [32]). One complication that was encountered was a partner hospital that provided a patient's admit time and discharge time, but the transfers between wards were kept in a separate "transfer registry." The transfer registry was an Excel document that had one tab for each day of the time horizon that contained all the transfers that occurred in that day. Significant data processing was required to combine the different data streams to be able to track each patient's entire care path.

Another important piece of data required is the calculation of arrival rates for both emergency and scheduled patients. These can be calculated directly from the data shown in Fig. 2.7. There are a number of data mining tools available to capture this information. In the case study and other analyses, we relied primarily on the open source tool Hillmaker (see [48]).

Finally, data about the hospital infrastructure and policies are needed. This includes the size (in terms of beds) of each ward of the hospital and whether or not certain beds have restrictions and/or special equipment associated with them. It is also important to know the hospital's transfer policies in the event that the patient's preferred ward is full. Another key piece of information needed for a successful modeling venture is what restrictions there are on admissions. This may include what

Pat ID	admit No.	Ward	start time at ward	end time at ward	Admitting Service	Admission Type	Projected Discharge Date
1	101	EHH	2008-11-27 12:53:	2008-11-27 17:17:	Cardiology	Scheduled Surg.	11/28/2008
1	101	A3	2008-11-27 17:17:	2008-11-27 19:17:	Cardiology	Scheduled Surg.	11/28/2008
3	103	SEH	2008-12-30 23:52:	2008-12-31 00:08:	Internal Medicine	Emergency	1/1/2009
3	103	A4	2008-12-31 00:08:	2008-12-31 11:36:	Internal Medicine	Emergency	1/1/2009
3	103	C4	2008-12-31 11:36:	2009-01-02 21:56:	Internal Medicine	Emergency	1/1/2009
6	106	B4	2008-12-31 01:59:	2008-12-31 05:52:	Neurology	Emergency	1/3/2009
6	106	B4	2008-12-31 05:52:	2008-12-31 12:38:	Neurology	Emergency	1/3/2009
6	106	C5	2008-12-31 12:38:	2008-12-31 21:01:	Neurology	Emergency	1/3/2009
6	106	B4	2008-12-31 21:01:	2009-01-01 15:12:	Neurology	Emergency	1/3/2009
10	110	SEH	2008-12-31 08:57:	2008-12-31 09:02:	General surg	Emergency Surg.	12/31/2008
10	110	A2	2008-12-31 09:02:	2008-12-31 20:00:	General Surg	Emergency Surg.	12/31/2008

Figure 2.7: Sample of hospital data that can be used to parameterize schedule optimization models.

the current block schedule looks like and which surgeons are amenable to changing their schedule and which are not, what are the upper limits for surgeries/admissions for each type of patient, and if there are any other restrictions on the admission process.

Data Characterization of Partner Hospital

The table in Fig. 2.8 shows the arrival rates (mean and standard deviation) for the partner hospital that used to validate the census model in Sec. 2.2.6.

Maintenance of an Implemented Scheduling System

In this section, we discuss some of the issues regarding ongoing maintenance of a scheduling system in a real-world hospital. The initial design of the scheduling system is intended to model the system dynamics of the partner hospital at the current time point. However, hospitals are dynamically changing entities so it is reasonable to consider updating the schedule to adjust to changes in the underlying system dynamics that may occur over time. Important changes to the system may include a change in the emergency arrival rate, building new beds or new wings, hiring new surgeons, adopting new processes for patient treatment that change length of stay and resource usage.

One mechanism used in similar scheduling systems to monitor the system for significant changes in underlying dynamics is to monitor key metrics using control charts (see for example [32]). Metrics that are especially relevant are occupancy levels, cancelation rates, and emergency department congestion among other things. If these metrics begin to exceed control limits around what the model predicts it indicates the possibility of a shift in the underlying dynamics.

In the event that the underlying system dynamic do change, the solution is to repa-

Mean	DOW	Elec Mean	Elec StdDev	Emerg Mean	Emerg Std Dev
A2	Sun	2.38	1.55	3.10	1.82
A2	Mon	4.87	2.49	4.54	2.88
A2	Tue	4.79	2.33	5.19	2.73
A2	Wed	3.79	1.84	5.25	2.95
A2	Thu	5.96	2.37	3.50	2.65
A2	Fri	5.17	2.27	5.15	2.48
A2	Sat	1.51	1.06	4.58	2.67
A3	Sun	2.11	1.27	4.02	2.82
A3	Mon	3.45	1.72	5.73	2.79
A3	Tue	4.66	1.94	5.79	2.44
A3	Wed	4.34	1.81	6.19	2.82
A3	Thu	4.89	2.71	6.40	3.21
A3	Fri	3.36	1.98	6.50	2.85
A3	Sat	0.98	1.17	4.44	1.99
A4	Sun	3.30	3.04	0.75	1.05
A4	Mon	5.15	2.66	1.15	1.45
A4	Tue	4.06	2.34	1.10	1.21
A4	Wed	3.91	1.87	0.89	1.25
A4	Thu	3.28	1.85	0.67	0.88
A4	Fri	4.30	2.11	0.96	1.10
A4	Sat	1.38	1.75	0.73	0.93
B1	Sun	0.00	0.00	0.00	0.00
B1	Mon	26.23	16.40	0.00	0.00
B1	Tue	22.68	12.03	0.00	0.00
B1	Wed	18.83	9.96	0.00	0.00
B1	Thu	28.17	14.63	0.00	0.00
B1	Fri	9.00	6.86	0.00	0.00
B1	Sat	0.00	0.00	0.00	0.00
B3	Sun	0.53	0.75	2.50	1.80
B3	Mon	3.19	1.81	3.37	2.03
B3	Tue	5.34	2.24	2.94	1.89
B3	Wed	3.53	1.67	2.42	1.74
B3	Thu	3.53	2.09	3.54	2.37
B3	Fri	1.49	1.32	4.08	3.28
B3	Sat	0.51	0.83	2.69	2.21
B4	Sun	5.85	3.80	6.50	3.77
B4	Mon	8.02	4.83	7.54	3.93
B4	Tue	10.43	5.21	7.17	3.97
B4	Wed	12.30	4.46	5.62	3.54
B4	Thu	10.21	4.38	6.04	4.09
B4	Fri	10.87	4.64	6.35	3.20
B4	Sat	6.09	2.98	6.31	3.49
C2E	Sun	0.00	0.00	0.00	0.00
C2E	Mon	27.53	7.89	0.02	0.14
C2E	Tue	32.23	3.13	0.04	0.19
C2E	Wed	30.96	5.75	0.21	0.45
C2E	Thu	31.98	6.24	0.19	0.53
C2E	Fri	26.83	5.04	0.12	0.32

Mean	DOW	Elec Mean	Elec StdDev	Emerg Mean	Emerg Std Dev
C2E	Sat	0.00	0.00	0.00	0.00
C2O	Sun	0.00	0.00	0.00	0.00
C2O	Mon	16.68	6.56	1.37	1.34
C2O	Tue	21.55	4.80	1.37	1.62
C2O	Wed	22.53	6.15	0.49	0.78
C2O	Thu	21.36	6.09	0.94	1.14
C2O	Fri	20.02	7.62	0.88	1.44
C2O	Sat	0.06	0.32	0.19	0.53
C4	Sun	0.30	0.59	3.81	2.18
C4	Mon	0.87	0.80	6.15	2.59
C4	Tue	1.19	0.90	5.48	2.25
C4	Wed	1.30	1.55	5.15	2.42
C4	Thu	1.49	1.27	5.40	2.81
C4	Fri	1.32	1.52	6.08	3.20
C4	Sat	0.30	0.62	3.73	2.35
C5	Sun	0.49	0.75	3.52	1.84
C5	Mon	4.15	2.43	3.73	2.26
C5	Tue	4.77	2.50	3.71	2.35
C5	Wed	4.11	2.35	3.53	2.66
C5	Thu	3.34	2.24	3.42	2.66
C5	Fri	3.53	1.95	4.38	2.70
C5	Sat	0.64	1.07	3.71	2.57
CCU	Sun	0.38	0.61	3.04	1.58
CCU	Mon	1.30	1.08	3.54	1.84
CCU	Tue	1.30	1.12	3.60	1.86
CCU	Wed	1.36	1.17	3.49	1.95
CCU	Thu	1.23	1.22	3.35	2.08
CCU	Fri	1.26	1.15	3.44	2.35
CCU	Sat	0.36	0.64	3.12	1.63
EHH	Sun	0.00	0.00	0.04	0.19
EHH	Mon	1.06	1.24	4.21	2.47
EHH	Tue	1.45	1.61	3.33	2.17
EHH	Wed	1.13	1.50	3.64	2.28
EHH	Thu	1.06	1.37	3.90	2.44
EHH	Fri	1.30	1.84	3.90	2.42
EHH	Sat	0.00	0.00	0.00	0.00
ICU	Sun	0.15	0.47	0.87	0.79
ICU	Mon	0.91	0.97	1.02	1.02
ICU	Tue	0.57	0.80	1.04	0.99
ICU	Wed	0.89	0.96	1.04	0.92
ICU	Thu	0.53	0.78	0.90	1.00
ICU	Fri	0.72	0.71	1.27	0.97
ICU	Sat	0.21	0.41	0.60	0.77
OPFL1	Sun	0.72	1.23	3.77	3.90
OPFL1	Mon	2.45	3.36	4.50	4.47
OPFL1	Tue	3.60	3.63	4.35	3.90
OPFL1	Wed	3.06	4.05	3.53	3.58
OPFL1	Thu	4.09	4.32	3.12	4.18

Mean	DOW	Elec Mean	Elec StdDev	Emerg Mean	Emerg Std Dev
OPFL1	Fri	3.11	3.31	3.62	4.35
OPFL1	Sat	0.57	1.10	5.27	4.73
PCHI1	Sun	0.00	0.00	0.00	0.00
PCHI1	Mon	0.09	0.28	0.02	0.14
PCHI1	Tue	4.94	1.72	0.00	0.00
PCHI1	Wed	0.09	0.46	0.00	0.00
PCHI1	Thu	4.30	2.07	0.00	0.00
PCHI1	Fri	0.02	0.15	0.00	0.00
PCHI1	Sat	0.00	0.00	0.00	0.00
PCHI2	Sun	0.00	0.00	0.00	0.00
PCHI2	Mon	0.00	0.00	0.00	0.00
PCHI2	Tue	2.74	1.54	0.00	0.00
PCHI2	Wed	0.00	0.00	0.00	0.00
PCHI2	Thu	0.02	0.15	0.00	0.00
PCHI2	Fri	0.00	0.00	0.00	0.00
PCHI2	Sat	0.00	0.00	0.00	0.00
PKNO1	Sun	0.00	0.00	0.00	0.00
PKNO1	Mon	0.00	0.00	0.00	0.00
PKNO1	Tue	0.00	0.00	0.00	0.00
PKNO1	Wed	0.09	0.35	0.00	0.00
PKNO1	Thu	0.00	0.00	0.00	0.00
PKNO1	Fri	0.00	0.00	0.00	0.00
PKNO1	Sat	0.00	0.00	0.00	0.00
PKNO2	Sun	0.00	0.00	0.00	0.00
PKNO2	Mon	0.02	0.15	0.00	0.00
PKNO2	Tue	0.00	0.00	0.00	0.00
PKNO2	Wed	0.00	0.00	0.00	0.00
PKNO2	Thu	0.00	0.00	0.00	0.00
PKNO2	Fri	0.00	0.00	0.00	0.00
PKNO2	Sat	0.00	0.00	0.00	0.00
PMHK1	Sun	0.00	0.00	0.00	0.00
PMHK1	Mon	0.36	0.67	0.02	0.14
PMHK1	Tue	0.49	0.75	0.00	0.00
PMHK1	Wed	0.21	0.51	0.02	0.14
PMHK1	Thu	0.49	0.78	0.00	0.00
PMHK1	Fri	0.49	0.69	0.02	0.14
PMHK1	Sat	0.02	0.15	0.00	0.00
PNEU1	Sun	0.00	0.00	0.00	0.00
PNEU1	Mon	0.64	0.92	0.00	0.00
PNEU1	Tue	0.00	0.00	0.00	0.00
PNEU1	Wed	0.00	0.00	0.00	0.00
PNEU1	Thu	0.00	0.00	0.00	0.00
PNEU1	Fri	0.00	0.00	0.00	0.00
PNEU1	Sat	0.00	0.00	0.00	0.00
PNEU2	Sun	0.00	0.00	0.00	0.00
PNEU2	Mon	0.00	0.00	0.00	0.00
PNEU2	Tue	0.00	0.00	0.00	0.00
PNEU2	Wed	0.00	0.00	0.00	0.00

Mean	DOW	Elec Mean	Elec StdDev	Emerg Mean	Emerg Std Dev
PNEU2	Thu	1.45	0.90	0.00	0.00
PNEU2	Fri	0.00	0.00	0.00	0.00
PNEU2	Sat	0.00	0.00	0.00	0.00
REC2	Sun	0.00	0.00	0.00	0.00
REC2	Mon	10.11	6.00	0.00	0.00
REC2	Tue	15.34	5.19	0.00	0.00
REC2	Wed	2.17	3.61	0.00	0.00
REC2	Thu	14.51	5.51	0.00	0.00
REC2	Fri	8.17	4.19	0.02	0.14
REC2	Sat	0.00	0.00	0.00	0.00
SEH	Sun	0.36	0.61	16.33	3.78
SEH	Mon	0.23	0.48	17.88	4.18
SEH	Tue	0.13	0.34	17.08	4.04
SEH	Wed	0.19	0.50	16.08	3.75
SEH	Thu	0.15	0.42	16.63	4.73
SEH	Fri	0.13	0.40	18.02	4.24
SEH	Sat	0.19	0.50	17.33	3.85
Total	Sun	16.57	5.51	48.23	8.44
Total	Mon	117.32	32.98	64.79	13.23
Total	Tue	142.26	17.98	62.17	11.64
Total	Wed	114.79	23.88	57.53	10.59
Total	Thu	142.04	28.93	58.02	13.89
Total	Fri	101.09	20.49	64.79	10.96
Total	Sat	12.83	4.76	52.69	8.94

Figure 2.8: Arrival vectors for elective and emergency patients by ward for partner hospital.

rameterize the model taking into account the new data that are reflecting the new dynamics. This can be done selectively or can constitute a full reparameterization depending on the extent of the changes in the underlying system. Once the system has been recalibrated, the optimization can be rerun to generate an adapted optimal schedule. One of the major benefits of having an analytical model, as opposed to a simulation model, is that rerunning the optimization is nearly instantaneous and doesn't require significant manual effort. Recalibrating the system and generating an improved schedule is much faster and requires less manual input with the analytical model. This also paves the way for a more automated, dynamically updating scheduling system.

To model this hospital, a full year's worth of data is used with identifying patient information removed and replaced by admission numbers. Given that our system is modeling a hospital based on its daily (midnight) census, we only consider patients that stayed in the hospital for at least one night. In 2008, 14,827 patients stayed *at least* one night. Out of these overnight patients, 7,016 were emergency patients while the remaining 7,811 were scheduled patients.

The input data contained the 14,827 patients' movements throughout the hospital. The patients transferred within the hospital 20,462 times, including the initial 'transfer' into the patient's first ward. The transfers within the hospital had been grouped into 23 ward codes by the partner hospital; to avoid unnecessary complexity we aggregate similar wards, where three aggregate wards (which make up 8 of the 23 wards) constitute the majority of transfers (the remaining 15 wards were rarely used by the patients). The first aggregate ward ("Ward A") is a surgical ward. The second aggregate ward ("Ward B") is a medicinal ward. The third aggregate ward ("Ward C") consists of the critical care unit (CCU) and intensive care unit (ICU)

Table 2.4: Transition probabilities for non-emergency and emergency patients.

	Non-Emergency				Emergency			
To:	–	A	B	C	–	A	B	C
From A	0.846	0.116	0.001	0.037	0.739	0.174	0.008	0.079
From B	0.847	0.004	0.147	0.002	0.699	0.026	0.271	0.004
From C	0.371	0.528	0.069	0.031	0.429	0.543	0.014	0.044

of the hospital. This aggregation works well because the statistical properties of patients at the sub-ward level do not differ significantly and the increased sample size for each ward provides better statistical estimates. Since the purpose of our approach is to evaluate system level properties, this loss of granularity has little effect on the system level outcome as noted in [63].

Transition Probabilities

For both emergency and non-emergency patients at each aggregate ward, one-step ward transition probabilities and length of stay parameters are presented. Table 2.4 gives the one-step transition probabilities for non-emergency and emergency patients. Here, "–" implies a discharged patient. Transitions from a ward back to itself (e.g. a transfer from Ward A to Ward A) may occur for multiple reasons. First, a patient's condition may change, which results in the patient's information (and possibly location within a given ward) being updated. Second, a patient may transfer from a ward in Ward A to another ward which also happens to be in Ward A – thus, even though the patient transfers from one ward to another, our method of aggregating the wards views such movements as internal transfers.

The length of stay data at each aggregate ward is calculated for both types of patients. The mean μ and standard deviation σ at individual wards are weighted appropriately in order to find aggregate versions of these parameters for non-emergency and emergency patients, which are listed in Table 2.5.

Arrival of Emergency Patients

Table 2.5: Average and standard deviation of the length of stay (in hours) for non-emergency and emergency patients.

	Non-Emergency		Emergency	
Ward	μ	σ	μ	σ
A	27.13	12.87	118.36	131.12
B	23.47	9.68	56.31	65.82
C	21.99	8.20	49.66	123.28

To properly model arrivals, one must consider that emergency patients may (i) arrive at Wards A, B, and C according to a distribution that differs from their intra-hospital transfer rates, and (ii) do not arrive uniformly throughout the week. Consequently, we compute these values as follows. First, we group emergency patients based on the *first* aggregate ward they visit (i.e. Ward A, B, or C). Many patients arrive first to a central triage, and then transfer within the hospital to one of the 23 wards. For each patient, we follow their transfers until they first enter one of the wards aggregated into Ward A, B, or C. For each of the three wards, we determine how many emergency patients are admitted between midnight and 2 p.m. (which we refer to as "AM") and 2 p.m and 11:59 p.m. ("PM") on *each day* of the year. Further, it is well-known that emergency patients do not arrive uniformly over the course of the day, which leads to increased queueing. The emergency patients in our study generally followed the daily arrival pattern found by previous studies – refer to [16] and [81] for empirical distributions.

After determining how many emergency patients arrive at each ward in the AM and PM time blocks for each day of the year, we find the mean and standard deviation of patients arriving at each ward in the AM and PM blocks by day of week. These values are summarized in Table 2.6. The arrival pattern of non-emergency patients is also of interest in order to evaluate the current system. As one would generally expect, emergency patients do not arrive uniformly over the course of the week, in

Table 2.6: Average number of arrivals, separated by ward and time of day.

	Non-Emergency						Emergency					
Ward:	A		B		C		A		B		C	
	AM	PM	AM	PM	AM	PM	AM	PM	AM	PM	AM	PM
M	8.44	1.38	2.04	1.94	0.35	0.13	1.94	3.17	2.19	2.48	1.15	1.63
Tu	8.45	1.45	6.15	2.19	0.25	0.11	2.03	2.81	1.89	2.53	1.02	1.60
W	7.85	1.32	5.15	1.75	0.25	0.09	1.96	2.87	1.43	2.00	0.92	1.60
Th	9.42	1.18	4.04	1.62	0.17	0.06	1.98	2.50	1.67	2.77	1.23	1.27
F	7.12	0.73	3.19	2.02	0.25	0.08	1.94	3.48	1.67	2.63	1.00	1.44
Sa	0.17	0.12	1.00	0.94	0.02	0.10	1.81	3.48	1.67	2.63	1.00	1.44
Su	1.13	4.87	0.87	1.38	0.12	0.08	1.58	2.17	1.65	2.23	1.12	1.40

addition to the heterogeneity over the course of a day. Subsequently, we use the mean number of arrivals (grouped by ward, day of week, and AM/PM) in order to model emergency arrivals using a Poisson distribution.

These calculations – the one-step transition probabilities and length of stay for both emergency and non-emergency patients at Wards A, B, and C, as well as the patients' arrival locations, grouped by day of week and time of day – all serve as inputs.

The optimal schedules for the minimum blockages and the maximum electives formulations contrasted with the original schedule from the case study presented in Sec. 4.7 are given below. An important feature of this particular case study is the fact that the partner hospital performed some admissions on the weekends. In the optimization, we constrained the maximum number of weekend admissions by the amount done historically in the original schedule. The limits were taken to be as follows: (1) General Surg admissions are capped at 0 for Saturday and 6 for Sunday, (2) Medicine admissions are capped at 2 for Saturday and 2 for Sunday.

Admitting Service	Sun	Mon	Tue	Wed	Thu	Fri	Sat
General Surgery	6	11	10	9	11	8	0
Internal Medince	2	4	9	7	6	5	2
Total	8	15	19	16	17	13	2

Table 2.7: Original schedule

Admitting Service	Sun	Mon	Tue	Wed	Thu	Fri	Sat
General Surgery	6	13	5	14	2	15	0
Internal Medicine	2	6	8	2	11	4	2
Total	8	19	13	16	13	19	2

Table 2.8: Minimum blocking schedule (32% reduction)

Admitting Service	Sun	Mon	Tue	Wed	Thu	Fri	Sat
General Surgery	6	13	6	12	5	13	0
Internal Medicine	2	8	8	6	9	6	2
Total	8	21	14	18	14	19	2

Table 2.9: Maximum elective admissions schedule (7% increase)

Economic Ramifications of Schedule Design

This section provides a high-level look at the economics of optimal schedule design. The purpose is not to provide a complete financial analysis of the hospital care delivery system, but rather to give a broad sense of the magnitude of impact of scheduling on hospitals.

When a hospital increases the volume of elective procedures done, as is the case in the maximum elective admissions optimization, the hospital benefits from both the procedure itself as well as the subsequent inpatient stay. Because the choice of which procedures to target for increased volume lies primarily in each individual hospitals strategic mission, we do not attempt to analyze every possible scenario; instead taking the average expected income generated across a number of common specialties. Table 2.10 gives the average revenue generated and relative volumes of procedures for 100 surgical centers (see [78]). From the table, the average revenue per procedure across these 5 services is around $1,829.

	General Surg	OB/GYN	Ortho	Plastics	Urology
Avg Revenue per Proc.	1545	1757	2443	1415	1435
Pecent of Volume	0.17	0.06	0.36	0.11	0.31

Table 2.10: Mean income per procedure generated by specialty and relative volume of each specialty across 100 surgical centers

In addition to procedure revenue, there is also the revenue generated by an inpatient stay. According to a survey of 114 hospital CFO's (see [37]), the average revenue generated by an inpatient stay is approximately $8,500.

In the maximum elective admissions optimal schedule in the hospital case study presented in Sec. 4.7, the hospital was able to perform an additional 310 additional elective procedures annual with the same level of access. The rough-cut financial impact of this increased procedure volume for the hospital is 310 procedures/year $\times$ ($1,829 per procedure + $8,500 per inpatient admission) = $3,201,990 in additional revenue annually. In addition, the hospital from the case study was very well managed already so the opportunities would likely be greater in a more typical hospital.

CHAPTER III

Design and Analysis of Hospital Admission Control for Operational Effectiveness

This chapter complements Chapter II by providing a solution to the dynamic control portion of the Hospital Admission Scheduling and Control (HASC) problem. In Chapter II, analytical methods were developed for the system-wide planning of scheduled elective admissions to stabilize workloads in downstream hospital resources. Even with an improved admission schedule, variability in length of stay and emergency arrivals will still impact the hospital. This chapter looks at dynamic, operational-level control of day-to-day hospital census as a complement to the planning and scheduling model presented in Chapter II.

Currently there are two major gateways for admission to a hospital: the ED and scheduled elective admission. Unfortunately, in highly utilized hospitals, excessive wait times make the scheduled gateway undesirable or infeasible for a subset of patients and doctors. As a result, this group often uses the ED gateway as a means to gain admission to the hospital. To better serve these patients and improve overall hospital functioning, we propose creating a third gateway: an expedited patient care queue. We first characterize an optimal admission threshold policy using controls on the scheduled and expedited gateways for a new Markov Decision Process model. We then present a practical policy based on insight from the analytical model that

yields reduced emergency blockages, cancelations and off-unit census via simulation based on historical hospital data.

3.1 Introduction

In the United States, health care lags significantly behind the manufacturing sector in process improvement practice. One consequence of this is that hospital care services are subject to significant, unnecessary and detrimental fluctuations in patient census and associated workload. This variability in patient census has been linked specifically to congestion and chaos in the Emergency Department (ED), excessive radiology backlogs, strains on nurse and ancillary staff, and overcrowding in the Post Acute Care Unit (PACU) to name a few. This system-wide congestion results in compromised quality of care, emergency patient blockage for lack of beds, excessive patient Length of Stay (LOS), and significant understaffing and overstaffing costs (see [20], [34], [84], [31], [64] and [75]).

This chapter introduces and evaluates mechanisms for managing the variability in hospital workload at a key source: inpatient admissions. Currently, most hospitals use only one mechanism for daily admission control; that is, they reactively cancel elective surgeries only when there are no more inpatient beds available. In such systems, no control exists to increase the short-term utilization of hospital resources. This occurs because hospitals often categorize admissions as either scheduled elective or emergency patients, foregoing the potential to redesign the ED and admissions to accommodate patients who need to be seen within a few days but are not true emergency cases. As a result, hospitals have some control over the arrival rate of scheduled patients, but very little control over the arrival rate of emergency patients.

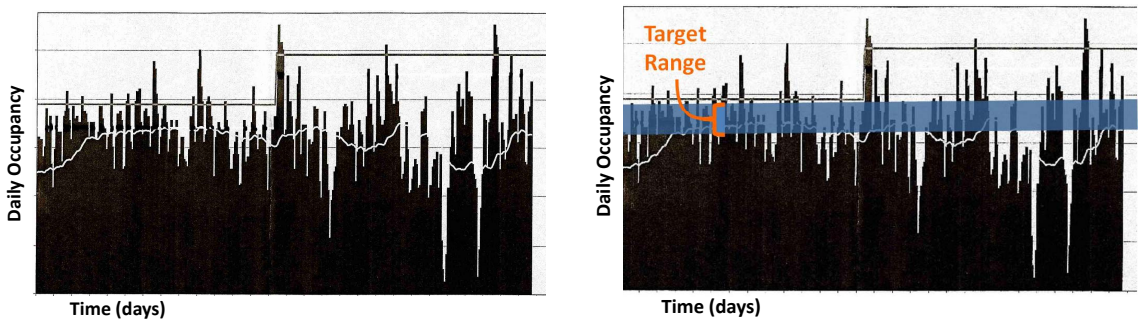
It can be counterproductive to lump the entire range of unscheduled patient types into one category: emergency. As has been noted, (see for example [70], [60] and [26])

one can identify a third category of patient that we refer to here as *expedited patients*. For patients in this category the acuity of their medical condition is less than most ED patients who are admitted, and their admission to the hospital can be delayed one to three days, for example, without compromising their treatment. Without an efficient expedited admission process, these patients are often admitted through the emergency department due to excessive waiting times if they seek admission as an elective patient.

This chapter makes strides towards developing a dynamic control structure that effectively employs a call-in mechanism for servicing this third class of patient and a controlled cancelation mechanism of elective patients. By properly servicing expedited patients, the excess load they placed on the ED during periods of peak congestion is reduced. This also reduces the arrival rate of uncontrollable random admissions. Thus the expedited call-in queue can be used to smooth hospital occupancy levels over time to increase utilization of a hospital's expensive resources, such as staff and beds. The effect of creating both a call-in and a proactive cancelation mechanism is to squeeze the hospital occupancy variation (see Figure 3.1), while leaving a sufficient capacity buffer to accommodate potential future emergency arrivals. This, in turn, increases the quality of care, facilitates patients' access to health care, and decreases the overall hospital costs resulting from occupancy variability.

This chapter develops a stochastic model for dynamic inpatient admission control that uses detailed information on the number of occupied beds by bed unit (ward) to show that:

1. Using both model-based (1) proactive cancelation and (2) call-in control mechanisms has advantages to using only reactive cancelation, as is the prevailing practice.



(a) System's Current Occupancy (b) System's Target Occupancy

Figure 3.1: Controlling census variability in hospitals.

2. An easily implementable multi-dimensional double threshold policy for controlling both these mechanisms can effectively balance the opportunity cost of unfilled beds against the potential for cancelation of electives and bed block for emergency and elective patients.

It is important to make a distinction between our proposed dynamic *control* approach versus *scheduling* or *planning* models. Admission scheduling models are typically concerned with decisions to generate efficient schedules (see [5], [43], [68]) or capacity plans (see [87]) for operating rooms, diagnostic labs (see [74]) or outpatient clinics (see [28]). The schedules are created by assigning patients to time slots, sequencing the operations or dynamically managing the capacity. In contrast, we propose a closed-loop (i.e. feedback driven) control model linking system-wide behavior to operational-level decisions to stabilize hospital occupancy. An optimized scheduling system is not a substitute for our proposed control approach, but rather a preamble for it. In practice, our control mechanism would take the schedule as an input and provide a planned response to the realization of the schedule and exogenous random events; we simulate such an approach in Section 3.4.

This chapter contributes to the literature a clearer understanding of the value of feedback control for inpatient admissions. First, it develops a stylized model

to generate insight into the structure of a typical admission system. Second, it argues the optimality of a double threshold admission policy for this stylized model and proves several properties for two new queueing operators. Third, the insight from the analytical model is used to develop a practical hospital admission policy with significant potential to improve health care delivery, which is demonstrated via simulation on historical hospital data.

The structure of this chapter is as follows. In Section 3.2 we discuss the motivation and background of hospital admission control systems and review the related literature. In Section 3.3 we develop and analyze two stochastic models for understanding the essential dynamics of hospital admission control systems. In Section 3.4 we propose a practical admission policy based on the insight from Section 3.3 and use a simulation framework to demonstrate the benefits for real-world applications.

3.2 Models for Hospital Admission Control

Based on our discussions with several major hospitals in the state of Michigan, 40% to 60% of inpatients are admitted through the ED. The Emergency Medical Treatment and Active Labor Act (EMTALA) requires U.S. hospitals, by law, to medically stabilize emergency patients. Though ED services are significantly impacted by inpatient admissions from other hospital services, hospitals often lack an effective system for coordinating of hospital admissions, bed units and the ED. [71] argue that a primary cause of ED overcrowding and ambulance diversion is a lack of empty inpatient beds for emergency patients. ED overcrowding negatively impacts the quality of patient care, leading to increased LOS and/or worsening patient disposition (see [20]) and in extreme cases, increased mortality rate (see [84]).

In the U.S., ED overcrowding is also exacerbated because many non-emergency patients use the ED as a convenient means to get admission into a highly utilized

hospital. While it is not currently common practice, we seek to show that ED congestion can be reduced by using a patient flow management system that effectively employs an expedited call-in queue and uses the hospital census levels in admission decisions. We review the literature on patient flow models that link census to admission decision making (Section 3.2.1) and the use of expedited call-in queues in hospitals (Section 3.2.2).

3.2.1 Patient Flow Modeling and Linking Admission to Census

[17], [12],and [82] represent early efforts in the modeling of hospital occupancy and admissions. [36] used a multistage stochastic approach to establish that variability in daily hospital occupancy in combination with high occupancy levels can increase the risk of hospital overflows.

Using simulation, [32] developed (1) an inpatient admission scheduling and control system to achieve high average occupancy subject to constraints on the number of cancelations and emergency diversions, and (2) prediction models for the maximum average occupancy attainable using their control system. [52] and [49] employed surveys to establish that an effective patient flow model can enable high patient throughput, low patient wait times, short LOS, and low clinic overtime. [30] provided a systematic approach to health care engineering based on four different levels of granularity: *macro*, *regional*, *center* and *department*. Our model targets the "center" (hospital level) and can be used at the departmental level. [76] used forecasting and simulation to predict and manage inpatient flows into hospital beds. This chapter builds upon the above work by using feedback driven control to optimize the use of reactive admission mechanisms with respect to system-level metrics. Recent advances in patient flow modeling can also be found in [7] and [85].

3.2.2 An Expedited Call-in Queue for Quick Response

Long wait times for elective (scheduled) admission – not only for the hospital but for primary care providers as well – can force patients to use the ED as an expedient means for hospital admission. A lack of bed availability sometimes causes doctors to work around the system by declaring an emergency – even if a true emergency does not exist – to get their patients admitted more quickly. This "work-around" behavior strains the ED and subverts proper admission control. Previous literature has established a significant need for better management of patients with time-sensitive, but not emergent needs (see for example [70, 33]). While there is opportunity through research to establish the current level of emergency department misuse we have verified through discussions with medical practitioners in several countries that the practice of using the emergency department to gain admission to the hospital persists. The exact magnitude of the problem is left to future research. Our goal is to extend the concept of a call-in queue to meet the needs of patients requiring "expedited" inpatient care and allow them to be seen within a few days (a faster turnaround than for a typical elective admission). The literature below demonstrates the efficiency of patient call-in queues in a variety of settings.

A call-in mechanism gives patients who need expedited care a new pathway into the hospital and thus reduces the load on the ED. Because a majority of hospital costs are fixed, an empty hospital bed carries a significant *opportunity cost* (see [64] and [57]). The expedited call-in queue can be used to increase the hospital occupancy during low occupancy periods to avoid this opportunity cost. Streamlined management of an expedited call-in queue for inpatients has been partially implemented before (see [70], [60] and [26]). Though call-in queues are not commonly used in the US, this practice is more prevalent in Europe. [86] analyzes the practice of using

inpatient call-in queues in the UK. [73] discuss how establishing a pool of on-call outpatients can improve resource utilization in an under-capacitated diagnostic center. On-call patient queues are also widely used for outpatient clinics (see [55]) and elective surgeries (see [6]).

3.3 A Markov Decision Process Model for Hospital Admission Control

Our dynamic admission control approach employs two mechanisms – elective admission cancelation and the call-in of expedite patients from a waiting list – to strike a balance between bed utilization and hospital congestion. The purpose of this section is to develop intuition into the structure of an optimal admission policy that is used to build the practical admission control mechanism presented in Section 3.4. For the sake of intuition we present a stylized Markov Decision Process (MDP) model that focuses only on key dynamics of an effective admission control system.

Of relevance to our methodology, [23] presented a dynamic programming model to optimize elective surgery schedules considering uncertainty in operating room capacity due to emergency arrivals. [25] used a dynamic programming approach to control admission of inpatients, outpatients and emergency patients into a diagnostic medical facility. [74] developed an approximate dynamic programming approach for scheduling multi-priority patients to a public diagnostic facility. [15] analyzes the effect of reserving slots for urgent patients in the context of primary health care practice using Markov Decision Process models. See [27] and [28] for a detailed survey on the use of Markov Decision models for appointment scheduling and controlling admissions in a variety of health care delivery environments.

3.3.1 The Model

Our proposed MDP models seek to balance the opportunity cost of unutilized resources with the penalties associated with heavy congestion in those resources.

In particular we consider balancing the opportunity cost of unfilled hospital beds with the penalties of (1) canceling elective admissions and (2) emergency arrivals that are blocked from entering a hospital bed. By analyzing the structure of this model, we determine the form of an optimal admission control policy to support the development of specific, practical policies for application.

For the general model consider a queueing system with 2 queues – a call-in queue and the hospital itself – as shown in figure 3.2. Scheduled patients and emergency patients are assigned hospital beds according to a Poisson process with rate λ_s' and λ_e' while the call-in patients are placed on the call-in queue according to a Poisson process with rate λ_q' and are then assigned a hospital bed (admitted) via the admission controller's call-in action. The controller also has the option of canceling an arriving scheduled patient. Patients receive service according to an exponential distribution with rate μ'.

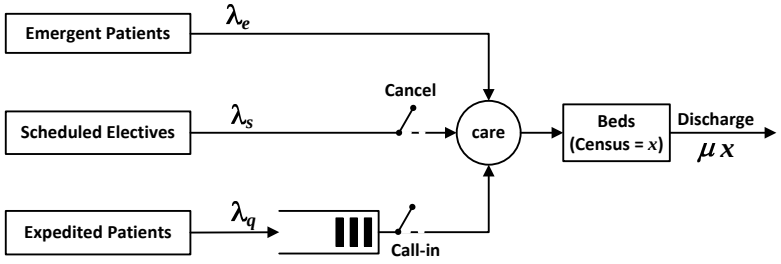


Figure 3.2: Two-dimensional Admission control system.

Let $\mathbf{X}^\pi(t) = (X_1^\pi(t), X_2^\pi(t))$ denote the number of patients in the system at time t under policy π, where $X_1^\pi(t) \in \mathbb{Z}^+$ is the number of patients in the hospital and $X_2^\pi(t) \in \mathbb{Z}^+$ is the number of patients on the call-in queue. Let B be the number of inpatient hospital beds. Let h_1' be the opportunity cost of an *empty* bed – quantified as $(B - X_1^\pi(t))^+$. In contrast, h_2' represents the per unit time (inventory) holding cost associated with holding a patient in the call-in queue.

Let τ' be the cost associated with having too many patients in the hospital – quantified as $(X_1^\pi(t) - B)^+$. When an emergency patient arrives at the door, it is forbidden by law to turn them away even if the hospital is full. This means that when a hospital reaches peak occupancy, emergency patients start to back up in the Emergency Department and/or are placed in the halls until a bed opens up. Similarly, scheduled surgeries remain in the Operating Room, occupying critical OR resources if there is no bed available for them post operation. These are adverse situations for the hospital and the patient and thus a penalty of τ is assessed for each patient that is forced to wait in the OR, the ED, or in the hallway for a bed to open up.

Finally, let c' be the cost of canceling a scheduled patient. If we let $N^\pi(t)$ be the counting process for the number of cancelations by time t under policy π, we can formulate the average cost per unit time objective function for policy π as

(3.1)
$$Z^\pi = \limsup_{T\to\infty} E\frac{1}{T}\left[\int_0^T \left(h_1'(B - X_1^\pi(z))^+ + \tau'(X_1^\pi(z) - B)^+ + h_2' X_2^\pi(z)\right)dz + c' N^\pi(T)\right]$$

A policy is then optimal if it achieves the optimal cost among all admissible policies Π as follows:

(3.2)
$$Z^* = \inf_{\pi\in\Pi} Z^\pi.$$

Our goal is to use the structure of the optimal policy for this model to gain insight into hospital management practice. The following theorem, proved in [44], shows this goal can be achieved by analyzing the finite horizon discounted version of the problem.

Theorem III.1. *For the average cost optimality equation, if $\lambda_e' + \lambda_q' < B\mu'$ then the following hold.*

(i) There exists an average-cost optimal stationary policy.

(ii) The optimal average cost can be computed by: $Z^* = \inf_{\pi \in \Pi} Z^\pi = \lim_{\beta \to 1^-} \lim_{n \to \infty} (1-\beta) V_{n,\beta}(x)$*, where* $V_{n,\beta}(x)$ *is the n stage discounted value function*

(iii) Let $\pi_{n,\beta}$*, denote an optimal policy for the n-period (discounted cost) problem. Then any limit point* π_β *of the sequence* $\{\pi_{\beta,n}\}_{n \geq 1}$ *as* $n \to \infty$ *is optimal for the infinite-horizon discounted cost. Moreover, any limit point of the sequence* $\{\pi_\beta\}_{\beta \in (0,1)}$ *(as* $\beta \to 1^-$*) is average-cost optimal.*

Theorem III.1 allows us to (1) guarantee the existence of an optimal stationary policy and (2) establish that the finite horizon problem converges to the infinite horizon average cost problem in both cost and policy. This means that the insights from analyzing the finite horizon problem are conferred to the original infinite horizon average cost optimal problem in Eq. (3.2). The condition $\lambda_e' + \lambda_q' < B\mu'$ is sufficient to guarantee that the birth death model of the Markov Chain induced by the certain policies is stable and thus the objective function cannot be infinite.

3.3.2 Markov Decision Process Formulation

The final step before beginning our analysis of the n-period discounted model is to apply uniformization ([61]) to formulate the discrete time equivalent of the discounted problem. To do so let the uniformization factor $\psi = \lambda_e' + \lambda_s' + \lambda_q' + B \cdot \mu'$. Let $\lambda_e = \lambda_e'/\psi$, $\lambda_s = \lambda_s'/\psi$, $\lambda_q = \lambda_q'/\psi$, $\mu = \mu'/\psi$ denote the discrete time parameters after uniformization corresponding to the transition probabilities in the embedded discrete time Markov chain (DTMC). Let α be the continuous time discount factor in the original problem and ξ be an exponential random variable with rate ψ (the length of time for one transition in the discrete chain). The equivalent discrete time

discount factor for the DTMC then becomes:

$$\beta = E[e^{-\alpha\xi}] = \int_0^\infty (e^{-\alpha t})(\psi e^{-\psi t})dt = \frac{\psi}{\psi+\alpha}. \tag{3.3}$$

This implies that the discrete time costs can be defined in terms of the continuous time costs as $h_1 = h_1'/(\psi+\alpha)$,$h_2 = h_2'/(\psi+\alpha)$, and $\tau = \tau'/(\psi+\alpha)$. Since c is assessed per event and not assessed per unit time set $c = c'$. Thus the discrete time instantaneous one stage cost can be written

$$\begin{aligned} C(\mathbf{X}) &= E\bigg[\int_0^\xi \big(h_1'(B-X_1)^+ + \tau'(X_1-B)^+ + \quad h_2'X_2\big)e^{-\alpha t}dt\bigg] \\ &= \frac{1-\beta}{\alpha}\big(h_1'(B-X_1)^+ + \tau'(X_1-B)^+ + h_2'X_2\big) = \frac{h_1'}{\psi+\alpha}(B-X_1)^+ \\ &+ \frac{\tau'}{\psi+\alpha}(X_1-B)^+ + \frac{h_2'}{\psi+\alpha}X_2 = h_1(B-X_1)^+ + \tau(X_1-B)^+ + h_2X_2. \end{aligned} \tag{3.4}$$

Now we can formulate the finite-horizon optimal expected discounted cost recursive optimality equation as:

$$\begin{aligned} V_{n+1,\beta}(\mathbf{x}) = C(\mathbf{x}) + \beta\Bigg\{ & \lambda_e \cdot V_{n,\beta}(\mathbf{x}+e_1) + \lambda_q \cdot \min\big\{V_{n,\beta}(\mathbf{x}+e_1), V_{n,\beta}(\mathbf{x}+e_2))\big\} + \\ & (x\wedge B)\mu \cdot \Big[\mathbb{1}_{\{x_2>0\}} \min\big\{V_{n,\beta}(\mathbf{x}-e_1), V_{n,\beta}(\mathbf{x}-e_2)\big\} + \mathbb{1}_{\{x_2=0\}} V_{n,\beta}\big((\mathbf{x}-e_1)^+\big)\Big] \\ & + \lambda_s \min\big\{V_{n,\beta}(\mathbf{x}+e_1), c/\beta + V_{n,\beta}(\mathbf{x})\big\} + (B-x_1)^+\mu\cdot \\ & \Big[\mathbb{1}_{\{x_1=0,x_2>0\}} \min\big\{V_{n,\beta}(\mathbf{x}), V_{n,\beta}(\mathbf{x}+e_1-e_2)\big\} + \big(1-\mathbb{1}_{\{x_1=0,x_2>0\}}\big)V_{n,\beta}(\mathbf{x})\Big]\Bigg\} \end{aligned} \tag{3.5}$$

Where $V_{n,\beta}(\mathbf{x})$ represents the optimal cost of the n-period β-discounted problem starting in state $\mathbf{x} = (x_1, x_2)$. $\mathbb{1}\{\cdot\}$ is the indicator function, and e_i represents the i^{th} unit vector. In other words, e_i is the vector that contains all zeros except for a 1 in the i^{th} position. The initial condition, $V_{0,\beta} \equiv 0$ is assumed for mathematical convenience and has no effect on the results for the infinite-horizon problem.

The first term of the value function is the one period instantaneous cost described in Eq. (3.4). The second term represents an emergency arrival, which is always

admitted. The third term represents the arrival of an expedited patient. When this event occurs, the controller either calls them in to the hospital and assigns a bed upon arrival (term 1 of the minimization) or places them on the call-in queue (term 2). The fourth term represents the discharge of a patient from the hospital. When a patient is discharged, a bed is freed up, so at this point the hospital can decide to admit an expedited patient and backfill the empty bed (term 2 of the minimization) or to free the bed for scheduled/emergency patients that may arrive in future periods (term 1). Note that, if there are no patients in the call-in queue, no call-in action is available (the second indicator of the discharge term). The fifth term represents a scheduled arrival and the decision to begin treatment and assign a bed to the patient (term 1 of the minimization) or to cancel the patient (term 2). The final term combines the uniformization term (at the end) with a call in decision when $x_1 = 0, x_2 > 0$. That is, an empty system may call in patients at a rate of $B\mu$. While this state has essentially zero probability for a realistic parametrization it is important in the case of $B = 1$ (see Section 3.3.5).

For convenience of analysis, we reformulate this n-period problem using *event-based dynamic programming* operators (see [56]). The cost, arrival, admission control and routing operators are defined as $T_{cost}f(\mathbf{x}) = C(\mathbf{x}) + f(\mathbf{x})$, $T_{A(1)}f(\mathbf{x}) = f(\mathbf{x}+e_1)$, $T_{AC}f(\mathbf{x}) = \min\left\{c+\beta\cdot f(\mathbf{x}), \beta\cdot f(\mathbf{x}+e_1)\right\}$, and $T_{R(\{1,2\})}f(\mathbf{x}) = \min\left\{f(\mathbf{x}+e_1), f(\mathbf{x}+e_2)\right\}$. Let $T_{unif}[f_1, f_2, f_3, f_4](\mathbf{x}) = \lambda_e\beta f_1(\mathbf{x}) + \lambda_q\beta f_2(\mathbf{x}) + \mu\cdot\beta f_3(\mathbf{x}) + \lambda_s f_4(\mathbf{x})$ be the uniformization operator. In addition we define a multi-server backfill operator that encompasses the dynamics of multiple servers as well as the dynamics of backfilling

a discharge with a patient from the call-in queue.

$$T_{MB}f(x) = \begin{cases} (x_1 \wedge B)\min\left\{f(\mathbf{x}-e_1), f(\mathbf{x}-e_2)\right\} + (B - x_1 \wedge B)f(\mathbf{x}) \\ \qquad\qquad \text{if } x_1, x_2 > 0 \\ B\min\left\{f(\mathbf{x}), f(\mathbf{x}+e_1-e_2)\right\} \\ \qquad\qquad \text{if } x_1 = 0, x_2 > 0 \\ (x_1 \wedge B)f\big((\mathbf{x}-e_1)^+\big) + (B - x_1 \wedge B)f(\mathbf{x}) \\ \qquad\qquad \text{if } x_2 = 0 \end{cases} \tag{3.6}$$

Using these operators we can rewrite the value function of Eq. (3.5) as

$$V_{n+1,\beta}(\mathbf{x}) = T_{cost} \cdot T_{unif}\big[T_{A(1)}V_{n,\beta},\ T_{R(\{1,2\})}V_{n,\beta},\ T_{MB}V_{n,\beta},\ T_{AC}V_{n,\beta}\big](\mathbf{x}). \tag{3.7}$$

The following sections develop intuition into the structure of an optimal policy for the above system using the n-period discounted model described in Eq. (3.5). After discussing a few modeling assumptions, we begin our analysis with a simplified case of the general model that disregards a detailed model of expedited call-in patients to isolate the effect of hospital occupancy alone on decision making. This model provides a sense of "what's best" for the hospital. Next we analyze the general formulation to gain insight into a policy that balances the hospital's needs for operational efficiency with the needs of the patients on the call-in queue.

3.3.3 MDP Modeling Assumptions

To reflect practice, we focus on the equilibrium properties of the steady state MDP model, rather than finite horizon effects. To model arrivals, we follow [36], who showed that a Poisson process is a good model both for elective and emergency arrivals to hospital beds. The Poisson assumption for call-in arrivals follows directly from Bernoulli splitting of the Poisson process for emergencies. As in [36], we are not addressing elective admission *scheduling* optimization and so we assume that the

hospital has no control over the elective admission schedule. This is in part to isolate the value of a call-in queue and in part because this is a reality in many hospitals in the U.S. due to decentralized control that allows surgery services to schedule independently of the rest of the hospital. The surgery schedule is constantly changing even up until the last minute, random case times also lead to a nearly Poisson process of elective patients (surgical and medical) requesting a bed. It is important that we are modeling only the flow of elective patients into beds and do not explicitly model the operating room or prior activities.

Generally the arrival rates vary in a non-stationary manner by time of day and day of week. However, since the purpose of the model is to gain insight into the structure of a policy that balances the tradeoffs between utilization and congestion and not to solve the MDP for specific values, making the stationary Poisson assumption for the aggregate flow of patients does not detract from our purpose. We allow for non-stationary arrivals that match historical arrival patterns in a real hospital in Section 3.4, where we analyze a practical policy that harnesses the insight presented in this section.

The length of stay in the hospital is assumed to be exponential for the purposes of tractability. Note that the policies developed in Section 3.4 based on the insight gained from this queueing model and using distributions fitted from historical hospital data still show significant benefits to the hospital despite the modeling simplifications made for the MDP analysis.

3.3.4 Isolating Hospital System Efficiency with Occupancy Based Decision Making

To isolate the effects of occupancy on hospital decision making and analyze in isolation the value of census-based cancelation and call-in admission control, we simplify the modeling of call-in queue patients with the following model changes:

(1) there is always a patient for the hospital to call-in and (2) there is no waiting penalty for call-in patients. While these assumptions lack the realism of a true hospital, they enable us to analyze the hospital's "best case" control options with a simple, insightful model which is later shown to provide reliable insights for complex models.

The value function in Eq. (3.8) represents the uniformized version of this parsimonious model. We proceed by analyzing the n-period discounted problem for this special case of the general model of Section 3.3.1.

$$V_{n+1,\beta}(x) = C(x) + \beta\Big\{\lambda_e V_{n,\beta}(x+1) + \lambda_s \min\left\{V_{n,\beta}(x+1), c/\beta + V_{n,\beta}(x)\right\} + (x \wedge B)\cdot\mu \min\left\{V_{n,\beta}(x-1), V_{n,\beta}(x)\right\} + (B-x)^+\mu \cdot V_{n,\beta}(\mathbf{x})\Big\}. \tag{3.8}$$

Note that the state $x \in \mathbb{Z}^+$ now represents only patients in the hospital, as we are not modeling the length of the call-in queue. $C(x)$ and the event operators are modified accordingly. To simplify the analysis, we do not allow call-ins when the system is empty. This is a good approximation because we are interested in large bed size B, so this state has negligible probability of occurrence.

We show that the value function is convex by demonstrating the closure of the event operators from Section 3.3.2 under convexity, since $V_{n,\beta} = T_1 \cdots T_k V_{0,\beta}$ is just a composition of a combination of the above operators with the initial convex value function V_0.

Theorem III.2. *The value function as defined in Eq. (3.8) is convex.*

Proof. $V_{0,\beta} \equiv 0$ is trivially convex. The closure under convexity of the operators T_{cost}, $T_{A(1)}$, T_{AC}, $T_{R(\{1,2\})}$, and T_{unif} is shown in [56] under the operators of the same name. All that remains is the multiple server backfill operator T_{MB}. To show closure of T_{MB} under convexity we consider four combinations of minimizers a_1 and a_2 of

the left hand side of

$$(x-1)\min\Big\{\underbrace{f(x-2)}_{a_1},\underbrace{f(x-1)}_{a_2}\Big\}+\big(B-(x-1)\big)f(x-1)+$$

$$(x+1)\cdot\min\Big\{\underbrace{f(x)}_{a_1},\underbrace{f(x+1)}_{a_2}\Big\}+\big(B-(x+1)\big)f(x+1)\geq 2x\cdot min\Big\{f(x-1),f(x)\Big\}$$

$$+2\cdot\big(B-x\big)f(x)$$

where a_1 represents the case where the discharge action minimizes the function and a_2 represents the case where the call-in action minimizes the function. We ignore the discount factor, β, because it can be divided out on both sides. To show closure under convexity, it suffices to show closure for every combination a_i, a_j on the left hand side. Case a_1, a_1 reduces to the operator T_{MD} shown to be closed under convexity in [56]. Case a_2, a_2 follows directly from the convexity f and from the properties of minimization. Case a_1, a_2 is easy, because for convex f, $f(x-2)\leq f(x-1)\Rightarrow f(x)\leq f(x+1)$, therefore $\min\{f(x-2),f(x-1)\}=f(x-2)\Rightarrow\min\{f(x),f(x+1)\}=f(x)$. The final case, a_2, a_1, follows from

$$\begin{aligned}
&(x-1)f(x-1)+\big(B-(x-1)\big)f(x-1)+(x+1)f(x)+\big(B-(x+1)\big)f(x+1) &&\geq\\
&\big(B-(x+1)\big)2f(x)+2f(x)+2f(x-1)+(x-1)\big[f(x-1)+f(x)\big] &&=\\
&\big(B-x\big)2f(x)+(x+1)f(x-1)+(x-1)f(x) &&\geq\\
&\big(B-x\big)2f(x)+(x+1)\min\big\{f(x),f(x-1)\big\}+(x-1)\min\big\{f(x),f(x-1)\big\} &&=\\
&2\big(B-x\big)f(x)+2x\min\big\{f(x),f(x-1)\big\}.
\end{aligned}$$

The first inequality follows by combining terms 2 and 4 and using convexity. Next, the equality involves rearranging terms. The following inequality follows from properties of minimization, and the final equality again involves rearranging terms and using properties of minimization. If $x > B$ then T_{MB} reduces to $B\min\{f(x-1),f(x)\}$

which preserves convexity because it is equivalent to T_{AC} where $c = 0$. The case $x = B$ can be shown for cases a_2, a_1 and a_2, a_2 directly using the properties of minimization and convexity. For the boundary, where $x = 1$ the result follows directly from convexity for the only two cases on the LHS: a_2, a_1 and a_2, a_2. Since V_0 is convex and all the operators that make up the t-stage value function, $T_{(.)}$, are closed under convexity, the value function itself is convex for all t. □

Corollary III.3. *The optimal policy for the value function as defined in Eq. (3.8) is (i) of double threshold type for both the cancelation action as well as the call-in action with a call-in threshold, θ^S, and a cancelation threshold, θ^C, and moreover (ii) $\theta^S \leq \theta^C + 1$.*

Proof. (i) The threshold structure of the optimal policy for both actions follows directly from convexity of the value function.

(ii) At θ^S the discharge action is cheaper, i.e., $f(\theta^S) - f(\theta^S - 1) \geq 0$, and the call-in action is cheaper at $\theta^S - 1$, i.e., $f(\theta^S - 1) - f(\theta^S - 2) \leq 0$. At θ^C, the cancelation action is cheaper, i.e., $f(\theta^C + 1) - f(\theta^C) \geq c$, and the admit action is cheaper at $\theta^C - 1$, i.e., $f(\theta^C) - f(\theta^C - 1) \leq c$. From these equations, it is clear that at $\theta^C + 1$, the discharge action will dominate the call-in action since $f(\theta^C + 1) - f(\theta^C) \geq c \geq 0 \Rightarrow f(\theta^C + 1) \geq f(\theta^C)$. Since θ^S is the smallest state at which the discharge action will dominate the call-in action, clearly $\theta^S \leq \theta^C + 1$. □

From Corollary III.3-(i) one can use thresholds θ^S and θ^C to decompose the hospital occupancy into zones based on the optimal admission control actions. Corollary III.3-(ii) suggests that the call-in threshold generally lies below the cancelation threshold (i.e. $\theta^S \leq \theta^C + 1$). When $\theta^S < \theta^C$, the state space can be decomposed specifically into three zones as shown in Figure 3.3: a *call-in* zone, a *steady* zone, and a *cancel* zone. In the call-in zone, the controller admits expedited and scheduled

patients. In the steady zone, the controller admits the scheduled patients but does not call in expedited patients. In the cancel zone, the controller cancels the scheduled patients and does not admit expedited patients. In practice, the solution is elegant and the hospital need only determine what zone the census falls into and take action accordingly.

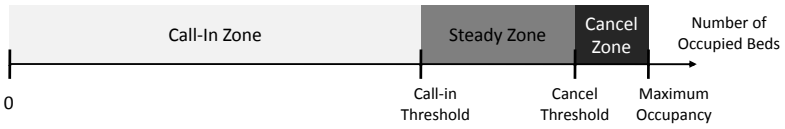


Figure 3.3: Zones for Zone-Based Admission Control versus number of filled beds

3.3.5 Balancing Hospital Efficiency and Call-in Patient Service

In this section, we capture the additional effect of the length of the call-in queue to penalize the waiting time of expedited patients as in the general model formulated in Equation 3.5. Recall the state space is $\mathbf{x} = (x_1, x_2) \in \mathbb{Z}^+ \times \mathbb{Z}^+$, where x_1 is the number of patients in the hospital and x_2 is the number of patients in the call-in queue.

The key result is that we identify properties of the value function in general and again find a double threshold policy to be optimal, except that in this model we have a two-dimensional double threshold policy. This section proves the structure for a special (but non-trivial) case of one bed, and the next section numerically extends the insight for an arbitrary number of beds. That is, there is a call-in and cancelation threshold in both number of occupied beds *and* in number of patients on the call-in queue. To extend the insight about threshold policies from the previous parsimonious model, the following properties are sufficient for a threshold policy in two dimensions to be optimal with respect to the call-in and cancel actions: (1) supermodularity, (2) superconvexity, and (3) convexity in x_1 and x_2. For notational convenience, we denote the class of functions that has these properties as $\mathfrak{F}$.

A supermodular function has the property $f(x+e_1)+f(x+e_2) \leq f(x)+f(x+e_1+e_2)$. A superconvex function has two symmetric properties called (1) *superconvexity 1* defined as $f(x+e_1)+f(x+e_1+e_2) \leq f(x+e_2)+f(x+2e_1)$ and (2) *superconvexity 2* defined as $f(x+e_2)+f(x+e_1+e_2) \leq f(x+e_1)+f(x+2e_2)$ (see [56]).

In general, the interaction between the *call-in action* dynamics and the *multiple server* dynamics disrupts the propagation of the properties of $\mathfrak{F}$. In fact, Section 3.3.6 reveals counterexamples where the $V_{n,\beta} \notin \mathfrak{F}$ even though the threshold structure is still optimal; moreover, the two-dimensional threshold structure holds in all cases of a large random test suite.

Following the form of analysis from Section 3.3.4 we first show that $V_0(\cdot), C(\cdot) \in \mathfrak{F}$. Then we show that if $f \in \mathfrak{F}$ then $T_i f \in \mathfrak{F}$ for all operators. Note that combining supermodularity and superconvexity gives convexity in both components immediately, so we omit the proof of convexity in x_1 and x_2.

Lemma III.4. $V_{0,\beta}, C(\cdot) \in \mathfrak{F}$

Proof. $V_{0,\beta} \equiv 0$ is trivially supermodular and superconvex. The proof of these properties for $C(\cdot)$ is straightforward and can be easily checked by using the definition of $C(\cdot)$ both above, below and at the boundary, B. Convexity in x_1 and x_2 follows directly from supermodularity and superconvexity, thus $V_0, C(\cdot) \in \mathfrak{F}$. □

$T_{cost}, T_{unif}, T_{A(1)}, T_{R(\{1,2\})}, T_{AC}$ are standard operators in [56], whose closure under the properties of the functions in $\mathfrak{F}$ was shown in that paper (as long as $C \in \mathfrak{F}$). It remains to show that the backfill operator, T_{MB}, is also closed under the properties of $\mathfrak{F}$. This is a new queueing operator, though it shares similarities with [29], whose closure under various properties has not yet been shown to the best of our knowledge. The closure of the backfill operator, T_{MB}, under supermodularity and superconvexity

is shown first for $x_1 > 0, x_2 > 0$ and then for the boundary cases $x_1 > 0, x_2 = 0$, $x_1 = 0, x_2 > 0$, and $x_1 = 0, x_2 = 0$.

Lemma III.5. *T_{MB} as defined in Eq. (3.6), is closed under supermodularity for $B = 1$.*

Proof. Suppose $f \in \mathfrak{F}$ and let $x_1 > 0, x_2 > 0$. For supermodularity we need to show the following:

$$T_{MB} \circ f(x) + T_{MB} \circ f(x + e_1 + e_2) \geq T_{MB} \circ f(x + e_1) + T_{MB} \circ f(x + e_2) \ . \tag{3.9}$$

In order to prove Equation (3.9) we need to consider four different cases based on the specific minimizing actions of the left hand side. We can re-write Equation (3.9) as:

$$\begin{aligned} & min\Big\{ \underbrace{f(x - e_1)}_{a_1}, \underbrace{f(x - e_2)}_{a_2} \Big\} + min\Big\{ \underbrace{f(x + e_2)}_{a_1}, \underbrace{f(x + e_1)}_{a_2} \Big\} \\ & \geq min\Big\{ f(x), f(x + e_1 - e_2) \Big\} + min\Big\{ f(x - e_1 + e_2), f(x) \Big\} \ . \end{aligned} \tag{3.10}$$

Where a_1 represents a discharge where the "no backfill" action minimizes the function. In a_2 the "backfill" action minimizes the function. Cases a_1, a_1 and a_2, a_2 follow directly from the supermodularity of $f \in \mathfrak{F}$. Cases a_1, a_2 and a_2, a_1 follow from convexity in x_1 and x_2 respectively.

Boundary Conditions $x_1 > 0, x_2 = 0$: Here we want to show:

$$f(x) + \min \Big\{ f(x - e_1 + e_2), f(x) \Big\} \leq f(x - e_1) + \min \Big\{ \underbrace{f(x + e_2)}_{a_1}, \underbrace{f(x + e_1)}_{a_2} \Big\}$$

For case a_1 the LHS $\leq f(x) + f(x - e_1 + e_2) \leq f(x - e_1) + f(x + e_2)$ and using change of variable $y = x - e_1$ it is easily seen that this equation reduces to supermodularity of f. For case a_2 the LHS $\leq 2f(x) \leq f(x - e_1) + f(x + e_1)$ by convexity in x_1.

Boundary Condition $x_1 = 0, x_2 \geq 0$: The inequality holds by observing that at the boundary $T_{MB}f(x) = T_{MB}f(x + e_1)$ and $T_{MB}f(x + e_2) = T_{MB}f(x + e_1 + e_2)$. $\square$

The following lemma establishes closure of T_{MB} under superconvexity. This result holds for some intuitive conditions on the costs as shown.

Lemma III.6. *If $h_1 \leq \lambda_e \beta \tau$ and $h_2 \geq 0$ then T_{MB} is closed under superconvexity for $B = 1$*

Proof. Suppose f is superconvex. First we need to show that the operator T_{MB} satisfies the first equation of superconvexity :

$$T_{MB} \circ f(x+e_2) + T_{MB} \circ f(x+2e_1) \geq T_{MB} \circ f(x+e_1) + T_{MB} \circ f(x+e_1+e_2) \ . \tag{3.11}$$

We proceed as before, re-writing Equation (3.11) as:

$$\begin{aligned} & min\Big\{ \underbrace{f(x-e_1+e_2)}_{a_1}, \underbrace{f(x)}_{a_2} \Big\} + min\Big\{ \underbrace{f(x+e_1)}_{a_1}, \underbrace{f(x+2e_1-e_2)}_{a_2} \Big\} \\ & \geq min\Big\{ f(x), f(x+e_1-e_2) \Big\} + min\Big\{ f(x+e_2), f(x+e_1) \Big\} \ . \end{aligned} \tag{3.12}$$

Similar to Theorem III.2, we can show that case a_1, a_2 need not be considered. By the superconvexity of f we have $f(x) - f(x-e_1+e_2) \leq f(x+e_1) - f(x+e_2) \leq f(x+2e_1) - f(x+e_1+e_2) \leq f(x+2e_1-e_2) - f(x+e_1)$. The first two inequalities follow from superconvexity (1) and the second follows from superconvexity (2). Thus if $f(x-e_1+e_2) \leq f(x)$, as is implied by a_1, a_2, then $f(x+2e_1-e_2) \geq f(x+e_1)$ so a_1, a_2 can be eliminated. Cases a_1, a_1 and a_2, a_2 follow directly from the superconvexity of $f \in \mathfrak{F}$. Case a_2, a_1 follows directly from the properties of minimization.

The arguments for the proof for superconvexity (2) are completely symmetric to those used to prove superconvexity (1) so the proof is omitted.

Superconvexity (1)-Boundary Conditions: For the first equation of superconvexity, boundary case $x_1 > 0, x_2 = 0$ reduces to

$$f(x) + \min\Big\{ f(x+e_2), f(x+e_1) \Big\} \leq \min\Big\{ \underbrace{f(x-e_1+e_2)}_{a_1}, \underbrace{f(x)}_{a_2} \Big\} + f(x+e_1)$$

For case a_1 the LHS $\leq f(x) + f(x + e_2) \leq f(x - e_1 + e_2) + f(x + e_1)$. Using change of variable $y = x - e_1$ this equation reduces to superconvexity of f. Case a_2 follows directly from the properties of minimization.

Since $T_{MB}f(x + e_2) = T_{MB}f(x + e_1 + e_2)$, the boundary case $x_1 = 0, x_2 > 0$ reduces to

$$\min\Big\{f(x), f(x + e_1 - e_2)\Big\} \leq \min\Big\{f(x + e_1), f(x + 2e_1 - e_2)\Big\}$$

It can easily be shown that if $h_1 \leq \lambda_e\beta\tau$ and $h_2 \geq 0$, then f is increasing in x_1 and therefore we get that $f(x + e_1) \geq f(x) \geq \min\Big\{f(x), f(x + e_1 - e_2)\Big\}$ and likewise $f(x + 2e_1 - e_2) \geq f(x + e_1 - e_2) \geq \min\Big\{f(x), f(x + e_1 - e_2)\Big\}$

Superconvexity (2)-Boundary Conditions: For the second equation of superconvexity, the boundary case $x_1 > 0, x_2 = 0$ reduces to

$$\begin{aligned}&\min\Big\{f(x - e_1 + e_2), f(x)\Big\} + \min\Big\{f(x + e_2), f(x + e_1)\Big\}\\ &\leq f(x) + \min\Big\{\underbrace{f(x - e_1 + 2e_2)}_{a_1}, \underbrace{f(x + e_2)}_{a_2}\Big\},\end{aligned}$$

where in case a_1 the LHS $\leq f(x - e_1 + e_2) + f(x + e_2) \leq f(x) + f(x - e_1 + 2e_2)$. Using change of variable $y = x - e_1$ this equation reduces to superconvexity of f. Again case a_2 follows directly from the properties of minimization. The boundary case $x_1 = 0, x_2 > 0$ reduces to

$$\begin{aligned}2\min\Big\{f(x + e_2), f(x + e_1)\Big\} \leq &\min\Big\{\underbrace{f(x)}_{a_1}, \underbrace{f(x + e_1 - e_2)}_{a_2}\Big\} +\\ &\min\Big\{\underbrace{f(x + 2e_2)}_{a_1}, \underbrace{f(x + e_1 + e_2)}_{a_2}\Big\}.\end{aligned}$$

Case a_1, a_2 follows directly from supermodularity, cases a_1, a_1 and a_2, a_2 follow directly from convexity in x_2. Case a_2, a_1 need not be considered because $f(x) - f(x + e_1 - e_2) \leq f(x + 2e_2) - f(x + e_1 + e_2)$ by using superconvexity (2) twice. Thus if

$f(x+e_1-e_2) \leq f(x)$ then $f(x+e_1+e_2) \leq f(x+2e_2)$, so a_2, a_1 can be eliminated. The final boundary case, $x_1 = 0, x_2 = 0$ reduces to

$$2\min\Big\{f(x+e_1), f(x+e_2)\Big\} \leq f(x) + \min\Big\{\underbrace{f(x+2e_2)}_{a_1}, \underbrace{f(x+e_1+e_2)}_{a_2}\Big\}.$$

Case a_1 follows from supermodularity and case a_2 follows from convexity in x_2. □

Our main result is rigorously proven for the case of only one bed, which is not trivial because, all actions and states are possible and all costs play a role in the optimal policy. The numerical testing in the next section suggests that this result is general.

Theorem III.7. *If $h_1 \leq \lambda_e \beta \tau$ and $h_2 \geq 0$, then for the special case of $B = 1$ server, a threshold policy in x_1 and x_2 is optimal for the backfill action, the cancel action and the call-in action.*

3.3.6 Numerical Results

Since there are cases where $V_t \notin \mathfrak{F}$, we designed a test suite (shown in Table 3.1) composed of several representative cases (Case 1-3) and a randomized test suite (Case 4) to investigate whether the threshold structure holds in a wide range of environments. We check these cases via the numerical computation of the MDP (value iteration) with a large finite state space. From Theorem III.1, we know that the value iteration method will converge to the infinite-horizon optimal average cost value and policy that we are interested in. Case 1 describes a typical community hospital. Cases 2 and 3 consider extremely high and low values for the expedited call-in queue holding cost (h_2) respectively. To test the sensitivity of our model to the arrival process, within each case, we consider subcases (denoted x-1 in the second line of Table 3.1) where 60% of patient arrivals are emergencies and 40% are scheduled and its reverse (subcases x-2) where 40% of arrivals are emergencies and 60% are

scheduled. In the first randomized test suite (Case 4-1), scheduled and emergency arrival rates are generated from a uniform distribution between 0 and 1, then the call-in arrival rate is taken to be a random, $U[0,1]$, percentage of the smaller of the two. This ensures that the call-in arrival rate is not higher than either of the two primary inputs to the hospital. In the second test suite (Case 4-2), this assumption is relaxed and the three rates are all generated from a U[0,1] distribution and then normalized with respect to the random utilization. In Table 3.1, ρ represents the hospital utilization determined by $(\lambda_e + \lambda_s + \lambda_q)/(B\mu)$.

Parma	Case 1		Case 2		Case 3		Case 4	
	1-1	1-2	2-1	2-2	3-1	3-2	4-1	4-2
λ_e	57%	38%	57%	38%	57%	38%	U[0,1]	U[0,1]
λ_s	38%	57%	38%	57%	38%	57%	U[0,1]	U[0,1]
λ_q	5%		5%		5%		$\min(\lambda_s, \lambda_e)$*U[0,1]	U[0,1]
ρ	82%		82%		82%		U[0,1]	U[0,1]
B	160		160		160		U[100,250]	U[100,250]
h_1	1		1		1		U[0,100]	U[0,100]
h_2	1.5		100		0		U[0,100]	U[0,100]
c	34		34		34		U[0,100]	U[0,100]
τ	40		40		40		U[0,100]	U[0,100]

Table 3.1: Parameters used for the test suite.

For each data set given in Table 3.1, we recursively solve the average cost per unit time MDP and record the optimal actions at each state. To illustrate the thresholds, we plot the minimizing action versus number of occupied beds (x_1, the horizontal axis) and number in the call-in queue (x_2, the vertical axis). Figure 3.4 exhibits switching curves as a function of the state for Cases 1, 2 and 3, where Table 3.2 explains the definition of each action.

Action Code	Arr. Call-in Patient	Scheduled Patient	Backfill	Action Code	Arr. Call-in Patient	Scheduled Patient	Backfill
7	admit	admit	yes	3	queue	admit	yes
6	admit	admit	no	2	queue	admit	no
5	admit	cancel	yes	1	queue	cancel	yes
4	admit	cancel	no	0	queue	cancel	no

Table 3.2: Definition of control actions.

Figure 3.4: Optimal actions. The vertical axis represents the number of patients on the call-in queue and the horizontal axis represents the number of patients in the hospital

In Figure 3.4, when hospital occupancy is low, the optimal policy is to call-in patients from the expedited queue in addition to admitting regular scheduled patients (actions 6 and 7). At higher occupancies, the optimal policy is to admit the scheduled patients but hold off the new call-in arrivals by placing them on the queue (actions 2 and 3). Finally at critically high occupancy levels, the optimal policy seeks to avoid congestion by canceling scheduled patients and placing new call-in arrivals on the queue (actions 0 and 1). Note that as the emergency arrival rate increases, the control system starts canceling scheduled patients at lower occupancy levels (i.e. the black area in Figure 3.4 expands to the left). Furthermore, when the holding cost of a patient on the call-in queue increases, the system gives more attention to new call-in arrivals, cancels scheduled patients and admits new call-in arrivals (actions 4 and 5) at medium-high occupancy levels. On the other hand, when the holding

cost of patients on the call-in list is zero (Case 3), scheduled patients are given more attention and action 2 becomes optimal at medium-high occupancy levels.

The action plots of the optimal policies in Figure 3.4 all clearly show that a threshold policy is optimal in both x_1 and x_2 for both the cancelation and call-in actions. Even though in some cases $V_t \notin \mathfrak{F}$, in all 2,000 random problem instances (1,000 each from Cases 4-1 and 4-2), the threshold policy for both actions in both dimensions is optimal. Thus, a threshold structure policy is at the very least optimal for a broad class of parameters.

3.4 Simulation Study of a Partner Hospital

The MDP model discussed in the previous sections provides intuition into the double threshold properties of a hospital admission control system. This section develops a practical admission control mechanism based on the zone-based admission control policy introduced in Section 3.3, but specifically modeling 3 bed units. We demonstrate the benefits of such a policy with a custom-designed C++ simulation study based on historical data from a partner hospital. We use the patient flow simulation framework described in [39], so the features are explained at a high level to outline our approach. Figure 3.5 presents an abstraction of the simulation patient flow model.

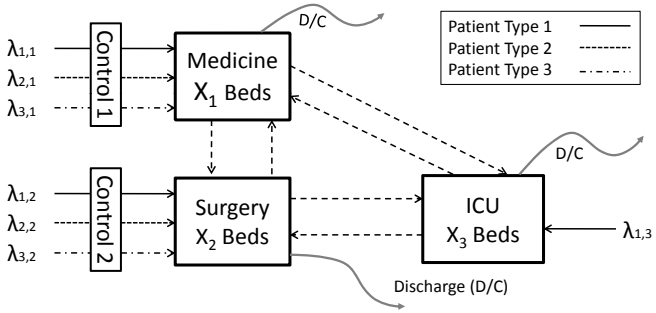


Figure 3.5: Abstract Hospital Patient Flow Simulation Model.

The primary building block of our patient flow simulation model is the set of

units, U, each having multiple beds. We allow for an arbitrary number of patient types, n, with patients of type $y \in \{1, \ldots, n\}$ arriving exogenously to these units according to a non-stationary Poisson Process with rate $\lambda_{y,d,t}$ where d represents the day of the week, t represents time of day, and $y \in \{1, \ldots, n\}$. After completing their initial segment of treatment, patients are transferred between units u_i and u_j with probability $p^y_{u_i,u_j}$ or are discharged from the hospital with probability $1 - \sum_{k \in U} p^y_{u_i,u_k}$. The empirical length of stay distributions are based on patient type. If the unit to which a patient seeks admission is full, the patient is placed on a non-preferred unit that has the capability to care for the patient (e.g. a surgical patient placed on a medicine unit), as is common practice in most hospitals.

The patient flow simulation framework allows for custom controls to be designed and attached to framework elements. For our admission control study, we attach the zone-based admission control module to the arrival streams as shown in Figure 3.5. Our model's purpose allows us to exclude detailed modeling of the ED and OR, allowing us to focus on the admission control and bed occupancy dynamics.

In order to verify and validate the accuracy of our simulation model, we use strategies suggested by [80]. Verification proceeded via a series of white-box and black-box testing schemes. First we verified the correct operation of each of the components shown in Figure 3.5. Using detailed patient location/transition output that we generated for each unit for every turn of the simulation clock, we are able to verify that the correct number of patients are flowing through the system on a daily basis.

We validate our model of the system by comparing it against actual "real-world" hospital operations. That is, given a year's worth of hospital admissions data, we are able to extract the weekly scheduling pattern for elective medical and surgical

patients and subsequently implement this policy in our model. Comparing the key features of the system (average daily census by day of week, volume of emergency and scheduled patients, and so forth) as in [63] we find that our simulation closely matches the actual hospital operations.

Our simulated hospital is a medium sized non-teaching hospital with approximately 50% emergency and 50% elective volume. There are three primary interacting units: surgery, medicine and ICU. The majority of the hospital's patients can be divided into four surgical types and three medical types. Arrival rates, length of stay distributions and transfer probabilities are determined from one year of historical data.

In early discussions, several systemic problems were identified in this hospital so this study focuses primarily on mitigating these issues: (1) ED crowding, (2) excessive surgical cancelations, (3) difficulty admitting medicine patients in the middle of the week and (4) too many patients placed *off* the preferred unit for their condition (e.g. a surgical patient on a medicine unit).

3.4.1 A Zone-Based Admission Control Mechanism and Call-in Queue Operation

Based on the analysis in Section 3.3, we propose a zone-based admission control mechanism in which the hospital state is divided into three zones (cancel, steady and call-in) as in Figure 3.3. However, the model of our partner hospital assigns each bed unit its own zone and also allowed the zones to vary by day of week to add refinement. The thresholds from the MDP model in Section 3.3.5 gave a rough estimate for initial zone control values, which were refined using a heuristic simulation-based search. A single run of the simulation for 100 weeks takes less than a minute on a regular office PC. The simulation-based neighborhood search to find the admission zones takes approximately a half an hour, depending on the quality of the initial policy

generated by the MDP. While the MDP model is simplified, it is very useful for approximating the more detailed control policy in the simulation.

An important practical feature of the call-in queue is that it should admit the patients on the queue within an advertised amount of time, which we set to 3 days in our simulation. In order to uphold this guarantee, we add a feature to the call-in queue that monitors each patient's length of time on the queue. When a patient reaches their maximum length of stay on the queue, their priority is upgraded and they are automatically admitted into the hospital; we call this a "force-in". Because this mechanism limits the maximum waiting time on the call-in queue, our zone-based admission control can focus purely on the occupancy level. Of course a significant number of call-in queue force-ins in practice signals that the system needs to be revised.

For practicality of implementation, we activate the control once per day (11 AM for example) to allow enough time to call in patients from the call-in queue or cancel surgeries scheduled for later that day. This occupancy evaluation should include, if possible, projected discharges, surgeries scheduled and in progress, emergency patients boarding in the ED and waiting for a bed and possibly other factors. At each review point, the hospital evaluates the predicted occupancy and applies the control policy; calling in patients up to the call-in threshold and canceling if census is above the cancelation threshold.

The expedited patients we identified are currently part of our partner hospital's emergency patient stream. The arrivals to the emergency department are modeled as a Poisson Process with rate λ, with a proportion, q, of those arrivals being expedited patients. We can use Bernoulli splitting of the emergency Poisson Process to generate two independent Poisson Processes with rate $(1-q) * \lambda$ for true emergency arrivals

and $q * \lambda$ for call-in queue arrivals. This splitting models the reality that a certain portion of ED arrivals occur because of the lack of an expedited mechanism for hospital admission within a short time frame.

3.4.2 Simulation Study Analysis

Using the current admission policies, Figure 3.6 shows one simulation sample path to visually demonstrate the effect of the expedited queue and zone-based admissions compared the sample path of the current admissions system.

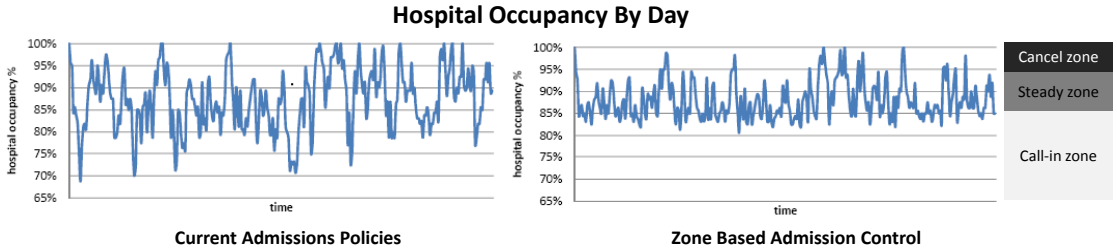


Figure 3.6: Simulation results - comparing current system with zone-based admission control.

Next we quantify the benefits of zone-based admission control in terms of the following key metrics: cancelations, emergency blockages, and off-unit census. Figure 3.7-(a) compares one year of operation under the current admission policies (Current State) with the same system using zone-based admission control (Admission Controlled). It conservatively models (based on previous literature) only 5% of the current emergency population as eligible for the expedited call-in queue.

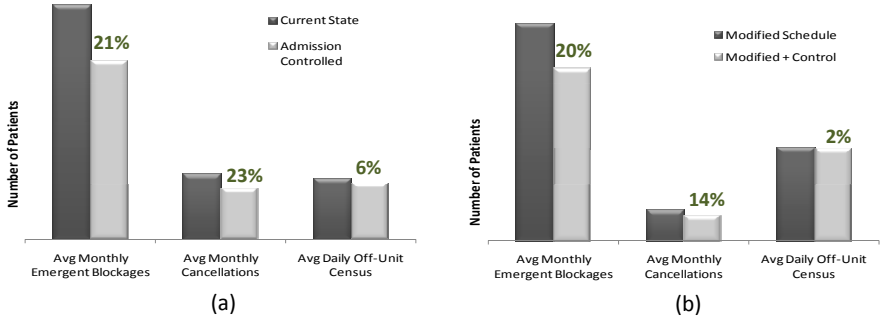


Figure 3.7: Simulation results: comparing the key hospital metrics for (a) current system vs. zone-based admission control and (b) a system with improved elective schedule with and without admission control.

Figure 3.7 clearly demonstrates that admission control can improve all three hospital metrics simultaneously. The most significant decreases are seen in cancelations and emergency blockages. The majority of cancelations and blockages occur in the middle of the week when the hospital is already full from a surgery schedule that emphasizes high volumes on Monday and Tuesday and few on Friday (which is a common practice). By placing call-in eligible candidates on a call-in queue during this occupancy spike we can avoid both emergency blockages and cancelations.

In general, an environment with a period of high volume scheduled electives followed by a period of low volume scheduled electives allows the expedited queue to improve the system in several ways. The expedited call-in queue delays a portion of patients who would otherwise exacerbate the ED crowding. This reduces the volume of non-electives and better accommodates the period of high elective volume, which in turn reduces the chance of emergency diversion and elective cancelation. Then as the occupancy starts to drop during the low volume period, expedited patients can be called in to bolster and smooth the occupancy level; with a host of benefits outlined in the introduction.

To further reduce the emergency blockages, the cancelation threshold can provide a small buffer on days where emergency blockages are frequent. This type of cancelation would probably be used sparingly (the cancelation zone is small) due to the lost revenue and organizational pain associated with canceling elective admissions, but even a small buffer properly placed provides significant relief for emergency blockages. Furthermore, it is likely that many hospitals serve elective patients at the expense of beds for emergency patients, given the prevalence of ambulance diversions [71], so our approach helps to rationalize this system and protect emergency patients.

There are many possible zone schemes that provide different proportions in the

tradeoff between the three metrics. We chose the current scheme to provide benefit to all metrics and attempted to achieve fairly equal reductions in cancelations and emergency blockages. By manipulating the zone scheme one can achieve other reduction proportions in accordance with the desires of the partner hospital.

These results suggest that zone-based admission control can effectively improve hospital *system level metrics*; however, one must also consider the implications for *service* to the expedited patients. Prior to introducing the call-in queue, expedited patients using the emergency department as a work around to gain admission were bed blocked at a rate of 11%. During blockage these patients experienced excessive waiting in the ED, frequently being boarded in the hall, and the added stress of spending long periods of time in an overcrowded emergency department. After adding the call-in queue, the blockage rate for these patients dropped almost to zero. From a qualitative perspective, call-in patients are able to wait in the comfort of their own home, get some of the necessary tests *before* admission rather than in the ED. This potentially decreases their length of stay, care costs, and they have a bed *ready* for them when they arrived to the hospital. Hence the call-in queue not only improves system performance, it also offers a valuable functionality to the expedited patients.

3.4.3 Sensitivity Analysis

In the previous section we assumed that only 5% of emergency patients would be eligible for the call-in queue and the patient mix was 50% emergency and 50% scheduled. Under the heading "Call-in Volume", Table 3.3 compares the improvement over the current system for the case where 5%, 10%, and 15% of the emergency patients are eligible for the expedited call-in queue. The 3 columns on the right labeled "Emergency Volume" test the system by varying the mix between emergency and

scheduled patients (while keeping the original 5% call-in fraction of the emergency load fixed). The values are in terms of percent reduction versus the current system in the three key hospital metrics we consider in this section. The system improves in all metrics as the volume of call-ins increases and the hospital gains more control over its random arrivals, but the system, as expected, exhibits decreasing marginal returns as the percent of call-in eligible patients increases. As might be expected, in the right three data columns of Table 3.3, when there is a higher volume of emergency patients, the zone-based control mechanism provides greater relative benefits than lower emergency volumes.

% Reduction vs Current	Call-in Volume			Emergency Volume w/ call-ins at 5% fraction		
	5%	10%	15%	40%	50%	60%
Cancelation	22%	42%	55%	13%	22%	25%
Emergency Blockage	17%	28%	45%	13%	17%	25%
Off-Unit Census	6%	7%	8%	6%	6%	2%

Table 3.3: Sensitivity analysis of call-in queue and emergency volume.

In this chapter we have emphasized reactive admission control and not scheduling. One may conjecture that reactive control would not be necessary if elective schedules are properly managed. On the contrary, our analysis in Figure 3.7-(b) shows that there are still significant benefits to using a zone-based admission control mechanism even when the elective admission schedule is controlled to improve performance. The improved elective admission schedule we consider still has no elective admissions on the weekends but attempts to level the average daily census (or workload) across the weekdays to accommodate the prevailing practice. Note that the dark gray in Figure 3.7-(b) represents the improved schedule without admission control, while the light gray represents the same improved schedule but using zone-based admission control as well.

Figure 3.7 demonstrates that zone-based admission control improves efficiency of

health care delivery both in a hospital that does not control its own surgical schedule *and* in a hospital that has worked to improve its scheduling to increase operational efficiency. In fact, the sensitivity analysis of Table 3.3 shows that, though the degree of efficacy varies, in all the cases considered zone-based admission significantly benefits multiple aspects of hospital care delivery.

3.5 Conclusion and Future Research

By analyzing the structure of an admission control system, we have proven the optimality of a two threshold policy in a stylized model based only on census, and a special case of a more general model based on census and length of the call-in queue. Though the MDP modeling in this chapter considers a hospital with only one type of bed, the threshold concept can be extended to include multiple bed types as we have done in the simulation study of Section 3.4 with very good results. Using simulation based on historical hospital data, we showed that a practical zone-based admission control policy can simultaneously reduce the number of emergency patients blocked, the number of canceled elective patients and the amount of off-unit census. In all of our test cases, a zone-based admission control policy significantly improved key hospital performance metrics compared to current practice.

For future work, one may wish to consider if using more than three zones is worth the additional complexity. For example, one can further differentiate expedited patients into different queues based on severity of their condition. One can also consider different cancelation levels for different classes of scheduled patients.

For a hospital systems solution to be implementable and sustainable, the overhead and complexity should be minimized. The simple structure of a threshold/zone-based system fits this paradigm nicely, because it is easily understood and implemented. Analysis of historical hospital data and hospital dyanamics/protocols can determine

the threshold values, so the admissions department only needs to monitor the census and take action. This requires a centralized admissions decision making body and census information. If the underlying dynamics of the hospital system change significantly, the zones will need to be adjusted; for example a change in patient population, adding a new hospital wing, hiring a new surgeon or holiday patterns.

CHAPTER IV

Fast-tracking Priority Customers through Queueing Networks with an Application to Destination Hospitals

The previous two chapters developed theoretical methodology and practical approaches for the planning and control of patient flow through networks of wards in hospitals. This chapter extends this core patient flow theory to networks of specialist services in a broader context. We motivate the work by the networks of specialist outpatient services provided at a destination hospital, but the approach can be generalized to general networks of care services. One organizational structure that could benefit beyond the destination setting is the increasingly popular model of accountable care/bundled payments for treatment that are found in the new Accountable Care Organization (ACO) construct.

This chapter focuses on prioritizing fast-track patients by assigning them priority scheduling slots in a destination hospital with a large network of specialist services. At a destination hospital, patients travel long distances to receive treatment. An important service metric for these national and international patients is that they complete their treatment segment before the weekend, when clinical service shut down causing non-value added patient waiting. We call this is called *itinerary completion*. The *itinerary completion* concept motivates this research, which addresses a larger class of problems in healthcare and other applications concerning the fast-

track scheduling of priority customers receiving services within a queueing network with hard deadlines for service completion.

The approach decomposes the problem into two consecutive stages: (1) workload smoothing, and (2) itinerary completion. The first stage begins with a queueing network blocking model that is built from a stochastic arrival-location model of demand for services across the network of services. Blocking measures are linearized and the analytical queueing network model is transformed into a deterministic linear program. The result is a smoothed workload that reduces system-wide congestion that can cause patient appointments to be delayed. The second stage develops a phase-type model of itinerary completion that is parameterized by the blocking probabilities computed in stage one and also transformed and solved via linear programming. Our methodology for the two stage control of queueing networks was tested on a case study for breast cancer patients, increasing itinerary completion from 74% to 88%.

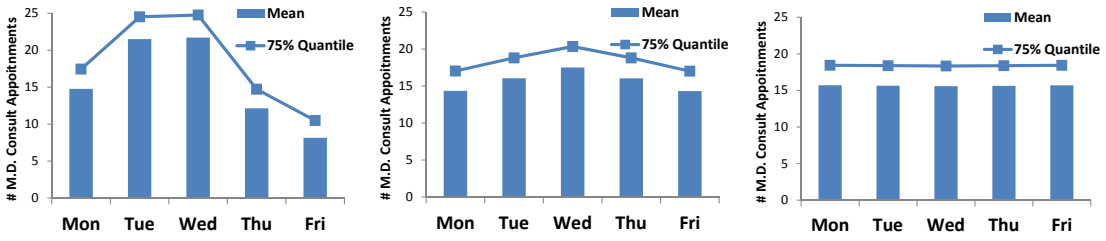
4.1 Introduction

Effective and efficient delivery of healthcare is a major societal concern in the US and throughout the world and the paradigm of coordinated care delivery has gained increasing visibility in recent years. Because the entire healthcare system in the US (and in other countries as well) is coming under increasing cost pressure to increase efficiency, there is also a financial incentive to transition to larger more coordinated network care models such as those found at the Mayo Clinic.

This chapter develops innovative stochastic modeling methods to forecast the workload across the network of outpatient care resources for any given patient schedule. This stochastic location model serves as the mechanism for linking admission decisions to outpatient service workload. Combining the stochastic location model with

a controlled arrival stream yields what we call a controlled-arrival-location model (CALM). In this chapter, we develop the CALM model in such a way as to allow for optimization methods that can determine improved patient schedules in a tractable manner.

Though large providers enjoy many benefits, there are also great operational and management challenges of coordinating the complex network of resources that comprise these healthcare organizations. A primary controllable driver of efficiency in these complex systems is patient scheduling. The way in which patients are scheduled has a major effect on the workload experienced in an organization's various network resources. As an example, consider the workload in the breast diagnostic clinic for three different schedules that generate three different workload profiles. The output based on historical data from the Mayo Clinic is shown in Figure 4.1.



(a) Unbalanced (b) Partially unbalanced (c) Completely Balanced

Figure 4.1: The effect of the patient schedule on the workload at the breast diagnostic clinic.

The schedule that generates an unbalanced workload profile (Fig 4.1 (a)) requires over 33% more resources on average in the middle of the week than the schedule that generates a balanced schedule. This causes the organization to adopt an unbalanced staffing profile (which requires more resources in aggregate and staffing inconvenience), or to accept access failures and treatment bottlenecks. The management challenge is in designing a schedule that can achieve a balanced workload such as the one in Figure 4.1 (c). Because patient care pathways are complex and

dynamically changing with their disease condition and the amount of information providers have about their disease it is often difficult to predict the effect scheduling various patients. Another complicating factor is that patients follow a stochastic path through a general network of care services. For example, breast cancer patients at the Mayo Clinic as a population require care from over 77 different outpatient services over the course of their treatment.

After developing the general analytical models for controlling patient schedules and smoothing hospital workloads, we demonstrate how this approach can be applied to a critical problem in destination healthcare organizations in general and of particular interest to our partner, the Mayo Clinic: *Itinerary Completion*. By definition, many patients of destination healthcare organizations come from geographically distant locations. These patients are classified as national or international patients. With these patient types an important access metric is that they complete their treatment segment before the weekend. Since most clinical services are not available on the weekend, failure to complete care within the work week forces the patients to pay for hotels and stay over the weekend without any treatment progress, which is emotionally challenging for patients and families at a very vulnerable time of their life. Itinerary completion is an important part of patient satisfaction and poor healthcare delivery performance; in fact, some patients decide to return home without completing treatment. The itinerary completion metric was developed to measure the ability of the healthcare provider to complete treatment for national/international patients treatment before the weekend begins.

The core model is divided into two stages as shown in Fig. 4.2. The goal of the workload smoothing stage is to reduce blocking in all of the important services by stabilizing the workload across the week. Because the clinics are staffed at a constant

level over the days of the week, eliminating the workload spike should give patients better access to care in the middle of the week, avoiding delays to getting an appointment and thereby flowing unhindered through their treatment path. The itinerary completion stage establishes the virtual fast-track priority schedule to maximize the probability that national/international patients will complete their treatment within the work week. In Fig. 4.2, for example, the national patients have been mostly moved to the beginning of the week.

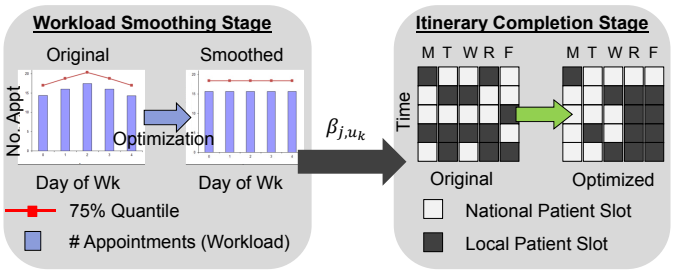


Figure 4.2: High level approach to a two stage fast-track model.

We demonstrate the applicability of our patient flow paradigm to the itinerary completion problem by developing schedules that fast-track national/international patients; however, the approach is generalizable to any healthcare organization in which patients with different needs / priority levels arrive according to a controllable patient schedule and flow through a network of care services. This applies not only to destination healthcare organizations, but to all healthcare organizations that manage multiple types of care services.

In Section 4.2, we provide some background on the different approaches to managing scheduled arrivals to stabilize flow and workloads, primarily in healthcare systems. In Sec. 4.3 we developed the queueing network model of patient flow through a network of services. Sec. 4.4 builds upon the stochastic models of Sec. 4.3 by developing queueing network blocking models and optimization methods for smoothing

workloads across a network. Sec. 4.5 presents a phase-type modeling approach for capturing itinerary completion of fast-track patients flowing through a congested queueing network with blocking. Sec. 4.6 incorporates the flow models from Sec. 4.3 and 4.4, and the phase-type itinerary completion model from Sec. 4.5 into a second schedule optimization model that fast-tracks priority patients. We demonstrate how the two optimization models can be used to solve the Mayo Clinic itinerary completion problem with a detailed case study in Sec. 4.7. Finally, we draw conclusions and discuss the contribution of this work in Sec. 4.8.

4.2 Patient Flow Management and Optimization in Highly Stochastic Systems

While the goal of this work is to manage the flow of patients through outpatient clinical services, our work differs significantly from the classical outpatient scheduling literature in terms of scope (single clinic vs network) and timing (single day vs multiple day planning horizon). Traditional outpatient scheduling literature typically focuses on the scheduling of a single clinic, which is often modeled as some kind of queueing system (see the survey paper [9]). Our problem focuses on a heterogeneous network of specialist care services, which we model in the form of a general *queueing network*. Traditional outpatient scheduling approaches focus on scheduling patients to time slots within a day with respect to various individual characteristics (e.g. no-shows, doctor availability etc.; see [9]). Our problem focuses on patients that have multiple visits to different services across days and even weeks and our goal is to smooth and stabilize that aggregate number of appointments at each service for each day of the week. We believe this chapter presents a novel approach that expands and deepens the outpatient scheduling literature by analyzing the burgeoning care delivery paradigm of coordinated care in the context of a network of outpatient care

services.

Perhaps surprisingly, the healthcare literature that has much more in common with the outpatient service network problem is the hospital inpatient scheduling literature. Much of the work in patient flow modeling and the control of patient schedules at the aggregate level has been developed for hospitals, rather than outpatient service networks. Early stochastic models of hospital census include simulation [26, 33, 32, 60] and probabilistic approaches [12]. The early work relied on heuristic schedule improvement, but were effective in characterizing the hospital census levels for any given admission schedule. Due to the complexity of the general stochastic network of resources required to serve hospital patients, simulation has continued to be a preferred modeling approach for hospital occupancy (see for example [35, 40]). One recent approach uses a genetic algorithm to identify improved schedules, but the method was shown to be computationally burdensome even for a small hospital (see [42]).

While simulation is highly effective in a descriptive context, tractable simulation-based optimization for such a large scale problem appears out of reach. Our modeling paradigm shares more commonality with the literature surrounding analytical models of hospital census. A recent interest in elective patient admission schedule optimization has led to a number of papers that attempt to linearize and solve the problem of improving hospital census focused metrics by designing better admission schedules. [21, 22] developed the early analytical models of hospital census for incorporation into a mixed integer programming framework. Recently, others have expanded upon this approach to solve problems for a variety of objectives (see [11, 1, 4]). These models, however, focus on a single ward / resource or a feed-forward network and lack the generality required to model the network of care services provided by a large

coordinated care provider.

To model the outpatient service network, this work builds upon the literature on stochastic arrival-location models developed for applications in the telecommunications industry (see [65, 66, 59]) and recently for hospital inpatient admission schedule optimization (see [46]). We extend the analytical work done in this area in two important ways. First, the aforementioned models are designed for capturing aggregate level offered load statistics. In our paradigm, both aggregate level metrics (e.g. patient access service level) and patient path specific metrics (e.g. itinerary completion) are important. Thus in our approach we add a customer (patient) focused layer that interacts with the aggregate level modeling output. This allows us to capture something that has not previously been captured in this context and yet is incredibly important for service oriented healthcare organizations: the individual customer experience. Secondly, at the aggregate level we allow for "batch" appointments, where a patient can visit more than one outpatient service in a given day, which leads to new arrival-location models and novel analysis of stochastic arrival-location model systems.

The modeling contributions of this work include (1) the development of new blocking models for the customer sojourn times in queueing networks for which service time is not continuous, but is characterized by the number of shifts/days until service can occur, (2) analytical models for congested queueing networks that are amenable to schedule optimization at the aggregate level and the individual customer level, (3) new linearizing approximations that capture blocking in queueing networks in a manner that can be incorporated into a linear programming optimization framework.

In the next section, we develop and analyze our analytical framework for statistically characterizing patient flow trajectories and using them to build a workload

profile for the entire network of clinical services. We call this model in the naming convention of [65] and [46], the outpatient controlled-arrival-location model (CALM).

The impact of this work in the area of effective and efficient healthcare delivery extends beyond destination hospitals; though we use the case of a destination hospital as part of the motivation and demonstration of the technique. In the context of the current healthcare climate, where reimbursement is moving toward bundled payment for treatment and the newly formed Accountable Care Organizations seek to provide a network of outpatient and inpatient services to manage all aspects of patients' treatment, this approach of capturing and planning for workloads across a disparate network of services will have the opportunity for broader impact.

4.3 Outpatient Controlled-arrival-location Model (CALM)

We model the flow of patients through an outpatient service network using the offered load approach. A simplified diagram of breast cancer patient flow for 5 key breast cancer services (out of a total of 77 services that breast cancer patients may potentially use) is given in Figure 4.3. New patients arrive to the healthcare provider to begin a treatment segment according to an initial set of scheduled appointments. A new patient is defined as either a patient that is new to the care provider or an existing patient that is returning to the provider because of a new diagnosis. A treatment segment is defined as a contiguous segment of care in which the last appointment is separated from a subsequent appointment by at least one business week.

After patients arrive for a new treatment segment, they flow through the network of services being scheduled for appointments according to the dynamically evolving information about their condition. As more tests come back and more specialists

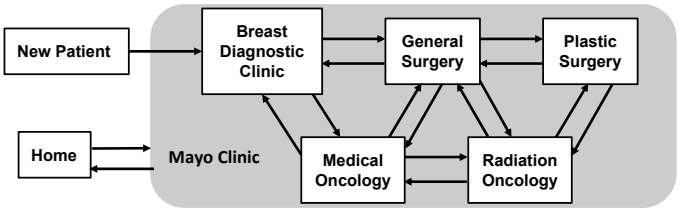


Figure 4.3: Simplified example of offered load flow model for breast cancer patients.

are consulted, new appointments are made to address the improved understanding of the patient's condition. At destination healthcare organizations, this path is not often known in advance. For example, a new breast cancer patient will typically arrive and receive an appointment at the breast diagnostic clinic. After the series of examinations at the breast diagnostic clinic, further appointments at other specialist service are scheduled. Once the patient's treatment segment is complete the patient will return home, but will then have the possibility of revisiting the healthcare provider for follow-up treatment or new treatment. In modeling breast cancer patients, for example, we consider care pathways of up to a year or longer. This is important because admitting a new patient doesn't just affect clinic load for their first treatment segment, but for their entire treatment path.

In this section, we develop an offered load model of patient flow to capture these system dynamics and statistically characterize the load on each of the clinical services over time. We build the aggregate workload model by first building a model of the care pathways for each type of patient. This model provides the probabilities on whether or not a given service is required at a given time after their initial appointment. Patient type can be general, but for our purposes we let patient type be related to specific patient diagnoses. We combine the care pathways for each patient type with the admission stream for that type to forecast the system's steady state offered load.

4.3.1 System Design and Modeling Assumptions

The system presented here functions by allocating a certain number of appointment slots of each type to each service by day of week. As a modeling assumption we assume these slots will always be filled, which creates a deterministic arrival stream. This is a reasonable assumption for high demand healthcare organizations.

In terms of assigning slots, we also consider a repeating weekly schedule. This is for the purposes of application rather than a modeling limitation. In most clinical services, weekly repeating schedules are preferable so that other physician activities, such as research, can be scheduled with regularity as well. This means that we are modeling the workload by day of week as a cyclostationary system.

Another mild assumption we make is that patient care paths are independent of one another. Clearly the condition and service requirements of one patient shouldn't affect the condition and requirements of another. Further, because we are using an offered load model we have the infinite server assumption and so patients don't block one another or otherwise interact. To capture blocking effects we later superimpose capacities on our offered load model – similar to the modified offered load approach proposed by [67]. In this way we avoid the significant non-linearities associated with blocking models while still capturing blocking system effects.

Finally, in applying service level constraints, we use a normal approximation on the number of appointments requested on a given day so that we only need to calculate the mean and variance analytically. This approximation can be justified by the central limit theorem, but also also been validated in the literature for similar healthcare applications (see [12, 48]).

4.3.2 Developing the Out-PATTERN Stochastic Location Process

In this section we develop an analytical stochastic location process to model patient flow through the network of clinical services. The stochastic location process forms the basis of the stochastic-arrival-location model of clinical service workload. Since the location model refers to the clinical needs of a single patient over time, we call the model the Outpatient Temporal Resource Requirements (Out-PATTERN) stochastic location process, consistent with the naming convention of [46].

Assume there are M clinical services provided. At the Mayo Clinic M is quite large, as breast cancer patients alone seek treatment over a range of 77 clinical services. These services span a wide range of specialties and typically have dedicated staff, given that most of the services provided are specialist services. Let the vector state space for the Out-PATTERN stochastic location process for clinical services be $\mathcal{S}^0 = \{[x_1, x_2, \ldots, x_M] : x_i \in \mathbb{Z}^+ \; \forall i\}$. This is a deviation from the stochastic location models employed in the research mentioned previously, where the state space was a scalar representing the location of the patient. We let the full state space be $\mathcal{S} = \mathcal{S}^0 \bigcup \{\Delta\}$, where Δ represents that the patient has no appointments (e.g. has returned home or has not yet become a patient).

In the outpatient setting there are two major differences that require this novel formulation. First, a patient can have more than one appointment at a given clinical service in a day. Though we allow each x_i to take values in $\mathbb{Z}^+$, typically the number of appointments at a given service on a given day will be few and tightly bounded. Second, the state is a vector rather than scalar because we allow for the fact that a patient could have appointments at multiple clinical services on a single day.

The state space could, of course, be simplified to the same state space used in previous literature by decreasing the discrete time step from a day to hours or min-

utes. However, prescribing outpatient schedules to the hour or minute does not leave enough flexibility for the organization, the physicians and other staff, or the patient thereby creating barriers to implementation. We instead focus on providing aggregate flow guidelines and targets that allow the organization and individuals significant flexibility in meeting those guidelines.

After the arrival of a new patient, the Out-PATTERN stochastic location process is an $\mathcal{S}$-valued function, $L_{s,k}(t)$, where s is the day that the patient started treatment and k is the patient type. To define the stochastic process, $L_{s,k}(t)$, we first define the space of outcomes. Let Σ_s be the set of right-continuous functions with left limits that map $\mathbb{R}$ to $\mathcal{S}$ such that $\sigma_s \in \Sigma_s$ has the property that $\sigma_s(t) = \Delta \ \forall t < s$ and $\sigma_s(s) \in \mathcal{S}^0$. Thus, Σ_s contains the care paths of all patients who start new treatment at the clinic at time s.

The function space Σ is the collection of all Σ_s. Now we need to define a probability measure on Σ that is associated with the stochastic location process, $\mathbb{P}_s : \Sigma \to [0, 1]$, that assigns 0 probability to Σ_t for $t \neq s$ and for $\Gamma \subseteq \Sigma_s$, it assigns the probability associated with the set of location functions Γ. As an example, consider the set $\Gamma_t^{u,2} = \{\sigma : \sigma \in \Sigma_s \text{ for } s < t, \mathbf{e}_u \cdot \sigma(t) = 2\}$ where $\mathbf{e}_u$ is the unit vector with all 0's and a 1 in the u^{th} column. In words, $\Gamma_t^{u,2}$ is the set of all location processes in which the patient requires two appointments in clinical service u at time t. Therefore $\mathbb{P}_s(\Gamma_t^{u,2})$ is the probability that a patient who initiates a new treatment at time s requires two appointments in clinical service u at time t. For notational convenience we let

$$\mathbb{P}(L_{s,k}(t) \cdot \mathbf{e}_u = m) = p_{s,k,u}(m, t - s), \tag{4.1}$$

where $p_{s,k,u}(\cdot, \cdot)$ is calculated based on historical data in a method similar to [38].

Remark IV.1. There is an important technical detail regarding the calculation of the stochastic location process probabilities from historical data. We propose to use the stochastic location process in an offered load model with infinite capacity, yet historical patient flow data are clearly capacitated. To eliminate this endogeneity, we take advantage of the fact that demand for healthcare services typically follows a seasonal pattern, which is confirmed by the data from the Mayo Clinic. Thus to calculate the true stochastic location process probabilities, we can simply use data from low utilization months during which time patients rarely have difficulty getting an appointment – thereby approximating uncapacitated flow.

Remark IV.2. It should be noted that the more information a healthcare organization has in advance of a patient arrival, the less uncertainty there is in regards to the patients path. In the case of breast cancer, for example, knowing information about the specifics of the patient's disease upfront can help the clinic plan ahead for what specialists need to be seen. While spending the effort to collect this information prior to the patient's arrival may have significant benefits in reducing workload variability and improving system performance, this avenue is left to future research.

4.3.3 The Out-PATTERN d-CALM Clinical Service Workload Process

To characterize the workload in terms of number of appointments by day of the planning horizon, we combine the Out-PATTERN stochastic location process from Section 4.3.2 with a controlled deterministic arrival stream to develop the Out-PATTERN deterministic controlled-arrival-location model (d-CALM).

Recall that $L_{s,k}$ is the stochastic location process for a patient of type k who had their initial treatment at time s, where $k \in \mathfrak{D}$, which is the set of all patient types. For our application, we consider a repeating planning horizon of 5 days to match the Mayo Clinic's work week during which almost all of the outpatient activity occurs.

For any given appointment slot schedule, Θ, the Out-PATTERN d-CALM model characterizes the demand for clinical service u on day d_1 of week t, D^t_{u,d_1}. The d-CALM model can be formulated as a point process as in [46], however we prefer an equivalent approach by writing the demand for clinical services as a sum of indicators:

$$D^t_{u,d_1}(\Theta) = \sum_{d_2=1}^{5} \sum_{k \in \mathfrak{D}} \sum_{j=1}^{\Theta_{k,d_2}} \sum_{n=0}^{t} \mathbf{e}_u \cdot L^{j,n}_{d_2+5n,k}(d_1 + 5t) \tag{4.2}$$

$$D^{\infty}_{u,d_1} = \lim_{t \to \infty} D^t_{u,d_1}. \tag{4.3}$$

Eq. 4.2 represents a t-week horizon and Eq. 4.3 represents an infinite horizon cyclostationary steady state model. $L^{j,n}_{s,k}(\cdot)$ is the $(j,n)^{th}$ i.i.d instance of the location process $L_{s,k}(\cdot)$, representing a patient who was the j^{th} patient of that type that was admitted on week n. The first sum in Eq. 4.2 is over the days of the planning horizon. The second sum is over all patient types. The third sum is over the number of patients of type k that were admitted on day d_2 of the planning horizon. The final sum is over weeks from week 0 to week t.

This concludes the presentation of the fundamentals of the Out-PATTERN d-CALM model. In the next section we analyze this model in an attempt to obtain an offered load characterization that is linear in the decision variable Θ, as the goal is to optimize the appointment slot allocation across the week to improve patient flow. The approach we take is to calculate the moments of the Out-PATTERN d-CALM process in Section 4.3.4 and use them in a normal approximation of demand for clinical services that can be incorporated into an optimization model to minimize appointment blocking in Section 4.5.

4.3.4 Moments of the Outpatient d-CALM Process

The key to the d-CALM approach is that the moments of the process can be calculated analytically (and linearly), which elminates the need to simulate this complex system. Let M_u be the maximum number of visits to resource u in a given day. Then the mean workload in service u on day d_1 follows from the monotone convergence theorem:

$$\begin{aligned}\mu_{d_1,r}(\Theta) &= \mathbf{E}\Big[\sum_{d_2=1}^{5}\sum_{k\in\mathfrak{D}}\sum_{j=1}^{\Theta_{k,d_2}}\lim_{t\to\infty}\sum_{n=0}^{t}\mathbf{e}_r\cdot L^{j,n}_{d_2+5n,k}(d_1+5t)\Big] \\ &= \sum_{d_2=1}^{5}\sum_{k\in\mathfrak{D}}\Theta_{k,d_2}\cdot\sum_{n=0}^{\infty}\sum_{m=1}^{M_r} m\cdot p_{d_2+5n,k,r}(m,d_1-d_2+5(t-n)). \qquad (4.4)\end{aligned}$$

Eq. 4.4 is linear in the decision variable Θ. The following theorem shows that the variance is also linear in the decision variable. For notational convenience let $\hat{p}^{n,k,m,t}_{d_1,d_2,u} = p_{d_2+5n,k,u}(m,d_1-d_2+5(t-n))$.

Theorem IV.3. *The variance in number of appointments requested (offered load) for resource u on day d_1 is given by*

(4.5)

$$\sigma^2_{d_1,u}(\Theta) = \sum_{d_2=1}^{5}\sum_{k\in\mathfrak{D}}\Theta_{k,d_2}\lim_{t\to\infty}\sum_{n=0}^{t}\sum_{m=1}^{M_u}\left[m^2\cdot\hat{p}^{n,k,m,t}_{d_1,d_2,u}\Big(1-\hat{p}^{n,k,m,t}_{d_1,d_2,u}\Big)-\sum_{q>m}2m\cdot q\cdot\hat{p}^{n,k,m,t}_{d_1,d_2,u}\cdot\hat{p}^{n,k,q,t}_{d_1,d_2,u}\right]$$

Proof. Each pair of location functions, $\big(L^{j_1,n_1}_{s_1,k_1}(t_1), L^{j_2,n_2}_{s_2,k_2}(t_2)\big)$ where $(j_1,n_1,k_1,s_1)\neq(j_2,n_2,k_2,s_2)$ represents two different patients. Therefore the two location processes are independent, which follows from our assumption that the care

paths of two different patients are independent. To see that each pairing in Eq. 4.2 does indeed represent two different patients, note that each patient's stochastic process is uniquely indexed by the patient type, k, the week in which they are admitted, n, the day of the week they were admitted, s, and their admission number on the day they are admitted, j. Therefore the the processes $L^{j_1,n_1}_{s_1,k_1}(t_1)$ and $L^{j_2,n_2}_{s_2,k_2}(t_2)$ are independent and their covariance term is necessarily zero. The variance of the ward census can be calculated by

$$\begin{aligned}
\sigma^2_{d_1,u}(\Theta) &= \mathbf{Var}\left[\sum_{d_2=1}^{5}\sum_{k\in\mathfrak{D}}\sum_{j=1}^{\Theta_{k,d_2}}\lim_{t\to\infty}\sum_{n=0}^{t}\mathbf{e}_r\cdot L^{j,n}_{d_2+5n,k}(d_1+5t)\right] \\
&= \sum_{d_2=1}^{5}\sum_{k\in\mathfrak{D}}\sum_{j=1}^{\Theta_{k,d_2}}\lim_{t\to\infty}\sum_{n=0}^{t}\mathbf{Var}\left[\mathbf{e}_r\cdot L^{j,n}_{d_2+5n,k}(d_1+5t)\right] \\
&= \sum_{d_2=1}^{5}\sum_{k\in\mathfrak{D}}\Theta_{k,d_2}\lim_{t\to\infty}\sum_{n=0}^{t}\sum_{m=1}^{M_u}\Bigg[m^2\cdot\hat{p}^{n,k,m,t}_{d_1,d_2,u}\Big(1-\hat{p}^{n,k,m,t}_{d_1,d_2,u}\Big)- \qquad (4.6)\\
&\qquad \sum_{q>m}2m\cdot q\cdot\hat{p}^{n,k,m,t}_{d_1,d_2,u}\cdot\hat{p}^{n,k,q,t}_{d_1,d_2,u}\Bigg]. \qquad (4.7)
\end{aligned}$$

The first equality follows from the monotone convergence theorem and from the independence of the stochastic location processes within the sum, allowing us to take the variance on the inside of the sum. The second equality follows from the following variance calculation for the Out-PATTERN stochastic location process:

$$\begin{aligned}
\mathbf{Var}\left[\mathbf{e}_r \cdot L_{s,k}(t)\right] &= \mathbf{E}\left[(\mathbf{e}_r \cdot L_{s,k}(t))^2\right] - \mathbf{E}\left[\mathbf{e}_r \cdot L_{s,k,u}(t)\right]^2 \\
&= \sum_{m=0}^{M_u} m^2 p_{s,k,u}(m, t-s) - \left(\sum_{m=0}^{M_u} m p_{s,k,u}(m, t-s)\right)^2 \\
&= \sum_{m=0}^{M_u} m^2 p_{s,k,u}(m, t-s) - \sum_{m=0}^{M_u} m^2 p_{s,k,u}(m, t-s)^2 \\
&\quad - \sum_{m=0}^{M_u} \sum_{n>m} 2mn \cdot p_{s,k,u}(m, t-s) \cdot p_{s,k,u}(n, t-s).
\end{aligned}$$

The final equality of the variance calculation follows by applying the multinomial formula.

□

In Sec. 4.4 we develop the first stage optimization that smooths the workload relative to capacity across the planning horizon incorporating the analytical flow models from this section.

4.4 Workload Smoothing Optimization

In the first stage of the optimization we consider smoothing the number of appointments (workload) over all clinical services across the week. The objective is to minimize the probability that a patient will not be able to get an appointment at any given care service within the network. To accomplish this, we develop analytical methods for calculating the blocking probability at each service based on the d-CALM offered load model from Section 4.3. The development and linearization of the blocking calculation are presented in Section 4.4.1. In that section we present a linear approximation of the square root function to calculate the standard deviation from the linear variance calculation. An example incorporating the analytical models and blocking metrics into a linear program to solve the workload smoothing problem

is presented in Section 4.4.2.

4.4.1 Blocking Calculations for Optimization Models

The blocking probability is typically difficult to compute and non-linear in the decision variable for d-CALM models. Thus such an approach is usually not suitable for incorporation into tractable optimization methods. To overcome this, we approximate the workload in each clinical resource by a Normal random variable with mean and variance parameters given by Eq's 4.4 and 4.6. The Normal workload approximation has been used effectively in many healthcare settings (see for example [48, 12]).

In service systems with Normally distributed workloads, service constraints can be defined in terms of mean μ, standard deviation σ, and the standard normal service-level factor n. Let n_α be the multiplier that matches the desired service level α and C be system capacity. Then the constraint $\mu + n_\alpha \cdot \sigma \leq C$ ensures that the system will encounter access block with probability no greater than $1-\alpha$. Since μ and σ are both a function of the patient schedule decision variable, Θ, to minimize the probability of an appointment delay one might consider $\max_{\Theta, n_\alpha} \{n_\alpha : \ \mu(\Theta) + n_\alpha \cdot \sigma(\Theta) = C\}$; however, this equation is not convex, which makes optimization difficult.

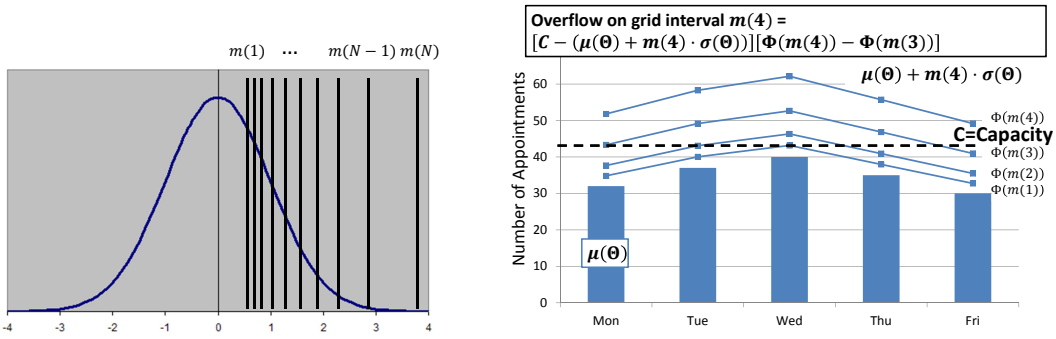
Instead, we propose the following approach to approximating the program just described. To remove the non-convex multiplication of n_α and $\sigma(\Theta)$, replace the single constraint, $\mu(\Theta) + n_\alpha \cdot \sigma(\Theta) = C$, by a set of constraints that discretize possible values of n_α over a grid. Doing so we can either calculate the expected overflow or the probability of overflow. We focus on the expected overflow and then show how the same technique can be used to approximate the probability of overflow. This approach eliminates n_α as a decision variable.

To calculate the expected overflow, let $\mathcal{M}' = \{1, 2, \ldots, N\}$ be an index that

creates a discrete grid with N sections. The grid need not (and in application is not chosen to be) linear. Thus we have a one-to-one mapping function $m(i) : \mathcal{M}' \rightarrow \mathcal{M}$ that maps the integer values of $\mathcal{M}'$ to the grid values $\mathcal{M}$. An example of a grid mapping for the grid $\mathcal{M} = \{0, 0.1, 0.2, 0.4, 0.6, 0.9, 1.2, 1.5, 1.8, 2.2, 2.6, 3.1\}$ is given in Table 4.1 and the Fig. 4.4 graphically depicts the approach.

$i \in \mathcal{M}'$	1	2	3	4	5	6	7	8	9	10	11	12
$m(i) \in \mathcal{M}$	0.0	0.1	0.2	0.4	0.6	0.9	1.2	1.5	1.8	2.2	2.6	3.1

Table 4.1: Sample grid mapping from the integers to the grid $\mathcal{M} = \{0, 0.1, 0.2, 0.4, 0.6, 0.9, 1.2, 1.5, 1.8, 2.2, 2.6, 3.1\}$.



(a) Grid on normal distribution

(b) Overflow approximation using grid

Figure 4.4: Example of a discrete grid that approximates the Riemann integral for the expected overflow. Solid bar = mean appointments, line = m(i) std. dev. above the mean

Fig. 4.4 (a) shows one type of grid over the standard normal, with $m(i)$ representing the i^{th} grid interval. In designing the grid, one approach that has nice intuition and behaves well in practice is to make the grid such that each interval contains an equal amount of probability density according to the standard normal. Fig. 4.4 (b) shows how the grid is used to approximate expected overflow. On the $m(i)$ interval, the overflow is calculated at the upper value of the overflow over the interval, $\mu(\Theta) + m(i)\sigma(\Theta)$. This overflow value is multiplied by the amount of density on the interval, $\Phi(m(i)) - \Phi(m(i-1))$. In the limit this approximation will converge to the exact overflow value.

All that remains is to design a set of constraints in the linear program that will set the decision variable, $\delta_{u,d,i}$, to the proper amount of overflow. The following constraint set, when combined with an objective function that minimizes expected overflow will achieve this result.

$$\mu_{u,d}(\Theta) + m(i) \cdot \sigma_{u,d}(\Theta) - C_{u,d} \leq \delta_{u,d,i} \text{ for } i \in \mathcal{M}' = \{1, 2, 3, \ldots, N\}. \tag{4.8}$$

If the desired metric is to minimize the blocking probability, the same set of constraints may be used except that $\delta_{u,d,i}$ is required to be a binary decision variable and is multiplied by a large constant value (e.g. Q) to ensure feasibility. Note that the constant Q need not be too large, since the probability that the demand will exceed a given level drops off quite sharply away from the mean. Thus very large values of demand can be ignored and Q can be determined directly from the structure of the grid.

Using the above approximation, we remove the first non-linearity of multiplying two decision variables together; however, the problem remains that $\sigma(\Theta)$ is still non-convex in the decision variable Θ. From the d-CALM model's Eq. 4.6 the *variance*, $\sigma^2(\Theta)$, of the workload can be calculated as a linear function of the decision variable Θ. To reconcile the issue of taking a square root of the variance with the tractability of the optimization model, we propose to approximate the square root of $\sigma^2(\Theta)$ using Newton's method. Letting $\hat{\sigma}(\Theta)$ be an initial guess for the standard deviation, Newton's method gives the approximation

$$\sqrt{\sigma^2(\Theta)} \approx \frac{1}{2}\left(\frac{\sigma^2(\Theta)}{\hat{\sigma}(\Theta)} + \hat{\sigma}(\Theta)\right). \tag{4.9}$$

Thus we can approximated the true workload standard deviation for any schedule Θ. If $\hat{\sigma}$ is chosen so that it is close to the actual standard deviation, then this

approximation can be quite accurate. One appropriate $\hat{\sigma}$ would be the standard deviation of the historical workload of the current system. Table 4.2 demonstrates the quality the approximation considering two different types of schedules: (1) *Worst Case:* The entire patient load is scheduled on the first day of the week, (2) *Likely Case:* The workload is divided evenly across all days the week and thus each day has the same level of workload. The data presented in Table 4.2 represented the workload from breast cancer patients for the top 4 clinical services that are used by breast cancer patients. The arrival rate, stochastic location processes, and workload are calculated based on historical data, creating a training set for model parametrization and a test set for validation.

DOW	Case	Med. Oncology		Diagnostic Clinic		Radiation Oncol.		Gen. Surg.	
		True	Approx	True	Approx	True	Approx	True	Approx
All	Likely	5.50	5.50	4.04	4.04	3.55	3.55	2.29	2.29
Mon	Worst	5.59	5.59	5.95	6.47	3.50	3.50	2.30	2.30
Tue	Worst	5.44	5.44	3.68	3.70	3.46	3.46	2.20	2.20
Wed	Worst	5.37	5.38	3.44	3.50	3.57	3.57	2.36	2.36
Thu	Worst	5.56	5.56	3.20	3.29	3.59	3.59	2.32	2.32
Fri	Worst	5.51	5.51	3.25	3.31	3.61	3.61	2.30	2.30

Table 4.2: Comparison of True standard deviation with Newton's method-based approximation for a likely case and the worst case by day of week

Clearly case (1) will produce the largest deviations in variance from the current system; the variance will be much higher than the current system on Monday and the other days of the week will have much lower variances. This represents the case in which the approximation should perform the worst, as the starting values will be farthest away from the actual value. Even in this unrealistically bad scenario, the worst error occurs in the diagnostic clinic and is an error of less than 9%. The next largest error is less than 3%. Outside of the diagnostic clinic, there is essentially no error to two decimal places of accuracy. For case (2) there is no error at all between the approximation and the actual to two decimal places. For this application, our

Newton's method-based approximation is extremely accurate.

It is important to note that the extremely unbalanced scenarios such as the one presented in case (1) will almost certainly not be optimal and so the error in the approximation is less important in these cases where the error is larger. Case (2) is a schedule that will produce good performance results and thus the fact that the approximation errors are negligible for this type of schedule indicates that this approximation is a good choice for linearizing the square root function.

4.4.2 Workload Smoothing Optimization Model

We begin this section with the notation required to develop the linear program. We then present one particular formulation of the outpatient workload smoothing optimization that stabilizes the workload (number of appointments at each clinical service) to improve system level access for outpatients. In this formulation time is discretized into days and we consider a planning horizon of $1, \ldots, 5$ to correspond to a weekly (business week) schedule. This is one example of how the above modeling can be applied to solve major patient flow problems, but certainly is not an exhaustive exposition of the potential for incorporating d-CALM stochastic models into optimization frameworks.

Sets

$\mathfrak{D}$ set of all patient diagnosis types

$\mathcal{U}$ set of clinical services

$\mathcal{M}'$ index for the grid for calculating service level

Parameters

$C_{u,d}$ service u capacity on day d

$\theta_{k,d}$ current admission volumes of type k on day d.

$\hat{\theta}_{k,d}$ maximum number of admissions of type k allowed on day d.

$\hat{\sigma}_{u,d}(\Theta)$ The standard deviation of workload under the current system.

$w_{u,d}$ the weight assigned to expected overflow in service u on day d

Decision Variables

$\Theta_{k,d}$ number of type $k \in \mathfrak{D}$ patients scheduled on day d

$\delta_{u,d,i}$ amount of workload overflow in service u on day d at service-level factor $m(i)$

Placeholder Variables for Exposition

$\mu_{u,d}(\Theta)$ the mean number of resources of type u required on day d under schedule Θ, given by Eq. 4.4

$\sigma^2_{u,d}(\Theta)$ the variance of the number of resources of type u required on day d under schedule Θ given by Eq. 4.6

For clarity, we substitute placeholder variables $\mu_{u,d}(\Theta)$ and $\sigma^2_{u,d}(\Theta)$ for the full linear equation, but we have argued previously that the equations are linear in the decision variable Θ. The probabilities associated with the Out-PATTERN stochastic location processes in Eq.'s 4.4 and 4.6 are calculated off-line and entered into the linear program as data.

(4.10)

$$\min_{\Theta,\delta_{j,k,\ell}} \sum_{u\in\mathcal{U}} \sum_{d=1}^{5} w_{u,d} \sum_{i\in\mathcal{M}'} [\Phi(m(i+1)) - \Phi(m(i))]\delta_{u,d,i}$$

s.t.

(4.11)

$$\mu_{u,d}(\Theta) + m(i) \cdot \frac{1}{2}\left(\frac{\sigma^2_{u,d}(\Theta)}{\hat{\sigma}_{u,d}(\Theta)} + \hat{\sigma}_{u,d}(\Theta)\right) - C_{u,d} \leq \delta_{u,d,i} \quad \forall u \in \mathcal{U}, i \in \mathcal{M}', d = 1,\ldots,5$$

(4.12)

$$\sum_{d=1}^{5} \Theta_{k,d} = \sum_{d=1}^{5} \theta_{k,d} \qquad \forall k \in \mathfrak{D}$$

(4.13)

$$\Theta_{k,d} \leq \hat{\theta}_{k,d} \qquad \forall k \in \mathfrak{D}, d = 1,\ldots,5$$

$$\Theta_{k,d} \in \mathbb{Z}^+, \delta_{j,k,\ell} \geq 0 \ \forall j,k,\ell$$

The objective function, Eq. 4.10, will drive the system to minimize the weighted approximation of the expected overflows across the week over all the clinical services. In fact, $\sum_{i\in\mathcal{M}'}[\Phi(m(i+1)) - \Phi(m(i))]\delta_{u,d,i}$ is a simple Riemann integral approximation, which we know converges to the true expected value as the intervals go to zero. Weights, $w_{u,d}$, are included for several reasons, the most obvious being that a blockage in one service may be more critical than a blockage in another; however, they are all equal to one in the case study we present later.

From constraint Eq. 4.11 it is clear that the smaller the $\delta_{u,d,m}$ variables are, the higher the probability that there will be enough capacity to serve an arriving patient. Thus minimizing the sum of the $\delta_{u,d,m}$'s achieves the goal of reducing appointment block. The details behind this discretization of the service-level factor by $m : \mathcal{M}' \rightarrow \mathcal{M}$ were provided in Sec. 4.4.1.

Eq. 4.11 is the constraint that calculates the service-level factor, for any given schedule Θ. If the workload level associated with $m(i)$ is below capacity, the variable $\delta_{u,d,m}$ can be set to 0. Using the standard normal lookup table it is possible to translate $m(i)$ directly into a blocking probability. Here we approximate the standard deviation with the highly accurate Newton's method-based approach detailed in Section 4.4.1. Eq. 4.12 constrains the optimal weekly volume to be equal to the current weekly volume. Eq. 4.13 ensures that an upper bound on the number of patients of each type arriving for treatment on a given day is respected.

4.5 Analytical Models for Itinerary Completion

In Section 4.3 we developed the stochastic models to capture the aggregate level system behavior. In this section, we develop and analyze the stochastic models to support the patient centered path-based layer of our paradigm. For purposes of exposition, this section focuses on our particular application, itinerary completion, though this modeling paradigm is generalizable to capture other objectives involving outpatients with differing needs and priorities.

Itinerary completion is defined as a patient completing a contiguous treatment path within one business week. The reason for this definition is tied to the particular dynamics of a destination healthcare organization. In such an organization, many patients come from a long distance. Nationally or internationally originating patients who can't complete their medical itinerary by Friday must stay over the weekend since most outpatient services don't operate on the weekends. This is a significant concern for the high level of patient satisfaction required for a sustainable destination healthcare service. Itinerary completion is impacted by both the level of congestion (through the ability to obtain an appointment in a timely manner) and the day of the week the patient is admitted. The congestion level was addressed in Sec. 4.4.

The day of week component is combined with the congestion modeling from Sec. 4.4 in this and the following section.

The following sections build toward a flow model for each patient's path through the congested network. By analyzing this model we are able to capture the patient's sojourn time from treatment initiation to completion. Importantly, we are able to capture this sojourn time linearly in the decision variable of when to schedule each type of patient.

4.5.1 Phase-Type Model for Critical Path Flow Times

To calculate the treatment length, including delay caused by blocking, for one contiguous segment of treatment that begins on day d, it is necessary to break the treatment into appointments along the patient's critical path. These are appointments that, if delayed, will delay the entire treatment cycle. Details of how to determine the critical path are discussed in Sec. 4.7.1.

The critical path is defined by the tuple $\mathcal{C} = (\mathcal{R}, \mathcal{P})$, where $\mathcal{R}$ are the services that lie along the critical path and $\mathcal{P}$ are the precedence relations. For example, $\mathcal{R} = \{u_1, u_2, \ldots, u_n\}$ and $\mathcal{P} = \{(u_i, c) : u_i \in \mathcal{R}, c = 1, 2, \ldots, N\}$, where the tuple (u_i, c) represents the scenario where an appointment at service u_i is required in the c^{th} step (order of precedence) of treatment. That is if $a < b$, then (u_i, b) can only occur after (u_j, a) have been completed for all j.

In the following three subsections we build up to a model to allow multiple appointments that can be done in parallel at each precedence level, but for the initial analysis we assume each precedence level has only one appointment. Let $\mathbf{B} = [\beta_{u_i,d}]$ be the matrix of blocking probabilities by clinical service (u_i) by day of week (d), which is calculated from the workload smoothing optimization (see Eq. 4.57 of Sec. 4.6). The total time to complete the treatment segment is denoted by $\delta_{\mathcal{C},d}(\mathbf{B})$, which

captures the sojourn time through a congested network with blocking matrix $\mathbf{B}$. $\delta_{\mathcal{C},d}(\mathbf{B})$ is comprised of $\delta_{u_i,d}(\mathbf{B})$, which is the time to complete the i^{th} appointment along the critical path, given that the patient began requesting an appointment on day d. Let K be the deadline for completing the itinerary. Then the probability a patient will not complete their itinerary within the planning horizon given they were admitted on day $d \in \{0, 1, \ldots, K\}$ of the horizon is:

$$\mathbb{P}(\delta_{\mathcal{C},d}(\mathbf{B}) > K - d) \tag{4.14}$$

4.5.2 Phase-type Base Model

In this section, a base model is developed in which all items along the critical path must be completed. The next two sections incorporate the possibility that not all patients need appointments at all of the services along the critical path and each stage may have multiple tasks. To calculate the CDF for the sojourn time $\delta_{\mathcal{C},d}(\mathbf{B})$ for the base model, we exploit the blocking probabilities computed from the workload smoothing stage. On each day d the patient is either able to get the appointment at service u (w.p. $1 - \beta_{u,d}$) or is blocked from getting the appointment (w.p. $\beta_{u,d}$). The time to complete an appointment at service u_i given an initial attempt to schedule on day d, $\delta_{u_i,d}(\mathbf{B})$, therefore follows a discrete phase-type distribution characterized by a Markov Chain with a phase transition probability matrix

$$\left[\begin{array}{c|c} \mathbf{T_u} & \mathbf{T_u^0} \\ \hline \mathbf{0} & 1 \end{array}\right] = \left[\begin{array}{ccccc|c} 0 & \beta_{1,u} & 0 & 0 & 0 & 1-\beta_{1,u} \\ 0 & 0 & \beta_{2,u} & 0 & 0 & 1-\beta_{2,u} \\ 0 & 0 & 0 & \beta_{3,u} & 0 & 1-\beta_{3,u} \\ 0 & 0 & 0 & 0 & \beta_{4,u} & 1-\beta_{4,u} \\ \beta_{5,u} & 0 & 0 & 0 & 0 & 1-\beta_{5,u} \\ \hline 0 & 0 & 0 & 0 & 0 & 1 \end{array}\right] \tag{4.15}$$

where $\mathbf{T_u} \cdot \mathbf{1} + \mathbf{T_u^0} = \mathbf{1}$. In particular, $\delta_{u_i,d}(\mathbf{B})$ is the time until absorption in the discrete time Markov Chain defined by the Eq. 4.15. The Markov chain, and therefore the discrete phase-type distribution, are completely characterized by $\mathbf{T_u}$, which is called the *generator matrix* of the phase-type distribution. This is because $\mathbf{T_u}$ represents the transitions between the transient states of the chain and $\mathbf{T_u^0}$ can be calculated by subtracting the row sum of $\mathbf{T_u}$ from 1 for each row of $\mathbf{T_u}$. It is well known (see [58]) that the CDF and pmf of $\delta_{u_i,d}(\mathbf{B})$ are given for integral x number of stages/days by

$$F(x) = 1 - \mathbf{e}_d(\mathbf{T}_u)^x \cdot \mathbf{1} \tag{4.16}$$

$$f(x) = \mathbf{e}_d(\mathbf{T}_u)^{x-1} \cdot \mathbf{T_u^0}, \tag{4.17}$$

where $\mathbf{e}_d$ is taken to be the unit column vector with 1 in the d^{th} position and 0 elsewhere. The intuition behind Eq. 4.16 is the following. From Markov chain theory, the probability that the chain has not left the transient states by time x is $\mathbf{e}_d(\mathbf{T}_u)^x \cdot \mathbf{1}$, where $\mathbf{e}_d$ is the state that the chain starts in and multiplying by $\mathbf{1}$ sums up the probabilities of being in each of the transient states at time x. The probability that the time to absorption is smaller than x is therefore one minus the probability

that the chain is still in the transient state at time x. The intuition behind Eq. 4.17 follows from the same line of reasoning.

The total delay for a patient is the sum of the correlated delays for each service along the critical path. Fortunately, it can be shown that this sum also has a phase-type distribution. The transition probability matrix that characterizes the phase-type distribution for total length of the critical path is given by the block diagonal matrix

$$\left[\begin{array}{c|c} \mathbf{T}_{\mathcal{C}} & \mathbf{T}^{\mathbf{0}}_{\mathcal{C}} \\ \hline \mathbf{0} & 1 \end{array}\right] = \left[\begin{array}{cccccc|c} \mathbf{T}^{\mathbf{1}}_{\mathbf{u_1}} & \mathbf{T}^{\mathbf{2}}_{\mathbf{u_1}} & \mathbf{0} & \mathbf{0} & \dots & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{T}^{\mathbf{1}}_{\mathbf{u_2}} & \mathbf{T}^{\mathbf{2}}_{\mathbf{u_2}} & \mathbf{0} & \dots & \mathbf{0} & \mathbf{0} \\ & \ddots & \ddots & & & & \vdots \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \dots & \mathbf{T}^{\mathbf{1}}_{\mathbf{u_n}} & \mathbf{T}^{\mathbf{0}}_{\mathbf{u_n}} \\ \hline \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \dots & \mathbf{0} & 1 \end{array}\right], \tag{4.18}$$

where $\mathbf{T}^{\mathbf{1}}_{\mathbf{u_i}}$ and $\mathbf{T}^{\mathbf{0}}_{\mathbf{u_n}}$ are defined as in Eq. 4.15. Additionally, Let

$$\mathbf{T}^{\mathbf{2}}_{\mathbf{u_i}} = \begin{bmatrix} 1-\beta_{1,u} & 0 & 0 & 0 & 0 \\ 0 & 1-\beta_{2,u} & 0 & 0 & 0 \\ 0 & 0 & 1-\beta_{3,u} & 0 & 0 \\ 0 & 0 & 0 & 1-\beta_{4,u} & 0 \\ 0 & 0 & 0 & 0 & 1-\beta_{5,u} \end{bmatrix}. \tag{4.19}$$

To see that this matrix does indeed represent the desired transition probability matrix of a Markov Chain whose time to absorption represents the length of a patient's care path, consider the phase transition diagram (Fig. 4.5) of the chain described by Eq. 4.18.

From this diagram it is clear that when a patient starts task 1 of their critical path on day d, they are either blocked from receiving an appointment (w.p. $\beta_{u_1,d}$) or

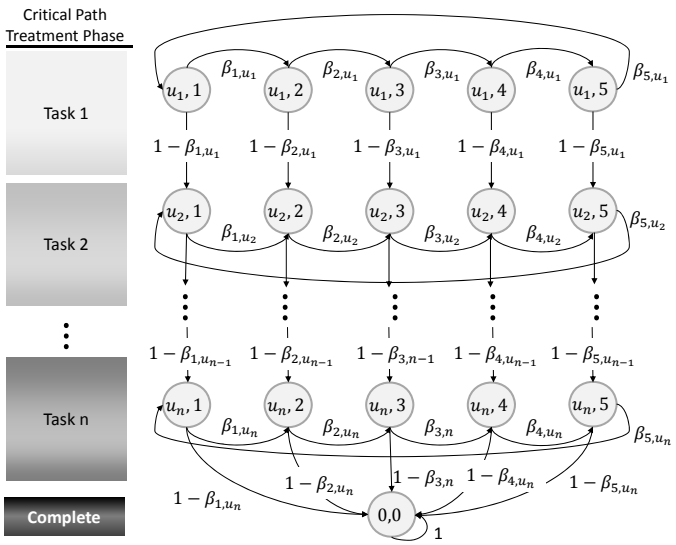


Figure 4.5: State transition diagram of a Markov Chain whose time to absorption represents the length of a patient's care path. The state is a tuple (task, day), where u_i is the task and the days are from $1, \ldots, 5$, representing weekdays

they complete their task (w.p. $1-\beta_{u_1,d}$). If the patient is blocked, the Markov Chain transitions to the next day, $(d+1) \mod 5$, but remains within task 1. If the patient is not blocked, the patient transitions to task 2. When the patient is on the final task of their critical path and they are not blocked, then the patient transitions to the absorbing state, $(0,0)$, which indicates that the treatment segment is completed. Thus the time to absorption is the time it would take for a patient to complete their critical path.

Hence, the time for a patient to complete their critical path given that they were initially scheduled to begin treatment on day d can be described by a discrete time phase-type distribution with pmf and CDF given by Eq.'s 4.20 and 4.21.

$$F(x) = 1 - \mathbf{e}_d \left(\mathbf{T}_{\mathcal{C}}\right)^x \cdot \mathbf{1} \tag{4.20}$$

$$f(x) = \mathbf{e}_d \left(\mathbf{T}_{\mathcal{C}}\right)^{x-1} \cdot \mathbf{T}_{\mathcal{C}}^{\mathbf{0}}. \tag{4.21}$$

Remark IV.4. With the phase-type structure it is also possible to capture precedence

relationships that require task j to begin no earlier than $n = 1, \ldots, N$ days after task i is completed (where N is the length of the planning horizon). This is useful if, for example, task i is a medical test that will require n days to process. This delayed precedence can be captured by modifying the matrix that transitions to task j, inputting mandatory transitions w.p. 1 for the required waiting time until the patient can transition to j. This holds for all future sections as well, enabling our complete phase-type model for itinerary completion to capture this feature.

4.5.3 Incorporating Probabilistic Resource Needs into the Phase-type Itinerary Completion Model

Based on the analysis of the Out-PATTERN stochastic location process it is clear that patients may only require each task along their critical path with a probability that is less than one. By employing the Out-PATTERN stochastic location process model it is possible to construct a more general critical path representation that contains the inherent stochasticity in care path transitions. We allow each visit along the critical path to be skipped (i.e., it was not needed for the given patient) with some probability that is fixed and calculated based on the Out-PATTERN location process.

Consider a typical uncapacitated patient flow path characterizing the probability that patients require certain clinical services over time in a setting without blocking (as discussed in Sec. 4.3.2). An example of such a flow path is given in Table 4.3. In this case, all patients have a visit to the breast diagnostic clinic on their first day, while 17% visit medical oncology. On their second day of treatment 24% of patients visit the breast diagnostic clinic and 4% visit medical oncology. etc.

The probability that a patient of type k will require task i (requiring service u_i) in precedence level d along their critical path, $\nu_{k,u_i}(d)$, is easy to compute from the data

Day	Breast Diagnostic	Med. Oncology	Radiation Onc.	Gen Surg	Plastic Surg
0	1	0.17	0.25	1	0.29
1	0.24	0.04	0.04	0.01	0.03
2	0.06	0.05	0.01	0.05	0.04
3	0.05	0.04	0.01	0.01	0.01
4	0.05	0.05	0.01	0.03	0.01

Table 4.3: Sample of a logistical care pathway model for breast cancer patients

(and will be discussed in Sec. 4.7). To combine this probability measure with the discrete phase-type distribution it is necessary to modify the prior Markov chain of Fig. 4.5. Figure 4.6 presents part of the modified Markov chain that is representative of the entire chain. When a patient completes any task j, they will move to the next task only if the task is required, w.p. $\nu_{k,u_{j+1}}(d)$. If task $j+1$ is not required (w.p. $1-\nu_{k,u_{j+1}}(d)$), then the chain can jump to task $j+2$. If both tasks $j+1$ and $j+2$ are not required, which occurs w.p. $(1-\nu_{k,u_{j+1}}(d))(1-\nu_{k,u_{j+2}}(d))$, then the chain can jump to task $j+3$ and so forth.

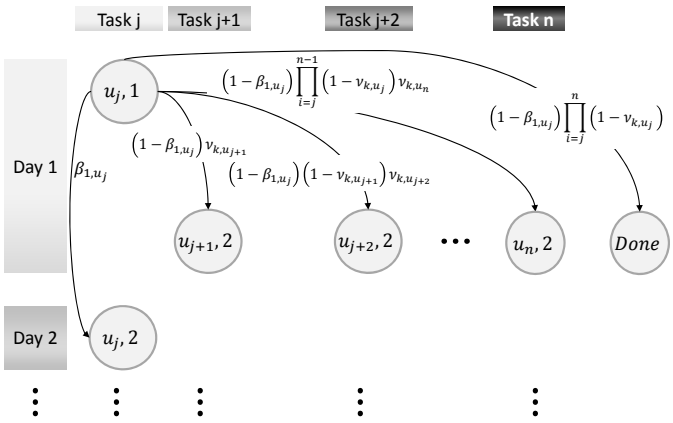


Figure 4.6: State transition diagram of the care path Markov chain that includes probabilities of not visiting a specific tasks.

The new transition probability matrix then has the upper triangular form

$$\left[\begin{array}{c|c} \mathbf{T}_{\mathcal{C}} & \mathbf{T}^{\mathbf{0}}_{\mathcal{C}} \\ \hline \mathbf{0} & 1 \end{array}\right] = \left[\begin{array}{cccccc|c} \mathbf{T}^{\mathbf{1}}_{\mathbf{u_1}} & \mathbf{T}^{\mathbf{2}}_{\mathbf{u_1}} & \mathbf{T}^{\mathbf{3}}_{\mathbf{u_1}} & \mathbf{T}^{\mathbf{4}}_{\mathbf{u_1}} & \dots & \mathbf{T}^{\mathbf{n}}_{\mathbf{u_1}} & \mathbf{T}^{\mathbf{0}}_{\mathbf{u_1}} \\ \mathbf{0} & \mathbf{T}^{\mathbf{1}}_{\mathbf{u_2}} & \mathbf{T}^{\mathbf{2}}_{\mathbf{u_2}} & \mathbf{T}^{\mathbf{3}}_{\mathbf{u_2}} & \dots & \mathbf{T}^{\mathbf{n}}_{\mathbf{u_2}} & \mathbf{T}^{\mathbf{0}}_{\mathbf{u_2}} \\ & \ddots & \ddots & & & & \vdots \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \dots & \mathbf{T}^{\mathbf{1}}_{\mathbf{u_n}} & \mathbf{T}^{\mathbf{0}}_{\mathbf{u_n}} \\ \hline \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \dots & \mathbf{0} & 1 \end{array}\right]. \tag{4.22}$$

$\mathbf{T}^{\mathbf{1}}_{\mathbf{u_i}}$ and $\mathbf{T}^{\mathbf{0}}_{\mathbf{u_n}}$ remain the same as in Eq. 4.18. For the remaining terms, we redefine them as

$$\mathbf{T}^{\mathbf{j}}_{\mathbf{u_i}} = \nu_{k,u_{i+j-1}}(j) \prod_{m=i+1}^{i+j-2} (1-\nu_{k,u_m}(m)) \begin{bmatrix} 1-\beta_{1,u_i} & 0 & 0 & 0 \\ 0 & 1-\beta_{2,u_i} & 0 & 0 \\ 0 & 0 & \ddots & 0 \\ 0 & 0 & 0 & 1-\beta_{5,u_i} \end{bmatrix}, \tag{4.23}$$

for $j = 2, \dots, n-i+1$. For the remaining terms

$$\mathbf{T}^{\mathbf{0}}_{\mathbf{u_i}} = \left[\prod_{m=i+1}^{n} \Big(1-\nu_{k,u_m}(m)\Big)\right] \begin{bmatrix} 1-\beta_{1,u_i} & 1-\beta_{2,u_i} & 1-\beta_{3,u_i} & 1-\beta_{4,u_i} & 1-\beta_{5,u_i} \end{bmatrix}'. \tag{4.24}$$

Eq.'s 4.23 and 4.24, use the convention that an empty product is equal to 1.

Beyond the change to the transition probability matrix, the data showed that the starting point of the patient's critical path is not deterministic. For a type k patient, let κ_k be the probability vector $\kappa_k(i)$ is the probability that the patient begins their critical path on service i.

$$\kappa_k = \left[\nu_{k,u_1}(1), \nu_{k,u_2}(2)(1-\nu_{k,u_1}(1)), \dots, \nu_{k,u_n}(n) \prod_{j=1}^{n-1} (1-\nu_{k,u_j}(j))\right]'. \tag{4.25}$$

Now let $\eta_{k,d}$ be the initial distribution on the starting location of a patient of type k that is scheduled for their initial appointment on day d. This distribution is given

by

$$\eta_{k,d} = [\nu_{k,u_1}(1)\mathbf{e}_d, \nu_{k,u_2}(2)(1-\nu_{k,u_1}(1))\mathbf{e}_d, \ldots, \nu_{k,u_n}(n)\prod_{j=1}^{n-1}(1-\nu_{k,u_j}(j))\mathbf{e}_d, 0]'. \tag{4.26}$$

The $(N \cdot n + 1) \times 1$ vector $\eta_{k,d}$ combines the initial service the patient enters (κ_k) with the day the patient enters that service ($\mathbf{e}_d$), to start the patient's path in the correct location at the correct time. Now, for a patient of type k, we can calculate the discrete time phase-type distribution of the probability that itinerary completes on or before stage/day x (CDF denoted by F_k) and the probability that the patient leaves the system at time x (pmf denoted by f_k).

$$F_{k,d}(x) = 1 - \eta_{k,d}\,(\mathbf{T}_{\mathcal{C}})^x \cdot \mathbf{1} \tag{4.27}$$

$$f_{k,d}(x) = \eta_{k,d}\,(\mathbf{T}_{\mathcal{C}})^{x-1} \cdot \mathbf{T}^{\mathbf{0}}_{\mathcal{C}}. \tag{4.28}$$

4.5.4 Tasks in Parallel: Maximum of Phase-type Distributions

The third and final feature we desire to add to the outpatient service model is that the critical path may contain non-identical tasks that can occur in parallel. That is, in precedence level d, there can be n parallel tasks with time to completion denoted by $X_1, X_2, \ldots, X_n$. From the modeling approach detailed in Sec. 4.5.2, each X_i follows a phase-type distribution characterized by generator matrix $\mathbf{T_i}$. The time to complete all tasks is given by $X = \max\{X_1, X_2, \ldots, X_n\}$. To analyze the distribution of X it is necessary to present the following definition (from [13]).

Definition IV.5. The Kronecker Product of matrices $\mathbf{A}$ and $\mathbf{B}$, $A \otimes B$, is an operation on two matrices, such that if $\mathbf{A}$ is an $m \times n$ matrix and $\mathbf{B}$ is a $p \times q$ then the $mp \times nq$ matrix is:

$$A \otimes B = \begin{bmatrix} a_{11}\mathbf{B} & \dots & a_{1n}\mathbf{B} \\ \vdots & \ddots & \vdots \\ a_{m1}\mathbf{B} & \dots & a_{mn}\mathbf{B} \end{bmatrix}. \tag{4.29}$$

Lemma IV.6 shows that the maximum of n phase-type distributions is a phase-type distribution. The proof builds upon Davio (see [13]), who shows the result for two phase-type distributions. While the generator matrix for the the maximum does have a closed form solution, it is extremely complex and so we present the more simple recursive formula for the generator matrix, from which it is easier to gain intuition.

Lemma IV.6. *Let $X_1, X_2, \dots, X_n$ be phase-type distributed random variables with generating matrices $\mathbf{T}_1, \dots, \mathbf{T}_n$ and the normalizing absorbing probability vector $\mathbf{T}_1^0$, $\dots, \mathbf{T}_n^0$. Let $X^{(n)} = \max\{X_1, X_2, \dots, X_n\}$. Then $X^{(n)}$ is phase-type distributed and the generator matrix for $X^{(n)}$ is upper triangular and has block form defined by the recursive relationship*

$$\mathbf{T}_{X^{(n)}} = \begin{bmatrix} \mathbf{T}_{X^{(n-1)}} \otimes \mathbf{T}_n & \mathbf{T}_{X^{(n-1)}} \otimes \mathbf{T}_n^0 & \mathbf{T}_{X^{(n-1)}}^0 \otimes \mathbf{T}_n \\ \mathbf{0} & \mathbf{T}_{X^{(n-1)}} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{T}_n \end{bmatrix}, \tag{4.30}$$

where $\mathbf{T}_{X^{(1)}} = \mathbf{T}_1$.

Proof. To show the result, we begin with the maximum of two phase-type distributions and then apply the relationship recursively to obtain the general result. The maximum of two phase type distributions, X_1, X_2, with generator matrices

$\mathbf{T}_1$ (with absorbing probability vector $\mathbf{T}_1^0$) and $\mathbf{T}_2$ (with absorbing probability vector $\mathbf{T}_2^0$) respectively can be shown to have a phase-type distribution by combining the states of the Markov Chains for each individual distribution. That is if the chain for X_1 has m states $a_1, \ldots, a_m$ and chain for X_2 has n states $b_1, \ldots, b_n$, then the combined chain would have $m \cdot n$ states $\{(a_i, b_j) : i \in \{1, \ldots, m+1\}, j \in \{1, \ldots, n+1\}\}/\{(a_{m+1}, b_{n+1})\}$, where a_{m+1} and b_{n+1} represent the state where chain X_1 and chain X_2 respectively have entered the absorbing state. We remove the case where both X_1 and X_2 have entered the absorbing state, because this is the absorbing state for the chain $max(X_1, X_2)$ and thus is not included in the generator matrix. The transitions are then defined by

(4.31)

$$\mathbb{P}((a_i, b_j) \to (a_k, b_\ell)) = \mathbb{P}(a_i \to a_k)\mathbb{P}(b_j \to b_\ell) \text{ for } i, k \in \{1, \ldots, m\}, j, \ell \in \{1, \ldots, n\}$$

(4.32)

$$\mathbb{P}((a_i, b_j) \to (a_k, b_{n+1})) = \mathbb{P}(a_i \to a_k)\left(1 - \sum_{r=1}^{n} \mathbb{P}(b_j, b_r)\right) \text{ for } i, k \in \{1, \ldots, m\}, \quad j \in \{1, \ldots, n\}$$

(4.33)

$$\mathbb{P}((a_i, b_j) \to (a_{m+1}, b_\ell)) = \left(1 - \sum_{r=1}^{m} \mathbb{P}(a_i, a_r)\right)\mathbb{P}(b_j \to b_\ell) \text{ for } i \in \{1, \ldots, m\}, \quad j, \ell \in \{1, \ldots, n\}$$

(4.34)

$$\mathbb{P}((a_i, b_{n+1}) \to (a_k, b_{n+1})) = \mathbb{P}(a_i \to a_k) \text{ for } i, k \in \{1, \ldots, m\}$$

(4.35)

$$\mathbb{P}((a_{m+1}, b_j) \to (a_{m+1}, b_\ell)) = \mathbb{P}(b_j \to b_\ell) \text{ for } j, \ell \in \{1, \ldots, n\}.$$

By arranging the transition matrix properly one can see that (1) Eq.'s 4.31 are

captured by the matrix $\mathbf{T}_1 \otimes \mathbf{T}_2$, (2) Eq.'s 4.32 are captured by $\mathbf{T}_1 \otimes \mathbf{T}_2^0$, (3) Eq.'s 4.33 are captured by $\mathbf{T}_1^0 \otimes \mathbf{T}_2$, (4) Eq.'s 4.34 are precisely $\mathbf{T}_1$, and (5) Eq.'s 4.35 are precisely $\mathbf{T}_2$. Certain transitions are prohibited, such as $(a_{m+1}, b_j) \rightarrow (a_k, b_\ell)$ for $k \neq m+1$ because once chain X_1 enters the absorbing state it doesn't leave. This creates blocks of zeros in the transition matrix for $max(X_1, X_2)$. Again, arranging the states in the right order yields the generator matrix that captures all the transitions:

$$\mathbf{T}_{max(X_1,X_2)} = \begin{bmatrix} \mathbf{T}_1 \otimes \mathbf{T}_2 & \mathbf{T}_1 \otimes \mathbf{T}_2^0 & \mathbf{T}_1^0 \otimes \mathbf{T}_2 \\ \mathbf{0} & \mathbf{T}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{T}_2 \end{bmatrix}. \tag{4.36}$$

Since $max(X_1, X_2)$ is still a phase-type distribution, it allows us to apply Eq. 4.36 recursively. For example $max(X_1, X_2, X_3) = max(max(X_1, X_2), X_3)$. Thus we can construct the recursive definition from Eq. 4.30 □

From Lemma IV.6 the maximum of n phase-type distributions has block upper triangular form. To obtain the closed for solution for the maximum of n phase-type distributions, first consider the result that the Kronecker product is distributive among blocks of a block matrix. That is,

$$\begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{D} & \mathbf{C} \end{bmatrix} \otimes \mathbf{N} = \begin{bmatrix} \mathbf{A} \otimes \mathbf{N} & \mathbf{B} \otimes \mathbf{N} \\ \mathbf{D} \otimes \mathbf{N} & \mathbf{C} \otimes \mathbf{N} \end{bmatrix}. \tag{4.37}$$

Combining the result from Eq. 4.37 and Lemma IV.6, the closed form solution for the recursion in Eq. 4.30 can be obtained. The solution, however, is notationally complex and therefore is not presented here.

4.5.5 A Tractable Representation of the "Max" Phase-type Generator

One of the primary drawbacks of the general representation of the generator matrix for the maximum of L phase-type distributions given by Eq. 4.30 is the size of

the matrix. If each matrix in the maximum is size $N \times N$, then the generator matrix of the maximum would be of size $(N+1)^L - 1 \times (N+1)^L - 1$. The exponential growth of the size of the matrix in the number of terms of the maximum is a concern. In our case study each individual generator matrix is 5×5 and we consider 5 possible services along the critical path, so the generator matrix for the maximum of those 5 services would be of size $7,775 \times 7,775$, with $60,450,625$ entries. If we were to add just one more service, the result would be a matrix with $2,176,689,025$ entries.

Since representations for phase-type distributions are in most cases not unique, we will exploit the special structure of our phase-type model to obtain a representation, $\mathbf{V}$, that is significantly more compact. The matrix simplification follows by eliminating the possibility of certain transitions and thereby reducing the state space. First, in the general form of the phase-type distribution, transitions can occur from any state to any other state. For itinerary completion, however, when an attempt to get an appointment is blocked, the chain never transitions more than one day ahead (i.e., transitions occur only from Day 1 to Day 2, or Day 2 to Day 3, etc.). Arranging the states properly in the Markov chain representation, allows us to create a sparse matrix structure that provides for efficient computation of powers. Secondly, for the general form of the maximum of phase-type random variables, each individual phase-type random variable that comprises the maximum is allowed to start in any state, independent of the other random variables in the maximum. Thus it is possible that the first random variable of the maximum starts on day 1, while the second starts on day 3. In the itinerary completion model, however, we begin attempting to get each individual appointment on the same day. In the previous example, all random variables of the maximum would start on the same day. This allows us to eliminate states such as (1,3), where the first appointment is attempting to get scheduled on

day 1 while the second appointment is attempting to get scheduled on day 3. The compact matrix, $\mathbf{V}$, has the following form, which will be explained in the following paragraphs.

$$\mathbf{V} = \begin{bmatrix} & \text{Day 1} & \text{Day 2} & \text{Day 3} & \text{Day 4} & \text{Day 5} \\ \text{Day 1} & \mathbf{0} & \mathbf{V}_{1,2} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \text{Day 2} & \mathbf{0} & \mathbf{0} & \mathbf{V}_{2,3} & \mathbf{0} & \mathbf{0} \\ \text{Day 3} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{V}_{3,4} & \mathbf{0} \\ \text{Day 4} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{V}_{4,5} \\ \text{Day 5} & \mathbf{V}_{5,1} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \end{bmatrix}. \tag{4.38}$$

Define $i \oplus 1$ as 1 if $i + 1 = N + 1$ (N being the length of the planning horizon) and as $i + 1$ otherwise (similar to mod). The blocks of zeroes in Eq. 4.38 are a result of the fact that on day i, the only possible next state for each appointment is day $i \oplus 1$ or that the appointment has been completed. Each block, $\mathbf{V_{i,i\oplus 1}}$, represents the transitions on day i. These transitions simply account for how many of the L services the patient was able to obtain an appointment at on day i. The 2^L states represented by $\mathbf{V_{i,i\oplus 1}}$ take the form $(a_1, a_2, \ldots, a_L)$, where $a_j = 1$ means that the patient has completed their appointment at service j and $a_j = 0$ means that the patient has not yet gotten an appointment at service j. As an example, trying to get an appointment at services 1 and 3 on day i, having already completed an appointment at service 2 yields on possible transition

$$\begin{aligned} \mathbb{P}((0,1,0) \to (0,1,1)) &= \mathbb{P}(\text{Get Appt at Svc 3})\mathbb{P}(\text{Not Get Appt at Svc 1}) \\ &= \beta_{1,i}(1 - \beta_{3,i}). \end{aligned}$$

For each day ($\mathbf{V_{i,i\oplus 1}}$), we only consider which of the remaining services the patient is able to get appointments in. The events of whether or not a patient can get an

appointment on day i in a single service u_j can be captured by the simplified 2×2 t.p.m.

$$A_{u_j,i} = \begin{bmatrix} \beta_{u_j,i} & 1 - \beta_{u_j,i} \\ 0 & 1 \end{bmatrix}. \tag{4.39}$$

$\beta_{u_j,i}$ is the probability that the patient couldn't get an appointment in service u_j on day i. If this event occurs, the patient is directed to the following day (e.g., $i \oplus 1$) and attempts to get the appointment again. Otherwise, the patient has finished the task and enters the absorbing state (i.e. task complete) for that task. Thus, $A_{u_j,i}$ describes the daily transition for a single service. The transition probabilities, $\mathbf{V_{i,i\oplus 1}}$, for the possible outcomes of trying to get appointments in L different services (i.e. success or failure on day i) can be calculated by combining the single service t.p.m's, $\bigotimes_{j=1}^{L} A_{u_j,i}$.

Theorem IV.7. *Suppose there are L phase-type distributed random variables,* $X_{u_1}, \ldots, X_{u_L}$*, with generator matrices,* $\mathbf{T}_{u_1}, \ldots, \mathbf{T}_{u_L}$*, following the structure in Eq. 4.15. Let* $A_{u_j,i}$ *be the compact representation of the single service t.p.m.'s for day i given by Eq. 4.39. Then Eq. 4.38 is a generator for* $\max_j\{X_{u_j}\}$*, where*

$$\left[\begin{array}{c|c} \mathbf{V_{i,i\oplus 1}} & \mathbf{V_i^0} \\ \hline \mathbf{0} & 1 \end{array}\right] = \bigotimes_{j=1}^{L} A_{u_j,i}. \tag{4.40}$$

Proof. This can be shown via equivalence of the compact Markov Chain with the Markov Chain for the general solution in Eq. 4.30. For the sake of expositional clarity we present a proof for the case of the maximum of 2 phase-type distributions because the general case follows using the exact same arguments.

Consider 2 phase-type distributions, X_{u_1} and X_{u_2} with generator matrices $\mathbf{T}_{u_1}$ and $\mathbf{T}_{u_2}$ respectively. The state space for the general solution is given by the $\{(i, j) :$

$i, j \in \{1, 2, \ldots, 6\}\}$, where 6 denotes service copmleted. Due to the structure of $\mathbf{T}_{u_1}$ and $\mathbf{T}_{u_2}$, the only possible transitions and their probabilities are given by

$$\mathbb{P}\left((i,j) \to (i \oplus 1, j \oplus 1)\right) = \beta_{i,u_1}\beta_{j,u_2} \tag{4.41}$$

$$\mathbb{P}\left((i,j) \to (i \oplus 1, 6)\right) = \beta_{i,u_1}(1 - \beta_{j,u_2}) \tag{4.42}$$

$$\mathbb{P}\left((i,j) \to (6, j \oplus 1)\right) = (1 - \beta_{i,u_1})\beta_{j,u_2} \tag{4.43}$$

$$\mathbb{P}\left((i,6) \to (i \oplus 1, 6)\right) = \beta_{i,u_1} \tag{4.44}$$

$$\mathbb{P}\left((6,j) \to (6, j \oplus 1)\right) = \beta_{j,u_2} \tag{4.45}$$

$$\mathbb{P}\left((i,6) \to (6,6)\right) = 1 - \beta_{i,u_1} \tag{4.46}$$

$$\mathbb{P}\left((6,j) \to (6,6)\right) = 1 - \beta_{j,u_2} \tag{4.47}$$

$$\mathbb{P}\left((6,6) \to (6,6)\right) = 1. \tag{4.48}$$

For this chain, the difference between i and j will remain constant until one of them enters the absorbing state (State 6). Considering this and the fact that the initial distribution in our application ($\mathbf{e}_d$) starts the search for both appointments on the same day of the week further reduces the the state space that needs be modeled to the state space $\{(i,i) : i \in 1, \ldots, 5\} \cup \{(i,6) : i \in 1, \ldots, 6\} \cup \{(6,i) : i \in 1, \ldots, 5\}$ because there is no path to any of the states where $i \neq j$ for $i, j < 6$.

Without loss of generality, another way to compute the $\mathbf{V}_{ij}$ blocks of Eq. 4.38 for the example of $(i,j) = (1,2)$:

$$\mathbf{V}_{1,2} = \begin{bmatrix} State & (2,2) & (2,6) & (6,2) \\ (1,1) & \beta_{1,u_1}\beta_{2,u_2} & \beta_{1,u_1}(1-\beta_{2,u_2}) & (1-\beta_{1,u_1})\beta_{2,u_2} \\ (1,6) & 0 & \beta_{1,u_1} & 0 \\ (6,1) & 0 & 0 & \beta_{2,u_2} \end{bmatrix}, \tag{4.49}$$

where the states are listed on the top and side of the matrix for exposition. By examining the entries of Eq. 4.49 it is clear that the transition probabilities match

those from the general phase-type representation Eq.'s 4.41 - 4.45. Eq.'s 4.46 - 4.48 come from the transitions to the absorbing state, which can be calculated by subtracting each row sum from 1. In the case of day 1, the transitions to the absorbing state from states $(1,1)$, $(1,6)$, and $6,1$ have probabilities $(1-\beta_{1,u_1})(1-\beta_{2,u_2})$, $(1-\beta_{1,u_1})$, and $(1-\beta_{2,u_2})$ respectively. These match the transitions from the general phase-type representation of the maximum. All other days have the same structure as day 1, thus we have shown that the two Markov Chains are equivalent because they have the same t.p.m. □

Corollary IV.8. *Let $U(n)$ be the computational complexity of multiplying $n \times n$ upper triangular matrices together. Let N be the length of the planning horizon for the blocking phase-type distribution described by Eq. 4.38. The computational complexity of calculating the CDF of the itinerary completion phase-type distribution for the maximum of L phase-type distributions using the compact approach is given by*

$$O_{compact}(F(x)) = N \cdot x \cdot U(2^L) \tag{4.50}$$

Proof. We prove the result by construction, presenting an algorithm for computing $F(x)$ that achieves the required complexity. Consider a matrix that has block form

$$\mathbf{V} = \begin{bmatrix} \mathbf{0} & \mathbf{V}_{1,2} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{V}_{2,3} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \ddots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{V}_{N-1,N} \\ \mathbf{V}_{N,1} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \end{bmatrix}. \tag{4.51}$$

Each time $\mathbf{V}$ is taken to the x^{th} power, each block $\mathbf{V}_{i,i\oplus 1}$ is shifted up by $x-1$ blocks (wrapping around to the bottom when the block reaches the top most block in the matrix) and multiplied sequentially on the left by the non-zero blocks to the left of it (again wrapping around to the right as necessary). This is best demonstrated by a simple example.

(4.52)

$$\mathbf{V}^2 = \begin{bmatrix} \mathbf{0} & \mathbf{0} & \mathbf{V}_{1,2}\mathbf{V}_{2,3} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \ddots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{V}_{N-2,N-1}\mathbf{V}_{N-1,N} \\ \mathbf{V}_{N-1,N}\mathbf{V}_{N,1} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{V}_{N,1}\mathbf{V}_{1,2} & \mathbf{0} & \mathbf{0} & \mathbf{0} \end{bmatrix}$$

(4.53)

$$\mathbf{V}^3 = \begin{bmatrix} \mathbf{0} & \mathbf{0} & \mathbf{0} & \ddots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{V}_{N-3,N-2}\mathbf{V}_{N-2,N-1}\mathbf{V}_{N-1,N} \\ \mathbf{V}_{N-2,N-1}\mathbf{V}_{N-1,N}\mathbf{V}_{N,1} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \ddots & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{V}_{N,1}\mathbf{V}_{1,2}\mathbf{V}_{2,3} & \mathbf{0} & \mathbf{0} \end{bmatrix},$$

where the diagonal dots, $\ddots$, in Eq. 4.53 represent a continuation of the pattern from the columns to the left of it where non-zero blocks will appear along the diagonal.

Thus the calculation of $\mathbf{V}^x$ decomposes into the multiplication of x matrices for each non-zero block and shifting that block's position up by x blocks. Because each matrix $\mathbf{V}_{d,d+1}$ is a portion of the Kronecker product of L matrices of size 2×2, the size of $\mathbf{V}_{d,d+1}$ is $(2^L - 1) \times (2^L - 1)$. To calculate the non-zero block of each colum

of $\mathbf{V}^x$ it requires $x-1$ multiplications of the $(2^L-1)\times(2^L-1)$ matrices, which is of order $x \cdot U(2^L)$. Since there are $N-1$ columns, this procedure must be repeated $N-1$ times, leading to the desired complexity on the order of $N \cdot x \cdot U(2^L)$. □

First note that the computational complexity grows linearly in the length of the planning horizon, which is far slower than the traditional representation. As an illustration of the importance of the above decomposition method, we compare the computational complexity of the compact representation ($O_{compact}$) with the standard representation (O_{full}) of the phase-type distribution. Let $M(n)$ be the computational complexity of regular matrix multiplication. Then

$$O_{compact}(F(x)) = N \cdot x \cdot U(2^L) \tag{4.54}$$

$$O_{full}(F(x)) = x \cdot M((N+1)^L). \tag{4.55}$$

The size of the matrix multiplication in Eq. 4.55 grows far more quickly than in Eq. 4.54. We compact the state space and compute the matrix power by multiplying smaller sub-matrices instead of the entire matrix. With the 5 day planning horizon and 5 services that could be potentially visited in parallel, the *compact representation* only requires matrix multiplication of size 31×31 that have 961 entries, where as the *general representation* requires matrix multiplication of a size $7,775 \times 7,775$ matrix with $60,450,625$ entries. If we again consider adding just one more service the contrast becomes even more stark: $3,969$ entries versus $2,176,689,025$ entries.

4.5.6 Phase-type Model for Itinerary Completion

We have developed a phase-type model of blocking that considers (1) completing a sequence of appointments with precedence constraints, (2) some tasks may be completed in parallel, and (3) the idea that some appointments along the critical path may not be required by all patients (i.e. there is a probability that patients can

skip over certain appointments on the critical path). In this section we incorporate all of the features of itinerary completion into a phase-type approach to calculate the probability of whether or not a patient will complete their itinerary given an initial appointment on day d.

Let $\mathcal{R}_d \subseteq \mathcal{R}$ be a cluster of services that are required in precedence level d. Let $\mathbf{V}(\mathcal{R}_d)$ be the generator for the phase-type model of completion time per Eq. 4.38 from Sec. 4.5.4. The phase-type model that captures itinerary completion is given by replacing the generator, $\mathbf{T^1_{u_d}}$ for each individual service in Eq. 4.22 with $\mathbf{V}^1(\mathcal{R}_d)$. The transition from precedence level d to $d+i$, $\mathbf{V^i}(\mathcal{R_d})$, (or to the absorbing state $\mathbf{V^0}(\mathcal{R_d})$) is calculated the same way as $\mathbf{T}^i_{u_d}$ (or $\mathbf{T}^0_{u_d}$) from Eq. 4.23 by subtracting the row sum of each row of the matrix from 1 and multiplying by the probability the cluster of services in level $d+i$ is needed, $\mathbb{P}(\mathbb{1}\{\mathcal{R}_{d+i}\})$. It is too large to display here. Because this generator matrix has block form with a significant number of zero blocks, the computation of the power of the phase-type generator matrix for the entire path, $\mathbf{V}(\mathcal{C})$, can be calculated by multiplying the smaller blocks of the total matrix together using an algorithm that is quite similar to the one proposed in the proof of Corollary IV.8.

Let $\eta_{k,d}$ be the initial distribution for patients of type k beginning their itinerary on day d of the planning horizon. We define $\eta_{k,d}$ using the same form as previously given in Eq. 4.26, but with the proper dimensions. The probability that a type k patient's itinerary completes before the end of the work week given that they were admitted on day $d \in \{0, \ldots, 4\}$ is then given by

(4.56)

$$\mathbb{P}\left(\mathbb{1}_{k,d}\{\text{Itinerary Complete}\} = 1\right) = \mathbb{P}\left(\delta_{\mathcal{C},d}(\mathbf{B}) \leq 5 - d\right) = 1 - \eta_{k,d}(\mathbf{V}(\mathcal{C}))^{5-d} \cdot \mathbf{1}.$$

Recall that $\delta_{\mathcal{C},d}(\mathbf{B})$ is the sojourn time for a patient with critical path $\mathcal{C}$ who was

originally admitted on day d and $\mathbf{V}(\mathcal{C})^{(5-d)}$ can be calculated efficiently using the algorithm developed in the proof of Cor. IV.8. In Sec. 4.6 we develop an optimization model that takes the phase-type model as an input and maximizes itinerary completion for the priority fast-track patients.

4.6 Itinerary Completion Optimization

For the second stage optimization problem, we begin with the workload dynamics of the system resulting from the workload smoothing optimization and we add a prioritized class for fast-track scheduling. That is, for each national or international patient, we can calculate the probability that their treatment segment will not complete by Friday based on the day they initiated treatment, their critical path, and the amount of delay they experience. The latter depends on the blocking probabilities from the workload smoothing optimization, which parameterize the discrete phase-type CDF from Eq. 4.56.

To ensure that blocking probabilities in the itinerary completion optimization closely match those of the optimal schedule of the workload smoothing optimization, we add a constraint on the service level at each resource to the itinerary completion optimization. Let Θ^* be the optimal schedule from the workload smoothing optimization. Then the service level at service u on day d is given by

$$\tau_{u,d} = \frac{C_{u,d} - \mu_{u,d}(\Theta^*)}{\sigma_{u,d}(\Theta^*)}, \tag{4.57}$$

where $C_{u,d}$ is the capacity, $\mu_{u,d}(\Theta^*)$ and $\sigma_{u,d}(\Theta^*)$ are the mean and standard deviation of the workload under schedule Θ^*. By constraining the service level in the itinerary completion optimization to respect the service levels under Θ^*, we ensure that the itinerary completion model is properly parameterized with similar flow dynamics to Θ^*. This is important because $\bar{F}_k(t)$, the itinerary completion probability, is depen-

dent on the blocking probabilities from the solution of the workload smoothing stage. Thus, as long as the blocking is not much worse in the second stage optimization, $\bar{F}_k(t)$ will be an upper bound on the probability of not completing an itinerary.

The key to the itinerary completion fast-track optimization is that we enrich the patient type to include each patient's geocode (national/international vs local/regional). For example, the breast cancer patient type would become *local* breast cancer patient, *national* breast cancer patient, etc. instead of just breast cancer patient as in the first stage model. We use the same variable definitions as in the workload smoothing optimization, with the following changes.

Parameters

$\theta^N_{k,d}$ current admission volumes of type k national/international patients on day d.

$\theta^L_{k,d}$ current admission volumes of type k local/regional patients on day d.

$\hat{\theta}^N_{k,d}$ maximum number of admissions of type k national/international patients allowed on day d.

$\hat{\theta}^L_{k,d}$ maximum number of admissions of type k local/regional patients allowed on day d.

$\tau_{u,d}$ the maximum allowable coefficient of the workload standard deviation values obtained from the solution of the workload smoothing optimization)

$\bar{F}_k(t)$ the probability that a type k patient's critical path takes longer than t units of time. $\bar{F}_k(t) = 1 - F_k(t)$, where $F_k(t)$ is given by Eq. 4.56 for patient type k.

Decision Variables

$\Theta^N_{k,d}$ number of type $k \in \mathfrak{D}$ national/international patients scheduled on day d

$\Theta^L_{k,d}$ number of type $k \in \mathfrak{D}$ local/regional patients scheduled on day d

This optimization model maximizes the itinerary completion for national / international patients while constraining the system to the smoothed aggregate system flow from the workload smoothing stage.

(4.58)

$$\min_{\Theta^N,\Theta^L} \sum_{k\in\mathfrak{D}}\sum_{d=1}^{5} \Theta^N_{k,d}\bar{F}_k(6-d)$$

$s.t.$

$$\mu_{u,d}(\Theta^N+\Theta^L)+\tau_{u,d}\cdot\frac{1}{2}\left(\frac{\sigma^2_{u,d}(\Theta^N+\Theta^L)}{\hat{\sigma}_{u,d}(\Theta^N+\Theta^L)}+\hat{\sigma}_{u,d}(\Theta^N+\Theta^L)\right)\leq C_{u,d}+\epsilon$$

(4.59)

$$\forall u\in\mathcal{U}, d=1,\ldots,5$$

(4.60)

$$\sum_{d=1}^{5}\Theta^N_{k,d}=\sum_{d=1}^{5}\theta^N_{k,d} \qquad \forall k\in\mathfrak{D}$$

(4.61)

$$\Theta^N_{k,d}\leq\hat{\theta}^N_{k,d} \qquad \forall k\in\mathfrak{D}, d=1,\ldots,5$$

(4.62)

$$\sum_{d=1}^{5}\Theta^L_{k,d}=\sum_{d=1}^{5}\theta^L_{k,d} \qquad \forall k\in\mathfrak{D}$$

(4.63)

$$\Theta^L_{k,d}\leq\hat{\theta}^L_{k,d} \qquad \forall k\in\mathfrak{D}, d=1,\ldots,5$$

$$\Theta^L_{k,d},\Theta^N_{k,d}\in\mathbb{Z}^+.$$

The objective function, Eq. 4.58, minimizes the expected number of incomplete itineraries across all patient types. Itinerary completion of a patient of type k beginning their itinerary on day d is modeled as a Bernoulli random variable with probability $F_k(6-d)$, where F_k is given by Eq. 4.56. Eq. 4.59 is the constraint that enforces the service level factor, $\tau_{u,d}$, for any given schedule $\Theta^L+\Theta^N$ using our Newton's method-based approximation of the square root function from Section 4.4.1.

Essentially Eq.'s 4.59 ensure that the system will retain a flow that is characterized by $F_k(\cdot)$, the minimum blocking solution of workload smoothing. We add ϵ to ensure that there are sufficiently many solutions for CPLEX to solve the problem easily and not be hampered by round-off errors.

Note that $\Theta^L + \Theta^N$ encompasses all the patients that Θ did in the workload smoothing optimization. Eq.'s 4.60 and 4.62 constrain the optimal weekly volume to be equal to the current weekly volume for both national/international patients and local/regional patients. Eq.'s 4.61 and 4.63 ensure that an upper bound on the number of patients of each type arriving for treatment on a given day is respected for both national/international patients and local/regional patients.

4.7 Analysis and Case Study of Itinerary Completion Improvement

In this section, we present an application of our outpatient flow paradigm to improve itinerary completion for national/international breast cancer patients at the Mayo Clinic. In principle the methods can be applied to the ensemble of services offered by the Mayo Clinic, but given that this is a research project, we tested our model on the scheduling of new breast cancer patients.

4.7.1 Data and Model Parametrization

Clinical Resources. The first challenge in parameterizing the theoretical models developed in the chapter was to determine how to categorize the clinical resources that patients place a load on. Key considerations for effective resource categories include:

1. Are the definitions of resources amenable to measurement of patient workload (e.g. number of appointments, OR time, etc.)?

2. Is it possible to quantify the resource's capacity?

3. Is the resource a bottleneck; i.e. is the resource capacity constrained?

Taking into account these considerations, we decided to categorize resources by clinical service (e.g. general surgery, radiation oncology) and, where appropriate, by the particular type of appointment at the clinical service (e.g. surgical consult, registered nurse, diagnostic imaging). The workload and capacities at these services were quantified in terms of number of appointments per day as detailed in the following paragraphs.

Patient Types. Choosing the patient types defines the granularity of the model. It is important to have enough differentiation between patients to make patient types meaningful, but to have enough data for each patient type category to be able to effectively estimate the stochastic location processes. In our case study, we decided to further subdivide breast cancer patients into categories by geo-code (local, regional, national, international). Not only did patients from different regions exhibit different care paths and resource usage, the geo-coded region also facilitates the second stage optimization in which the goal is to maximize itinerary completion for national/international patients.

Logistic Care Pathways. To parameterize the breast cancer stochastic location processes, we desire pathways that are not distorted by access block. Thus we pulled patient flow data from periods of low congestion during 2006 - 2011. This set of breast cancer patients placed a load on 77 different clinical specialties over this time period, all of which were modeled. Breast cancer patients were further characterized by their geo-code (local, regional, national, international). An example of the care paths can be found in Appendix 4.9.2, Figure 4.11.

Clinical Service Capacities and Workloads. The total workload at each resource was measured in terms of number of appointments. Historical workloads for

two of the main breast cancer services are given in Figure 4.10 in Appendix 4.9.2.

To calculate the utilization, we divide the total appointments by the FTEs at the service to get the ratio of appointments per FTE. To estimate the capacity in each clinical service, we developed a quantile chart that plots the proportion of days that didn't exceed a certain level of appointments per FTE. We then chose an upper range as the "capacity" per FTE. It was mutually decided with the Mayo Clinic that the 90% quantile would be a good measure of staff capacity to serve patients in the Breast Diagnostic Clinic, because it is believed that patients are being "squeezed" in the top 10% of days worked. An example of the quantile plot for the Breast Diagnostic Clinic is given in Appendix 4.9.2, Fig. 4.12. The capacity, in terms of number of appointments, was then calculated by multiplying the maximum number of appointments per staff member by the number of staff members on duty for each day of the planning horizon.

An additional complication was that the clinical services that we studied did not exclusively serve breast cancer patients. In our case study, we only control the breast cancer patient arrival stream so we model the other patients that use the same services as exogenous, uncontrolled demand. The uncontrolled demand is approximated as a normal distribution with the mean and variance parameters estimated from historical data. Clearly controlling all of the demand is the ideal and would allow for the best possible results. However, in healthcare organizations there are many independent parties and not all participants are able or willing to change their scheduling practice. This inflexibility can lead to significant extra costs, but as we show in the following case study, great benefits can be achieved even by controlling a small (e.g. 25%) fraction of the demand.

Breast Cancer Critical Paths. One of the key challenges in this research lies in determining critical paths for patients in a scalable manner. Working with the Mayo Clinic, we investigated a number of approaches for determining the critical path for each patient type. For scalability, we decided upon a data driven approach that identifies resources that are:

1. *Commonly Used*. A significant proportion of patients from a given patient type visited at least once

2. *Caused Itinerary Completion Failure*. Comprised a significant percent of the visits that exceeded the one week mark. These can be deduced to be critical appointments.

For the breast cancer patients in our proof-of-concept case study, we identified 5 services that lay along the critical path that met both criteria: Breast Diagnostic Clinic, Medical Oncology, Radiation Oncology, General Surgery, and Plastic Surgery. A table of the top resources used on the 2nd week (i.e. caused an itinerary completion failure) is shown in Figure 4.13 of Appendix 4.9.2.

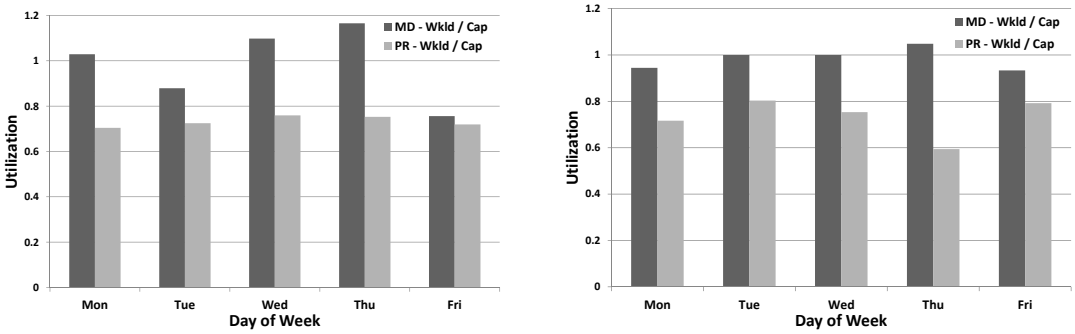
4.7.2 Case Study Results

In this section we present the results obtained by applying the (1) workload smoothing and (2) itinerary completion optimization models to the critical resources associated with breast cancer care. This case study provides a detailed analysis and solution to the breast cancer patient scheduling problem that would provide the decision rules for managing itinerary completion for new breast cancer patients. In this study only the new breast cancer patient schedules would be modified, and thus exogenous demand at each service was modeled but not controlled. In each service along the breast cancer critical path, breast cancer patients amounted to less than

25% of the total volume of appointments. The results demonstrate that significant gains can be achieved even with this small amount of control.

To illustrate how the two-stage method works, we first present the intermediate workload smoothing results (stage 1) for the Breast Diagnostic Clinic and discuss the implication of this stage of the optimization. Then we present the final results in terms of breast cancer patient schedule and itinerary completion improvement after both the workload smoothing and itinerary completion optimizations were run sequentially.

Workload Smoothing Optimization. In this section we present the intermediary result for the workload smoothing optimization. To illustrate the concepts at play, we focus on a detailed analysis of the Breast Diagnostic Clinic. Because we are studying breast cancer patients, the Breast Diagnostic Clinic provides a rich environment for illustrating the insights from this intermediary stage. At the clinic, the key appointments were (1) Physician Visit (MD) and (2) Diagnostic Procedure (PR). Fig. 4.7 shows the results in terms of the utilization of the physicians and diagnostic imaging machines by day of week.



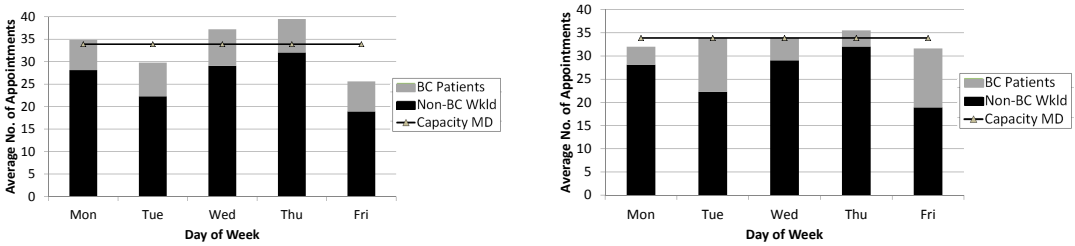
(a) Original schedule

(b) Smoothed schedule

Figure 4.7: Workloads for the physician (MD) appointment type and the diagnostic procedure (PR) appointment type. Observe that the MD capacity is the bottleneck

In Fig. 4.7 it can be seen that the physician visit is the bottleneck resource

at the Breast Diagnostic Clinic. In fact, the physician's workload at the Breast Diagnostic Clinic is over capacity on three out of the five weekdays (Fig 4.7 (a)). The optimal schedule (Fig. 4.7 (b)) smooths the physician workload relative to capacity - only slightly exceeding capacity on one out of the five weekdays. In the optimal schedule there is more variation in the diagnostic procedure service, but this is acceptable to achieve the gains in the physician workload because the average utilization of the diagnostic imaging machines is well below peak levels. Fig. 4.8 presents the optimization results for the physician schedule at the Breast Diagnostic Clinic in greater detail; showing the non-controllable (non-BC) and controllable (BC) appointment workloads.



(a) Original schedule

(b) Smoothed schedule

Figure 4.8: Physician appointments at the Breast Diagnostic Clinic for breast cancer (BC) and non-breast cancer (Non-BC) patients.

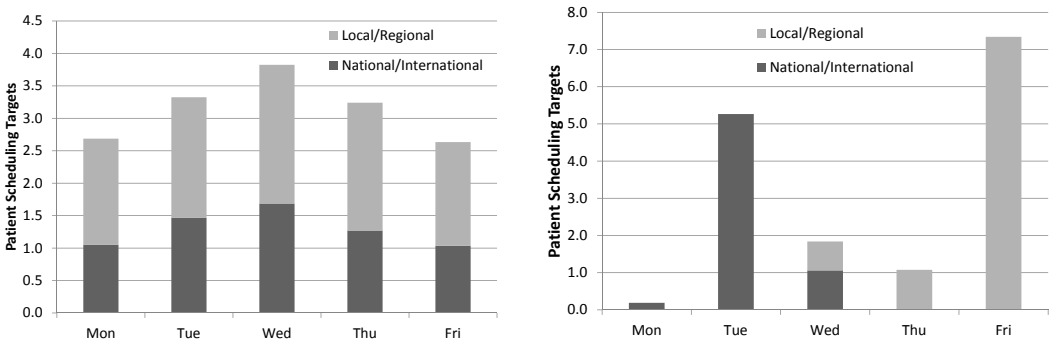
The key insight from Fig. 4.8 is the following: by controlling the schedules of a small fraction of the patient population, it is possible to mold the controlled workload to the exogenous workload like fitting pieces of a puzzle together to smooth workloads and reduce congestion in critical clinical services. Take, for example, Tuesday and Wednesday. In the original schedule (Fig. 4.8 (a)) the physicians experience about the same amount of BC patient workload on both days despite the fact that the exogenous workload is much higher on Wednesday than on Tuesday. After optimization (Fig. 4.8 (b)), patients are scheduled in such a way that the BC patient

appointment workloads are light on Wednesday when the exogenous workload is high, and higher on Tuesday when the exogenous workload is low. Thursday is still overutilized due to the extremely high level of non-controllable workload and the fact that any admissions of BC patients on Monday, Tuesday or Wednesday implies some follow-up workload on Thursday. This clearly demonstrates the potential value of controlling a larger segment of the patient population in a more comprehensive system-wide design.

Itinerary Completion Optimization. From the solution of the workload smoothing optimization, we calculated the (1) blocking probabilities to parameterize the national/international patient's phase-type care paths, (2) blocking quantiles to ensure that blocking probabilities are bounded above by the first stage solution (see Eq. 4.59), (3) the phase-type care paths with blocking incorporated. In setting the blocking quantile constraints, we relaxed the capacity by 1% of the original capacity (e.g., $\epsilon = 0.34$ for Breast Diagnostic Clinic).

The resulting schedule from the itinerary completion optimization in comparison with the original schedule is given in Fig. 4.9. The stage two optimization essentially pushes the national/international demand to the beginning of the week, while maintaining the workload smoothing properties of the stage one optimization so that mid-week blockages don't delay itinerary completion. This gives the national/international patients the best chance to complete their itinerary before the weekend. Note that this is the schedule of new patient starts (first day appointments) only. Subsequent and return return visits are accounted for in the workload forecast location model.

The result of this improved schedule increases the itinerary completion from 74% to 88%, as can be seen in Table 4.4. In the optimal schedule, the number of new breast



(a) Original schedule (b) Optimal schedule

Figure 4.9: Comparing the original schedule with the stage 2 optimal schedule for national / international patients versus local / regional patients.

	Current Schedule	Improved Schedule
Itinerary Completion	74%	88%

Table 4.4: Itinerary completion improvement using optimization

cancer patient appointments on Monday is low because there will be a significant number of follow-up appointments on Mondays, being driven by the fact that most of the local/regional patients are being shifted to Thursday and Friday. Because most national/international itineraries can be completed in 4 days or less if there are few appointment delays, starting the majority of "priority" patients on Tuesday doesn't greatly affect their chance of completing their itinerary by Friday.

The power of the approach is that patient service/throughput can be improved without adding capacity (the traditional approach to solving congestion problems in healthcare). Providing better and more efficient care with fixed capacity will become increasingly important as healthcare capacity becomes more and more constrained.

4.8 Conclusions

In this chapter we developed a two stage optimization approach to maximizing itinerary completion for a priority class (national/international) of patients in a destination hospital. The workload smoothing stage builds a stochastic arrival-location

model of patient flow along with linearizing approximations of blocking to optimize and smooth workloads across the critical services in the outpatient clinical service network. This approach to demand control in queueing networks transforms the stochastic problem into a determinstic one that can be solved via linear programming. The workload smoothing optimization stabilizes the environment through which patients flow. Within the stabilized environment designed by the workload smoothing optimization, the second stage allows for further differentiation of patients by "priority" (national/international vs local/regional). The second stage linear program maximizes itinerary completion based on the CDF of a phase-type distribution parameterized by the blocking probabilities endogenous to the smoothed workload stable environment.

The theoretical models were tested on breast cancer patients at the Mayo Clinic. A full model was developed and parameterized to model and control the flow of breast cancer patients only, accounting for non-breast cancer demand at each clinical service as exogenous demand. This model was able to achieve significant improvements in itinerary completion - increasing from 74% to 88% completion rate - while controlling less than 25% of the demand in each service.

While this approach was developed for and applied to itinerary completion of national/international patients at a destination hospital, it is generalizable to a prevalent problem in the healthcare industry: how to reduce delays and improve throughput by scheduling in a manner that uses expensive fixed capacity more effectively. Through analytical solutions, we have been able to design control systems for queueing networks that can enable hospitals to meet the growing need to serve patients for efficiently and effectively under increasing capacity constraints.

4.9 Appendix

4.9.1 Notation

The notation for the chapter is presented below, categorized by the section in which it first appears.

Section 4.3

M	the number of clinical services in the care network
$\mathcal{S}$	the vector state space for the Out-PATTERN stochastic location process consisting of $\mathcal{S}^0$, representing the clinical services and Δ representing being at home.
$L_{s,k}(t)$	the stochastic location process that represents random location of patient type k at time t given they began treatment at time s
Σ_s	the function space containing outcomes of the stochastic location process for patients beginning treatment at time s.
Σ	the collection of all Σ_s.
$p_{s,k,r}(j,t)$	the probability that a patient of type k who initiates a new treatment at time s requires j appointments in clinical service r at time t
$\mathfrak{D}$	set of all patient diagnosis types
Θ	a decision variable matrix representing the new appointment schedule over the planning horizon. $\Theta_{i,d}$ is the number of type i new patients to start on day d of the planning horizon.

M_r the maximum number of visits to resource r in a given day

$\mu_{d,r}(\Theta)$ the mean workload in terms of number of appointments in unit u on day d under schedule Θ

$\sigma^2_{d,r}(\Theta)$ the variance of the workload in unit u on day d

$D^t_{r,d}$ the random variable representing demand for clinical service r on day d of week t

$D^\infty_{r,d}$ the random variable representing the steady state demand for clinical service r on day d

Section 4.4

$\mathcal{M}'$ an index that creates a discrete grid with N sections.

$\mathcal{M}$ discrete grid values corresponding to the index $\mathcal{M}'$.

$m(\cdot)$ function mapping the grid index $\mathcal{M}'$ to the grid $\mathcal{M}$.

$X_{u,d}(\Theta)$ the normal approximation of the amount of demand for service u on day d given admission schedule Θ

$C_{u,d}$ the capacity in service u on day d.

$\delta_{u,d,i}$ as a decision variable that calculates the amount of overflow at $m(i)$ standard deviations above the mean

$\mathbb{E}[O_{u,d}]$ the expected overflow approximation in service u on day d

$\hat{\sigma}(\Theta)$ a guess of the workload standard deviation; can be chosen to be the historical standard deviation

$\theta_{k,d}$ current admission volumes of type k on day d.

$\hat{\theta}_{k,d}$ maximum number of admissions of type k allowed on day d.

$w_{u,d}$ the weight assigned to expected overflow in service u on day d

Section 4.5

$\mathcal{C}$ the tuple $(\mathcal{R}, \mathcal{P})$ defining a patient's critical path

$\mathcal{R}$ set of clinical services along a patient's critical path. $\mathcal{R} \subseteq \mathcal{U}$

$\mathcal{P}$ set of precedence relations between clinical services on the critical path

$\beta_{u_i,d}$ blocking probability at clinical service u_i on day d of the planning horizon

$\mathbf{B} = [\beta_{u_i,d}]$ the matrix of blocking probabilities by clinical service (u_i) by day of week (d)

$\delta_{u_i,d}(\mathbf{B})$ time to complete the i^{th} appointment (at service u_i) along the critical path given that the appointment was first requested for day d given blocking probabilities $\mathbf{B}$

$\delta_{\mathcal{C},d}(\mathbf{B})$ The total time to complete the treatment segment for an initial appointment on day d, given blocking probabilities $\mathbf{B}$

K deadline for completing the itinerary

$\mathbf{T}^1_{u_i}$ The phase-type generator matrix for the time to complete an appointment at service u_i

$\mathbf{T}^0_{u_i}$ The probability vector for transitioning to the absorbing state

$\mathbf{T}^j_{u_i}$ The matrix that routes the patient to service u_{i+j-1} after completing an appointment at service u_i

$\mathbf{T}_{\mathcal{C}}$ The phase-type generator for critical path $\mathcal{C}$

$\nu_{k,i}(d)$ The probability that a patient of type k will require task i along their critical path given that they were admitted on day s

κ_k the initial distribution of where the patient begins their critical path

$\eta_{k,d}$ the initial distribution on the starting location of a patient of type k that is scheduled for their initial appointment on day d

$\mathbf{V}$ compact representation of the maximum of phase-type distributions for the itinerary completion model

$\mathbf{V}_{i,j}$ Block of $\mathbf{V}$ representing the transition from day i to day j

$\mathbf{V}^i(\mathcal{R}_d)$ the compact representation equivalent of $\mathbf{T}^i_{u_d}$

$A_{u_j,i}$ a 2×2 phase-type generator matrix that represents the attempt to get an appointment in service j on day i

Section 4.6

$\theta^N_{k,d}$ current admission volumes of type k national/international patients on day d.

$\theta^L_{k,d}$ current admission volumes of type k local/regional patients on day d.

$\hat{\theta}^N_{k,d}$ maximum number of admissions of type k national/international patients allowed on day d.

$\hat{\theta}^L_{k,d}$ maximum number of admissions of type k local/regional patients allowed on day d.

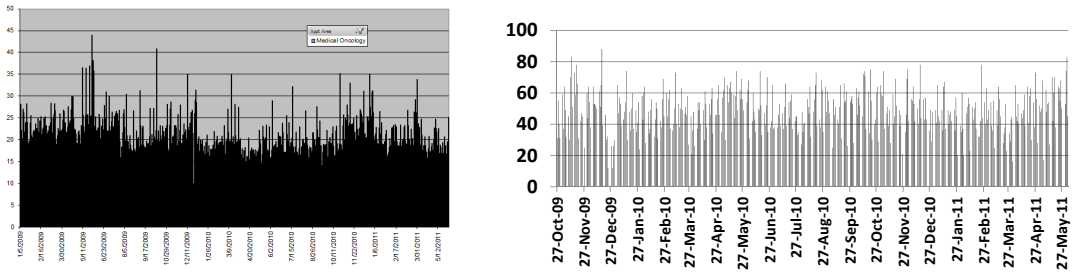
$\tau_{u,d}$ the service level factor for resource u on day d; calculated from the workload smoothing optimal schedule

$\bar{F}_k(t)$ the probability that a type k patient's critical path takes longer than t units of time. $\bar{F}_k(t) = 1 - F_k(t)$, where $F_k(t)$ is given by Eq. 4.20 for patient type k.

$\Theta^N_{k,d}$ number of type $k \in \mathfrak{D}$ national/international patients scheduled on day d

$\Theta^L_{k,d}$ number of type $k \in \mathfrak{D}$ local/regional patients scheduled on day d

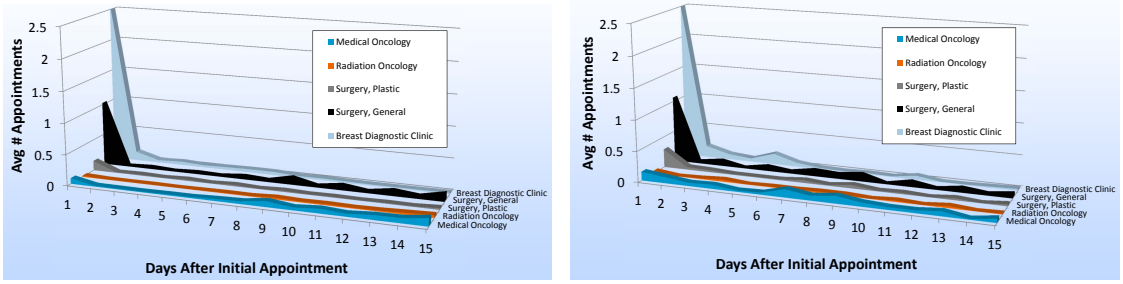
4.9.2 Mayo Clinic Breast Cancer Case Study Results and Analysis



(a) Medical Oncology appointment load (b) Breast Diagnostic Clinic appointment load

Figure 4.10: Workloads in two major breast cancer services over time.

The breast cancer pathways by patient type in Fig. 4.11 demonstrate the different ways in which local patients use resources versus international patients. The international patients use more resources early on in their treatment path but tend to taper of the further into the future on looks. On the other hand, the local patients use less resources at once but rather spread their care out over a longer period of time.



(a) Local (b) International

Figure 4.11: Representation of breast cancer patient logistic care pathways by patient type.

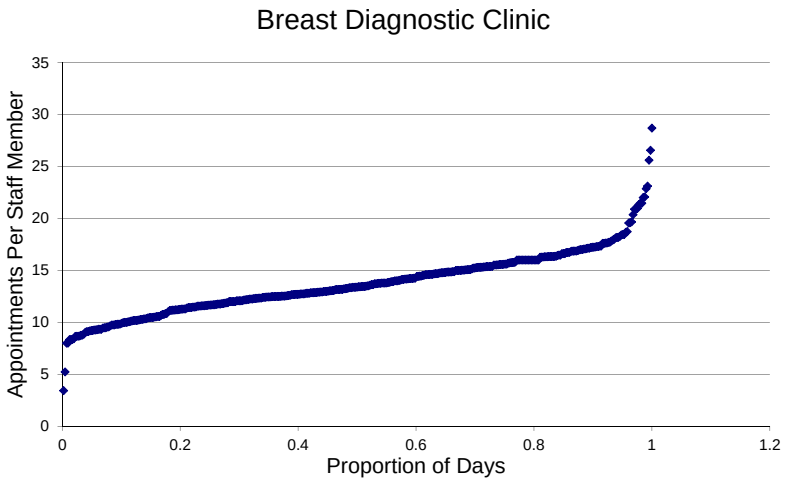


Figure 4.12: Plot of the cumulative proportion of days over a years worth of data that had up to a given patient load per staff member.

Clinical Service	% 2nd Week Visits
Surgery, General	20%
Medical Oncology	17%
Radiology, Nuclear Medicine	15%
Breast Diagnostic Clinic	13%
Surgery, Plastic	7%
Radiation Oncology	6%

Appointment Area	% 2nd Week Visits
MD - Staff Physician	33%
PR - Procedure/Diagnostic Testing	28%
IN - Individual	13%
CL - Clinic	12%

Appointment Type	% 2nd Week Visits
Consult/Limited Exam	35%
Subsequent Visit	31%
Radiology	22%
Procedure/Diagnostic testing	8%

Figure 4.13: Critical resources categorized by percent of 2nd week visits that occurred in each resource.

CHAPTER V

Conclusions

In healthcare, admission/scheduling practices have a great impact on care delivery performance in terms of cost, quality, and access. This body of work develops analytical tools for managing workloads in networks of healthcare resources to smooth workloads and enable more effective care delivery with fixed care resources. At a high level this work will allow hospitals to serve more patients and provide better access without expanding resource capacities. Through case studies developed using historical data from a number of hospitals the methods proposed here have been shown to

1. Reduce blocking and cancelation by 17%-32% while maintaining current patient volumes.

2. Serve 7% more patients without increasing blocking (a conservative estimate).

3. Increase the percent of patients completing their itinerary within a work week from 74% to 88%.

The theoretical contributions of the work lie in designing methods for tractable optimization of workloads in large and complex queueing networks. Chapter II develops a new controlled arrival-location model (CALM) and linearizing approximations

that transform stochastic metrics into a linear deterministic optimization problem. This approach allows for the optimization of elective inpatient admission schedules. Such optimization was not tractable using simulation-based optimization. Chapter III develops and analyzes a new stylized Markov Decision Process model, obtaining closure properties for a new queueing operator and threshold structure of the optimal policy. Chapter IV extends the CALM model of Chapter II and develops and analyzes a phase-type model for the flow of individual patients through a congested system.

In conclusion, this work has developed new methods for understanding flows and blocking in queueing networks that model the flow of patients in networks of healthcare services. These methods also provide insight and decision support for managing one of the highest impact controls in healthcare: the scheduling and admission of patients.

BIBLIOGRAPHY

BIBLIOGRAPHY

[1] I. Adan, J. Bekkers, N. Dellaert, J. Vissers, and X. Yu. Patient mix optimisation and stochastic resource requirements: A case study in cardiothoracic surgery planning. *Health Care Management Science*, 12(2):129–141, 2009.

[2] L.H. Aiken, S.P. Clarke, D.M. Sloane, J. Sochalski, and J.H. Silber. Hospital nurse staffing and patient mortality, nurse burnout, and job dissatisfaction. *Jama*, 288(16):1987–1993, 2002.

[3] P. Anderson, J. Meara, S. Brodhurst, S. Attwood, M. Timbrell, and A. Gatherer. Use of hospital beds: a cohort study of admissions to a provincial teaching hospital. *British Medical Journal*, 297(6653):910–912, 1988.

[4] R. Bekker and P.M. Koeleman. Scheduling admissions and reducing variability in bed demand. *Health Care Management Science*, 14(3):1–13, 2011.

[5] J. Beliën and E. Demeulemeester. Building cyclic master surgery schedules with leveled resulting bed occupancy. *European Journal of Operational Research*, 176(2):1185 – 1204., 2007.

[6] J. Bowers and G. Mould. The deferrable elective patient: A means of reducing waiting-lists in orthopaedics. *Journal of Management in Medicine*, 16:150–158(9), 2002.

[7] K.M. Bretthauer, H.S. Heese, H. Pun, and E. Coe. Blocking in healthcare operations: A new heuristic and an application. *Production and Operations Management.*, 2010. Forthcoming.

[8] K. Brownson and SB Dowd. Floating: a nurse's nightmare? *The Health Care Manag*, 15(3):10–15, 1997.

[9] T. Cayirli and E. Veral. Outpatient scheduling in health care: A review of literature. *Production and Operations Management*, 12(4):519–549, 2003.

[10] S.X. Chen and J.S. Liu. Statistical Applications of the Poisson-Binomial and conditional Bernoulli distributions. *Statistica Sinica*, 7(4):875–892, 1997.

[11] V.S. Chow, M.L. Puterman, N. Salehirad, W. Huang, and D. Atkins. Reducing surgical ward congestion through improved surgical scheduling and uncapacitated simulation. *Production and Operations Management*, 20(3):418–430, 2011.

[12] M.M. Connors. A stochastic elective admissions scheduling algorithm. *Health Serv Res*, 5(4):308–319, 1970.

[13] M. Davio. Kronecker products and shuffle algebra. *IEEE Transactions on Computers*, 30(2):116–125, 1981.

[14] R.W. Derlet, J.R. Richards, and R.L. Kravitz. Frequent overcrowding in u.s. emergency departments. *Acad Emerg Med*, 8(2):151–155, 2001.

[15] G. Dobson, S. Hasija, and E.J. Pinker. Reserving capacity for urgent patients in primary care. *Production and Operations Management.*, 2010. Forthcoming.

[16] M.A. Draeger. An emergency department simulation model used to evaluate alternative nurse staffing and patient population scenarios. In J.J. Swain, D. Goldsman, R.C. Crain, and J.R. Wilson, editors, *Proc. of the 1992 Winter Simulation Conference*, pages 1592–1600, Piscataway, New Jersey, 1992. Institute of Electrical and Electronics Engineers, Inc.

[17] R.G. Dunn. Scheduling elective admissions. *Health Serv Res*, 2(2):181–215, 1967.

[18] DM Fatovich, Y. Nagree, and P. Sprivulis. Access block causes emergency department overcrowding and ambulance diversion in Perth, Western Australia. *BMJ*, 22(5):351–354, 2005.

[19] G.B. Folland. *Real Analysis: Modern Techniques and Applications*. Wiley, New York, 1999.

[20] A.J. Forster, I. Stiell, G. Wells, et al. The effect of hospital occupancy on emergency department length of stay and patient disposition. *Acad Emerg Med*, 10(2):127–133, 2003.

[21] S. Gallivan and M. Utley. Modelling admissions booking of elective in-patients into a treatment centre. *IMA Journal of Management Mathematics*, 16(3):305–315, 2005.

[22] S. Gallivan, M. Utley, T. Treasure, and O. Valencia. Booked inpatient admissions and hospital capacity: mathematical modelling study. *BMJ*, 324(7332):280–282, 2002.

[23] Y. Gerchak, D. Gupta, and M. Henig. Reservation planning for elective surgery under uncertain demand for emergency surgery. *Management Science*, 42(3):321–334, 1996.

[24] Jan Grandell. Point processes and random measures. *Adv Appl Probab*, 9(3):502–526, 1977.

[25] Linda V. Green, Sergei Savin, and Ben Wang. Managing patient service in a diagnostic medical facility. *Operations Research*, 54(1):11–25, 2006.

[26] J.R. Griffith, W.M. Hancock, and F.C. Munson. *Cost Control in Hospitals*. Health Administration Press, Ann Arbor, MI, 1976.

[27] D. Gupta. Surgical suites operations management. *Production and Operations Management*, 16(6):689–700, 2007.

[28] D. Gupta and B. Denton. Appointment scheduling in health care: Challenges and opportunities. *IIE Transactions*, 40:800–819, 2008.

[29] A.Y. Ha. Optimal dynamic scheduling policy for a make-to-stock production system. *Operations Research*, 45(1):42–53, 1997.

[30] R. Hall, D. Belson, P. Murali, and M. Dessouky. Modeling patient flows through the healthcare system. In R. Hall, editor, *Patient Flow: Reducing Delay in Healthcare Delivery*, volume 91 of *International Series in Operations Research & Management Science*, pages 1–44. Springer, New York, 2006.

[31] J. Han, C. Zhou, and D.J. France. The effect of emergency department expansion on emergency dpartment overcrowding. *Academic Emergency Medicine*, 14:338–343, 2007.

[32] W.M. Hancock and P.F. Walter. The use of computer simulation to develop hospital systems. *SIGSIM Simul. Dig.*, 10(4):28–32, 1979.

[33] W.M. Hancock and P.F. Walter. *The "ASCS": Inpatient Admission Scheduling and Control System*. Ann Arbor, MI, 1983.

[34] C. Haraden and R. Resar. Patient flow in hospitals: Understanding and controlling it better. *Frontiers of Health Services Management*, 20(4):3–15, 2004.

[35] P.R Harper and A.K. Shahani. Modelling for the planning and management of bed capacities in hospitals. *J Oper Res Soc*, 53(1):11–18, 2002.

[36] G.W. Harrison, A. Shafer, and M. Mackay. Modelling variability in hospital bed occupancy. *Health Care Manag Sci*, 8(4):325–334, 2005.

[37] Merrit Hawkins. 2010 physician inpatient/outpatient revenue survey. http://www.merritthawkins.com/pdf/2010_revenuesurvey.pdf, 2011.

[38] J.E. Helm, A. Adisusanto, A. Chrzanowski, J. Pan, and M.P. Van Oyen. Hillsmoother: A hospital census forecasting and bed management tool. Working Paper. Department of IOE, University of Michigan, 2011.

[39] J.E. Helm, S. AhmadBeygi, and M.P. Van Oyen. The flexible patient flow simulation framework. In *Proceedings of the 2009 Industrial Engineering Research Conference*, Vancouver, British Columbia, Canada, May 2008. Institute of Industrial Engineers.

[40] J.E. Helm, S. AhmadBeygi, and M.P. Van Oyen. The flexible patient flow simulation framework. In *Proceedings 2009 IIE IERC Conf*, 2009.

[41] J.E. Helm, S. AhmadBeygi, and M.P. Van Oyen. Design and analysis of hospital admission control for operational effectiveness. *Production and Operations Management*, 20(3):359–374, 2011.

[42] J.E. Helm, M. Lapp, and B.D. See. Characterizing an effective hopsital admission scheduling and control management system: A genetic algorithm approach. In *WSC '10: Proceedings 42nd Conf on Winter Sim*, pages 2387–2398, 2010.

[43] J.E. Helm and M.P. Van Oyen. Design and optimization methods for elective hospital admissions. Technical Report 11-01. Deptartment of IOE, Univeristy of Michigan, Ann Arbor, MI, 2010.

[44] J.E. Helm and M.P. Van Oyen. Infinite horizon analysis of a hospital admission control model. Technical Report 10-01. Department of IOE, Univeristy of Michigan, Ann Arbor, MI, 2010.

[45] J.E. Helm, M.P. Van Oyen, and A. Adisusanto. Census smoothing through hospital admission control (presentation). In *21st Annual POMS Conf*, 2010.

[46] Jonathan E. Helm and Mark P. Van Oyen. Design and optimization methods for elective hospital admissions. Submitted to Operations Research, 2011.

[47] N.R. Hoot and D. Aronsky. Systematic review of emergency department crowding: Causes, effects, and solutions. *Ann Emerg Med*, 55(2):126–136, 2008.

[48] M.W. Isken. Modeling and analysis of occupancy data: A healthcare capacity planning application. *Int J Inf Tech Decis*, 1(4):707–729, 2002.

[49] S. Jacobson, S. Hall, and J.R. Swisher. Discrete-event simulation of health care systems. In R. Hall, editor, *Patient Flow: Reducing Delay in Healthcare Delivery*, volume 91 of *International Series in Operations Research & Management Science*, pages 211–252. Springer, New York, 2006.

[50] N.L. Johnson, S. Kotz, and N. Balakrishnan. *Discrete multivariate distributions*, volume 157. Wiley, New York, 1997.

[51] J.B. Jun, S.H. Jacobson, and J.R. Swisher. Application of discrete-event simulation in health care clinics: A survey. *J Oper Res Soc*, 50(2):109–123, 1999.

[52] J.B. Jun, S.H. Jacobson, and J.R. Swisher. Application of discrete-event simulation in health care clinics: A survey. *The Journal of the Operational Research Society*, 50(2):109–123, 1999.

[53] R.L. Kane, T. Shamliyan, C. Mueller, S. Duval, and T.J. Wilt. Nurse staffing and quality of patient care. *Evid Rep Technol Assess (Full Rep)*, 151:1–115, 2007.

[54] S. Keehan, A. Sisko, and C Truffer. Expenses for hospital inpatient stays: 2004. *AHRQ, Statistical Brief*, 164, 2007.

[55] K.J. Klassen 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.

[56] G. Koole. Structural results for the control of queueing systems using event-based dynamic programming. *Queueing Systems*, 30(3):323–339, 1998.

[57] M. Kroneman and J.J. Siegers. The effect of hospital bed reduction on the use of beds: A comparative study of 10 european countries. *Social Science & Medicine*, 59(8):1731 – 1740, 2004.

[58] G. Latouche and V. Ramaswami. *Introduction to Matrix Analytic Methods in Stochastic Modelling, 1st edition. Chapter 2: PH Distributions*. Society for Industrial and Applied Mathematics, Philadelphia, PA, 1999.

[59] K.K. Leung, W.A. Massey, and W. Whitt. Traffic models for wireless communication networks. *Selected Areas in Communications, IEEE Journal on*, 12(8):1353–1364, 1994.

[60] T. Lim, D. Uyeno, and I. Vertinsky. Hospital Admissions Systems: A Simulation Approach. *Simulation Gaming*, 6(2):188–201, 1975.

[61] S.A. Lippman. Applying a new device in the optimization of exponential queuing systems. *Operations Research*, 23(4):687–710, 1975.

[62] Y. Liu and W. Whitt. A network of time-varying many-server fluid queues with customer abandonment. *Submitted to Operations Research*, 2010.

[63] J.C. Lowery. Design of hospital admissions scheduling system using simulation. In *Proceedings of the 28th conference on Winter simulation*, pages 1199–1204, Coronado, California, USA, December 1996. IEEE Computer Society.

[64] K. Machlin, S.R. Carper. Expenses for hospital inpatient stays, 2004. In *Statistical Brief # 164*. Agency for Healthcare Research and Quality, 2007.

[65] W.A. Massey and W. Whitt. Networks of infinite-server queues with nonstationary Poisson input. *Queueing Systems*, 13(1):183–250, 1993.

[66] W.A. Massey and W. Whitt. A stochastic model to capture space and time dynamics in wireless communication systems. *Probability in the Engineering and Informational Sciences*, 8(04):541–569, 1994.

[67] W.A. Massey and W. Whitt. An analysis of the modified offered-load approximation for the nonstationary erlang loss model. *The Annals of Applied Probability*, 4(4):1145–1160, 1994.

[68] J.H. May, W.E. Spangler, D.P. Strum, and L.G. Vargas. The surgical scheduling problem: Current research and future opportunities. *Production and Operations Management.*, 2010. Forthcoming.

[69] M.L. McManus, M.C. Long, A. Cooper, J. Mandell, D.M. Berwick, M. Pagano, and E. Litvak. Variability in surgical caseload and access to intensive care services. *Anesthesiology*, 98(6):1491–1496, 2003.

[70] F.C Munson and W.M. Hancock. Problems of implementing change in two hospital settings. *IIE Transactions*, 4:256–266, 1972.

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

[72] L. Ou, L. Young, J. Chen, N. Santiano, L.S. Baramy, and K. Hillman. Discharge Delay in Acute Care: Reasons and Determinants of Delay in General Ward Patients. *Aust Health Rev*, 33(3):513–521, 2009.

[73] J. Patrick and M.L. Puterman. Dynamic multipriority patient scheduling for a diagnostic resource. Technical Report. Sauder School of Business, University of British Columbia, 2006.

[74] Jonathan Patrick, Martin L. Puterman, and Maurice Queyranne. Dynamic multipriority patient scheduling for a diagnostic resource. *Operations Research*, 56(6):1507–1525, 2008.

[75] C. Price, B. Golden, M. Harrington, R. Konewko, E. Wasil, and W. Herring. Reducing boarding in a post-anesthesia care unit. *Production and Operations Management.*, 2010. Forthcoming.

[76] N.C. Proudlove, S. Black, and A. Fletcher. Or and the challenge to improve the nhs: modelling for insight and improvement in in-patient flows. *J Oper Res Soc*, 58(2):145–158, 2007.

[77] NC Proudlove, K. Gordon, and R. Boaden. Can good bed management solve the overcrowding in accident and emergency departments? *BMJ*, 20(2):149–155, 2003.

[78] Becker's Hospital Revue. 100 surgery center benchmarks. http://www.beckershospitalreview.com/asc-turnarounds/100-surgery-center-benchmarks.html, September 2011.

[79] D.B. Richardson. Increase in patient mortality at 10 days associated with emergency department overcrowding. *Med J Australia*, 184(5):213–216, 2006.

[80] R.G. Sargent. Verification and validation of simulation models. In *Proceedings of the 37th conference on Winter simulation*, pages 130–143, Orlando, FL, USA, December 2005. IEEE Computer Society.

[81] B.D. See, S.-P. Liu, Y.-W. Lu, and Q. Pang. Staffing a pandemic urgent care facility during an outbreak of pandemic influenza. In M.D. Rossetti, R.R. Hill, B. Johansson, A. Dunkin, and R.G. Ingalls, editors, *Proc. of the 2009 Winter Simulation Conference*, pages 1996–2007, Piscataway, New Jersey, 2009. Institute of Electrical and Electronics Engineers, Inc.

[82] W. Shonick and J.R. Jackson. An improved stochastic model for occupancy-related random variables in general-acute hospitals. *Operations Research*, 21(4):952–965, 1973.

[83] P.C. Sprivulis, J Da Silva, I.G. Jacobs, et al. The association between hospital overcrowding and mortality among patients admitted via Western Australian emergency departments. *Med J Australia*, 184:208–212, 2006.

[84] P.C. Sprivulis, J. Da Silva, I.G. Jacobs, A.R. Frazer, and G.A. Jelinek. The association between hospital overcrowding and mortality among patients admitted via Western Australian emergency departments. *Medical Journal of Australia*, 184(5):208–212, 2006.

[85] D.L. White, C.M. Froehle, and K.J. Klassen. The effect of integrated scheduling and capacity policies on clinical efficiency. *Production and Operations Management.*, 2010. Forthcoming.

[86] D. Worthington. Hospital waiting list management models. *Journal of the Operational Research Society*, 42(10):833–843, 1991.

[87] B. Zhang, P. Murali, M.M. Dessouky, and D Belson. A mixed integer programming approach for allocating operating room capacity. *Journal of the Operational Research Society*, pages 1–11, 2008.