Design and Optimization Methods for Elective Hospital Admissions

Jonathan E. Helm
Mark P. Van Oyen

Technical Report 11-01
February 21, 2010

University of Michigan
Industrial and Operations Engineering
Ann Arbor, Michigan 48109
Design and Optimization Methods for Elective Hospital Admissions

Abstract: Hospitals in the US typically lack accurate enterprise level management of bed and care resources, which contributes to bed census levels that are statistically “out of control.” This system dysfunction manifests itself in the delay of treatment for emergency patients due to lack of beds (bed block), cancelation of scheduled surgeries, denial of admission for elective medical patients, ambulance diversions and operational chaos within the hospital. One of the key drivers of this variability in census is the unstable manner in which elective admissions are scheduled. Elective admissions exhibit significant variation both by day of week and from week to week. To reduce patient blockages and system congestion while maintaining throughput, elective admissions must be stabilized and coordinated with the other hospital subsystems, such as the emergency room and hospital beds. This paper details the design and optimization of a hospital admissions management system to coordinate hospital systems and stabilize patient flows to allow the hospital to maintain or increase throughput with less congestion and fewer blockages. To do so we develop an enterprise-wide stochastic patient flow model and an associated optimization model that can identify hospital admission policies that smooth hospital census over time. In particular, we formulate a new Poisson-arrival-location (PALM) model based on an innovative patient stochastic location that we develop called the Patient Temporal Resource Requirements (PATTERN) model. We further extend the PALM approach to the class of deterministic-arrival-location models (DALM) endowed with our PATTERN location process. This work provides the theoretical foundations of an efficient admissions management system as well as a practical decision support methodology to coordinate elective admissions with the emergency department and hospital beds.

Keywords: Hospital Admissions Management, Stochastic Patient Flow Modeling, Mixed Integer Programming, Census Smoothing

1 Introduction
1.1 Consequences of Census Variability

Hospital care services are subject to significant, unnecessary and detrimental fluctuations in patient census and associated workload. This variability in patient census impacts cost,
access, quality and safety in healthcare delivery. Variability in 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 Intensive Care Unit (ICU) and Post Acute Care Unit (PACU). Studies have confirmed that system-wide congestion results in compromised quality of care, emergency patient diversions for lack of beds, excessive patient Length of Stay (LOS), and significant excess costs [24, 31, 9, 17, 26, 29, 30]. The high census variability experienced by most hospitals is illustrated by the census time series in Figure 1 (a).

![Census variation over time](image1)

(a) Census variation over time

![Weekly occupancy pattern](image2)

(b) Weekly occupancy pattern

Figure 1: Controlling census variability in hospitals.

In addition to the considerable variation in census from day to day and week to week shown in Figure 1 (a), most hospitals also exhibit a weekly pattern with a census spike in the middle of the week followed by a sharp drop in census around the weekends (see 1 (b)). This weekly census spike or “hump” causes most hospitals to become overcrowded in the middle of the week despite an average census level (the dotted line in Figure 1 (b)) well below peak census. The frequent periods of peak census / overcrowding contribute significantly to one of the major systemic problems in hospitals: “bed block.”

Bed block describes the situation where a patient requires an inpatient bed but is blocked from being transferred into the bed because all the beds are full. Bed block significantly impacts emergency patients and emergency department performance as well as elective inpatients and transfers from other hospitals. When bed block occurs emergency patients are forced to remain in the emergency department or are placed in the hallway until a bed becomes available, contributing to emergency department overcrowding [10, 27, 12]. When emergency departments become overcrowded, patient wait times increase dramatically and
the emergency department experiences a significantly increased rate of accidents and mortalities [31].

The cyclic census spike, Figure 2 (a), and the ramifications in terms of unnecessary patient blockages and cancelations in the middle of the week, Figure 2(b), are demonstrated with the solid line based on data from one of our partner hospitals. The dotted line in these figures graphically demonstrates the potential benefit of smoothing census across the week. The two census plots exhibit the same average daily census; however, the smoothed census benefits from significantly reduced blockages. This paper proposes mechanisms that guide hospital policies to achieve such a smooth census level.

Figure 2: Benefits of census smoothing based on simulation using historical hospital data.

Quality of care also suffers in hospitals with high census variability. Census variability leads to highly variable workloads not only for the nursing staff but for most other members of the workforce, including ancillary services such as labs, radiology, and physical therapy. Overloaded nursing staff has been linked to mortality, nurse burnout and job dissatisfaction [2]. Nursing can utilize float nursing pools and other “chase” staffing strategies, but quality of care suffers and staff are often dissatisfied [5, 23]. Variable workloads often create large backlogs in radiology and ancillary services that can delay diagnosis, treatment and discharge of a patient [28]. Census variability has also been linked directly to increased LOS and/or worsening patient disposition [12] and even increased mortality rate [31, 30].

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. While this flexibility helps mitigate some of the variability in workload, it also leads to complex interactions with unintended consequences and patient safety is not well served by placing them “off-unit” [3]. For example, the decision to perform too many elective surgeries can lead to overcrowding in the Emergency
Department as surgical patients often overflow into medicine wards, blocking emergency department patients, for whom the medicine ward is the primary destination (see \[19\]).

While hospitals endeavor to limit the amount of congestion and blockage, they are also driven by a competing need to have high utilization of their expensive fixed resources. According to the Department of Health and Human Services (TK cite), the per diem for a typical hospital bed is $1,817 per day. Further, hospitals with higher throughput can provide better access to their community by serving more patients. The key to efficient hospital management is to enable high throughput while limiting the number of blockages. Smoothing the census levels over time is one method to achieve this goal. To smooth hospital census the hospital must address both (1) the weekly census “hump” that causes overcrowding in the middle of the week despite an average census level well below peak census and (2) the variability in admissions from week to week.

1.2 Causes of Census Variability

It is well known that both a weekly pattern of variation in elective admissions (see Figure 3 (a)) and the week to week variation in number of elective admissions on a given day (see Figure 3 (b)) significantly contribute to both the weekly census hump and the week to week variation in census and consequently to hospital congestion and patient blockages \[4\]. The weekly census “hump” is primarily generated by an elective admission pattern that is heavily front loaded early in the week and tapers off toward the end of the week as can be seen in Figure Figure 3 (a). The variability from week to week can be attributed to the significant variation in the number of elective admissions scheduled on a given day from week to week. Figure 3 (b) demonstrates that, on a given Monday over the course of a year, the number of elective admissions ranges from 8 to 170 with a wide range of admission values realized.

The highly non-stationary weekly pattern in elective admissions is due in part to the fact that hospitals generally do not perform elective admissions on the weekends. Because few elective admissions occur over the weekend, the census drops, which then leads the hospital to admit large numbers of elective patients early in the week to take advantage of the extra capacity in order to meet weekly throughput targets. Determining how “large” an amount is currently left up to individual surgical services and informed intuition and represents a major opportunity for modeling and optimization.

Because we know that each day of the week has, on average, significantly different vol-
ume, the week to week variability that generates significant noise around the mean census levels is described by measuring the number of admissions on each day. Table 1 provides specific measures to demonstrate the magnitude of variability in elective admissions by day of week. It may be shocking to note that elective admissions actually exhibit more variability (higher standard deviation) during the weekdays than emergency admissions. The fact that elective admissions are more variable than emergency arrivals is a well known phenomenon and, because it contributes significantly to census variability, must necessarily be addressed in our approach to census smoothing.

<table>
<thead>
<tr>
<th>Category</th>
<th>DOW</th>
<th>Arr. Std. Dev</th>
<th>Arr. Mean</th>
<th>Arr. CV</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Emergency</td>
<td>Elective</td>
<td>Emergency</td>
</tr>
<tr>
<td>Hospital Total</td>
<td>Sun</td>
<td>8.44</td>
<td>5.51</td>
<td>48.23</td>
</tr>
<tr>
<td>Hospital Total</td>
<td>Mon</td>
<td>13.23</td>
<td>32.98</td>
<td>64.79</td>
</tr>
<tr>
<td>Hospital Total</td>
<td>Tue</td>
<td>11.64</td>
<td>17.98</td>
<td>62.17</td>
</tr>
<tr>
<td>Hospital Total</td>
<td>Wed</td>
<td>10.59</td>
<td>23.88</td>
<td>57.53</td>
</tr>
<tr>
<td>Hospital Total</td>
<td>Thu</td>
<td>13.89</td>
<td>28.93</td>
<td>58.02</td>
</tr>
<tr>
<td>Hospital Total</td>
<td>Fri</td>
<td>10.96</td>
<td>20.49</td>
<td>64.79</td>
</tr>
<tr>
<td>Hospital Total</td>
<td>Sat</td>
<td>8.94</td>
<td>4.76</td>
<td>52.69</td>
</tr>
</tbody>
</table>

Table 1: Variation in Elective and Emergency Admissions

The variability in elective admissions from week to week seen in Figure 3 (b) and Table 1 primarily results from the fact that elective admissions are typically scheduled in a decentralized manner, with each service scheduling independently without visibility to the effect of their scheduling decisions on the downstream hospital resources. This is a likely cause for the fact that the daily number of elective admissions often appears to follow a Poisson distribution (with a different mean each day of the week) and indeed are often more
variable than emergency admissions in many hospitals [17]. As a complement to smoothing the weekly census hump, we investigate the benefits of centralizing the admission decision process and setting daily admission targets to reduce this week to week variability in elective admissions.

The fact that hospitals operate at reduced capacity on weekends is one of the main reasons that it is difficult to smooth hospital census. Properly managing admissions in the presence of cyclically varying loading and patient stays that span weeks and multiple hospital services requires a detailed analysis of the flow and workload that admitted patients place on hospital resources over the entire course of their stay. Another challenge to census smoothing lies in coordinating the typically siloed elective admission decisions. We develop a model that can link the temporal resource requirements to admission decisions as an effective coordination mechanism. The specific question that we address is how many patients to admit by day of week to achieve a census level that is as smooth as possible.

2 Smoothing Hospital Census

Non-smooth hospital census patterns typically take two forms: (1) non-stationary mid-week census spikes and (2) widely varying unstable census levels from week to week. The key to smoothing 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 measure of ward census levels over time for any particular mix and volume of patient admissions.

Prior literature establishes the importance of census levels and census variability to admission decision making. For example, [8] developed early stochastic models to link admissions decisions with hospital census. In a more recent study, [17] used a multistage stochastic approach to establish that variability in daily hospital census in combination with high census levels can increase the risk of hospital overflows.

Extensive research also demonstrates the effectiveness of patient flow management and admissions in particular as a tool for improving hospital care delivery. According to [22], an effective patient flow model can enable high patient throughput, low patient wait times, short LOS, and low clinic overtime. [16] developed and implemented a heuristic simulation-
based inpatient admissions scheduling and control system to achieve high average census subject to constraints on the number of cancelations and emergency patient blockages. Our work engages the same problem; however, because our methods are analytical rather than simulation-based, we are able to employ a non-heuristic optimization that solves much more quickly and can guarantee the optimality of the resulting admission schedules.

Recently, the relationship between elective admissions and census levels has been studied with the purpose of developing improved elective admissions schedules. [13, 14, 7, 1, 4], have all studied the impact of elective admissions on census levels in various wards. These papers, however, focus solely on a single ward or isolated subset of hospital resources. In each case, network dynamics were either not modeled or were assumed to take the form of a simple feed forward network. The scope of our work not only includes modeling the entire hospital, full patient care trajectories, and census levels by ward, it also includes the far more realistic generalized network dynamics of the hospital wards and the use of flexible wards to serve patients off-unit.

To capture the true hospital dynamics we model the hospital as a general network of interacting wards/units, considering the two primary types of interaction between wards that have not been 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. As an example of the first type of interaction, consider a patient who arrives for surgery and is initially placed in an ICU bed for recovery. Then when the patient’s condition improves they transfer to a surgery bed for the remainder of their stay. For the second type of interaction between wards, consider a surgical patient that leaves the operating room to find all the surgical beds (the patient’s preferred ward) full. The hospital would then place the patient “off-unit” in a medicine ward.

While these interactions between wards complicate the modeling of patient flows and subsequent elective admission schedule optimization, ignoring these interactions leads to omitting critical dynamics of hospital system functioning. Since patients often transfer between wards as their condition changes, each patient can place a load on more than one ward of the hospital over the course of their stay. In our partner hospital 56% of patients transfer wards at least once during their hospital stay and patients who transfer on average transfer 1.6 times per visit. Considering only the first ward, or a feed-forward subset of resources ignores a significant load that patients place on other hospital resources.
Considering off-unit interactions is important for several reasons. First, while it is feasible to place patients off-unit it is not desirable for the patient or the hospital. Off-unit patients are often not being cared for by professionals who specialize in their condition as they would if they were on their preferred ward. The nursing skills differ from ward to ward, so the mismatch when a patient is placed off their preferred ward detracts from quality of care (including safety) as well as nurse satisfaction. To understand this, note that nurses work hard to master every aspect and nuance of caring for the types of conditions assigned to their ward. An off-unit patient brings them out of their specialty and comfort zone, which causes the nurses considerable job stress [5]. Research has also shown that patients placed off their preferred ward have more bottlenecks to discharge, which increases their length of stay [3].

Secondly, the fact that patients use resources that are not a part of their preferred treatment path leads to consequences that may not be immediately obvious but nevertheless must be considered in order to reach the potential of a fully coordinated hospital. One pertinent example, which is also quite common in US hospitals due to the heavily front-loaded elective surgery schedule, demonstrates how myopic surgical scheduling can actually cause ambulance diversions and increase emergency department overcrowding. Consider that on Monday and Tuesday a large number of elective surgeries are planned so that the hospital’s surgical ward fills up. This triggers the off-unit mechanism, sending the overflowing surgical patients to the medicine ward that then also fills up. Since the medicine ward is the primary destination for patients being transferred to the hospital from the emergency department, emergency patients are blocked by the surgical patients that have filled the medicine ward as off-unit patients. This “bed block” causes the emergency patients being blocked to remain in the emergency department, taking up critical ED resources and contributing to ED overcrowding and ambulance diversions [31]. Analogously, bed block affects elective patients, causing cancelation of elective surgeries due to lack of beds.

Clearly the consequences of using the off-unit mechanism can be serious, but how prevalent is the use of the off-unit mechanism? In fact, the percent of off-unit patients in a hospital is often quite significant. One hospital we have worked with reserves around 17% of their total beds for patients that cannot enter their preferred ward due to lack of beds. For the above reasons, without modeling ward-to-ward transfers and off-unit mechanisms, critical hospital system dynamics are omitted and the hospital system as a whole cannot be
optimized.

In all of the previous research, the models that have considered total system flow are descriptive in nature and they lack an optimization component. On the other hand, the existing optimization models fail to consider complete patient trajectories through a general network of hospital wards and the off-unit interactions that occur in flexible wards. A primary contribution of this paper 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.

3 Characterization of the Stochastic Census Process

Figure 4 demonstrates 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 obtained from the network patient flow model to inform elective admission decisions while accounting for the resource requirements of the emergency patients. By coordinating elective admissions with the rest of the hospital (including ED and ward beds) we show that it is possible to determine the volume and mix of elective patients that will generate a consistent, stable workload and avoid blockages and cancelations.

Figure 4: A Simplified Sketch of a Conceptual Patient Flow Model.
To create an optimization framework from the conceptual model of Figure 4, we begin by developing a stochastic model of patient flow through the network of hospital wards. As a basic building block for such a model, 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.). Specifically, we build a model of patient flow that determines the temporal resource requirements for each patient type. That is, we generate a probabilistic map of the resources (beds) used by a patient of a given type over their entire stay in the hospital. Figure 7 shows the expected load a cardiology patient places on hospital wards over the course of their treatment.

![Partial Individual Care Path for Cardiology Patients](image)

Figure 5: Expected workload generated across the hospital by a cardiology patient

We then use these type-specific building blocks to characterize the elective census level in each ward for a given elective schedule. Next, the elective census process is combined with the emergency census process to characterize the total census level in each ward over time for a given elective admission schedule. 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.

3.1 System Design and Assumptions

To provide the clearest possible presentation, we begin by clarifying some of the modeling assumptions and the perspective of the approach. First we elaborate on the perspective of
the approach to modeling patient resource requirements. In the following sections we develop a stochastic process where element $t$ of the process represents the resources (beds) required to serve a patient $t$ time units after admission to the hospital. The key that makes the approach computationally tractable is that we do not follow each patient as they transition from one ward to another over time. Instead, we look at each time period after admission independent of where the patient was in prior time units and we characterize the probability distribution over the wards for that time period.

For example, imagine a patient is admitted and then we observe where the patient is $t$ time periods in the future without observing any part of the patient’s path up to that point. The stochastic process we develop characterizes the distribution on what ward (if any) the patient is in $t$ time units after admission. Mathematically, this means that we assume each element of the stochastic process is independent of previous elements. This does not mean, however, that the data used to calculate the distribution are uncorrelated.

Another key assumption for our modeling approach is that an admission plan is developed 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 consider a weekly cycle. Doctors usually have fixed clinic times, OR time, research time, etc. each week and so a repeating elective admission schedule fits well within the practical constraints of the hospital environment.

Another assumption, which is more a design element of our elective admission scheduling system, is that the admission schedule is followed exactly. That is, the number of elective patients to be scheduled is considered deterministic. As noted in Section 1.2, one of the primary causes of census variability is the extreme variability in the number of elective admissions from week to week. Figure 3 underscores the extent to which elective admissions deviate from the mean volumes from one week to the next. Overcoming this instability in scheduling requires organizational change to enforce operating discipline on the admissions decision makers, which can be difficult but is a critical component of our elective admissions system design. While determinism will not occur exactly 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, 16, 13]).

A further assumption we make is that the care path of each patient is independent of
other patients that have been admitted to the hospital. This is a mild assumption when one considers our approach to census modeling. Specifically, we initially develop models of demand for hospital services without considering the capacity required to serve those patients. 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. Of course the historical data we work with to parameterize our models comes from hospitals with finite capacity so our care paths are in fact approximations of an uncapacitated care path; however, we show in Section 3.6 that these approximations are in fact quite accurate in modeling an actual hospital’s census. Thus except in the case of infectious disease it seems reasonable that one patient would not impact the health or care requirements of another. In the case of an infectious disease, we assume that hospitals have taken the appropriate precautions to isolate such patients and prevent spread to other hospital patients.

Finally, we anticipate our elective admissions system being used more often at a daily granularity because this gives more flexibility and decision making power to admission planners and thus is more likely to increase acceptance. Therefore, in many of the examples that follow we present the elective admission system in terms of days, though the modeling approach is significantly more general.

3.2 Care Pathways – a PATTERN Stochastic Location Process Model

To understand the effect that scheduling decisions have on census level over time across the network of hospital wards, first 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 [25]. Let the state space be $S = \{1, \ldots, M \cdot n, \Delta^*, \Delta_*\}$, where state $i$ represents that the patient is currently in ward $i$, state $\Delta^*$ represents the state where the patient has left the hospital (i.e. discharged) and $\Delta_*$ represents the state where the patient has not yet arrived to the hospital. Patients move through the state space according to the $S$-valued stochastic location process model $\{L_s(t) : r \in \mathbb{R}\}$, where $s$ is the arrival time and $t > s$ is the time of interest. For notational convenience we let $S = S^0 \cup \{\Delta^*, \Delta_*\}$, so that $S^0$ 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 3.1.** 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.

To characterize the stochastic location process, let $\Sigma$ be the function space of right-continuous functions with left limits that represents the set of all possible sample paths of the stochastic location process $L_s(t)$. An element $\sigma \in \Sigma$ is a (deterministic) mapping $\sigma : \mathbb{R} \rightarrow S$ such that $\sigma(t)$ represents the location of the patient at time $t$. Figure 6 represents three different sample path functions. The solid line represents path $\sigma_1(t)$, a sample path of the process $L_{s_1}(t)$, the dashed line represents the path $\sigma_2(t)$, a sample path of the process $L_{s_2}(t)$, and the dotted line represents the path $\sigma_3(t)$, a sample path of the process $L_{s_3}(t)$. Path $\sigma_2(t)$, for example, represents a patient who arrives at $s_2$ to ward 1, then transfers to ward 2 for a brief stay and then back to ward 1 before being discharged slightly before time $t$. Note that a location function $\sigma \in \Sigma$ is a right-continuous step function that takes values in $S^0$ 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 $t \geq T_s$.

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 if $\Gamma_s \subseteq \Sigma$ is the set of all location functions that enter the hospital, $S^0$, at time $s$, then $P_s(\Gamma_s) = 1$ and $P_s(\Gamma_t) = 0$ for $t \neq s$.

We will demonstrate how this measure is used to find the probability that a patient is in ward $w$ at time $t$, given that they arrived to 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 the set of location functions that describe whether the patient is

![Figure 6: Emergency patient sample paths](image)
in ward \( w \) at some time \( t \). The measure of this set is the probability of being in ward \( w \) at time \( t \). To this end, define

\[
\Gamma_{t,w} = \left\{ \sigma_s \in \Sigma : \sigma_s(t') = \Delta_s \text{ for } t' < s < t, \exists T_s < \infty \text{ s.t. } \sigma_s(t') = \Delta^* t \geq T_s, \sigma_s(t') \in S^0 \text{ for } t' \in [s, T_s) \right\}.
\] (3.1)

This subset of \( \Sigma \) is essentially the set of all location functions that place a patient in ward \( w \) 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. 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 propose to use the design approach proposed in the first paragraph of Section 3.1. Each patient type will have their own stochastic location model, so for patient type \( k \)

\[
P_{s,k}(\Gamma_{t,w}) = p_{s,k,w}(t - s) \tag{3.2}
\]

\[
P_{s,k}(\Gamma_{t,\Delta^*}) = p_{s,k,\Delta^*}(t - s) = 1 - \sum_{j=1}^{W} p_{s,k,j}(t - s) \tag{3.3}
\]

where \( p_{s,k,w}(t) \) is the probability that a patient of type \( k \) who arrives at time \( s \) is in ward \( w \), \( t \) time units after admission to the hospital. These probabilities are easily obtained from historical data [18]. An example of these probabilities for discrete time points for cardiology patients is shown in Table 2. 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 \( t \).

### 3.3 The Emergency Census Process

There are several ways to characterize the emergency census process. [13] characterize the process in the same manner as the elective process. It is possible to adapt this approach to
Table 2: Patient Temporal Resource Requirements (PATTERN) Matrix for a cardiology patient

<table>
<thead>
<tr>
<th>Ward</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>\ldots</th>
</tr>
</thead>
<tbody>
<tr>
<td>A3</td>
<td>45.92%</td>
<td>37.54%</td>
<td>21.22%</td>
<td>19.16%</td>
<td>17.03%</td>
<td>\ldots</td>
</tr>
<tr>
<td>C20</td>
<td>6.17%</td>
<td>0.14%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>\ldots</td>
</tr>
<tr>
<td>CCU</td>
<td>7.10%</td>
<td>3.48%</td>
<td>2.34%</td>
<td>1.14%</td>
<td>0.92%</td>
<td>\ldots</td>
</tr>
<tr>
<td>ICU</td>
<td>0.14%</td>
<td>0.14%</td>
<td>0.14%</td>
<td>0.07%</td>
<td>0.07%</td>
<td>\ldots</td>
</tr>
</tbody>
</table>

To achieve this highly detailed description of the emergency census process, 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. Figure 7 depicts the queueing approach we take to characterize the emergency flow. Since we are first modeling just the demand for services (as in Section 3.5), 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 (see Figure 7). In queueing, this is denoted by $\left(M_t/G_t/\infty\right)^N/G_t$.

It has been shown that the non-stationary Poisson Process is a good model for emergency patient arrivals ([17]), 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 $\mathbb{W}_i(t) = Q_1^i(t) + Q_2^i(t) + \cdots + Q_n^i(t)$ represent the amount of emergency patient demand for ward $i$ at time $t$, where $Q_j^i(t)$ is the demand of type $j$ patients for ward $i$.

Analogously, $Q(t) = \sum_{i=1}^{M} \sum_{j=1}^{n} Q_j^i(t)$ 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.

### 3.4 PATTERN Poisson-arrival-location model (PALM)

To specify the PATTERN PALM model for the emergency census process we rely on the Poisson random measure approach proposed by [25]. 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 3.2. 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 [15]. The Poisson random measure is defined by it’s intensity measure, so we start by defining the space that the intensity measure lies in as

**Definition 3.1.** Let $\mathbb{M}$ be the set of measures $\mu$ on $\mathbb{R}^+$ such that $\mu(0) = 0$ and $\mu(t) < \infty$
for all \( t \in \mathbb{R}^+ \).

It is important for our analysis that a metric can be defined in \( M \) that makes \( M \) separable and complete [15]. Let \( \mathcal{B}(M) \) be the Borel \( \sigma \)-algebra on \( M \). To describe point processes, it is necessary to also define

**Definition 3.2.** Let \( N = \{ \mu \in M : \mu(t) \in \mathbb{Z}^+ \text{ for all } t \in [0, \infty) \} \) be the set of measures \( \mu \in M \) that yield integer values.

The space \( N \) 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 3.3.** A random measure is a measurable mapping from the probability space \((\Omega, \mathcal{B}, \mathbb{P})\) into the measure space \((M, \mathcal{B}(M))\), or \((N, \mathcal{B}(N))\) as the situation requires.

We characterize our PATTERN PALM Poisson random measure by extension of the standard Poisson random measure, \( \hat{M} \), which we now define as in [15]. Let \( \Omega = \mathbb{R}^\infty \) and an element \( \omega \in \Omega \) is defined by the process’ interarrival times: \( \omega = (s_0, s_1, s_2, \ldots) \). Let \( \mathcal{B} \) be the natural \( \sigma \)-algebra on \( \mathbb{R}^\infty \) and \( \mathbb{P} \) is defined as

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

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

\[
\hat{M}(t, \omega) = \begin{cases} 
k & \text{if } \sum_{j=0}^{k-1} s_j \leq t, \sum_{j=0}^{k} s_j > t \text{ and } \sum_{j=0}^{\infty} s_j = \infty \\
0 & \text{if } s_0 > t \text{ or if } \sum_{j=0}^{\infty} s_j < \infty
\end{cases}
\]  

Let the distribution of the standard Poisson random measure be \( \hat{\Pi} \), which is a probability measure on \((N, \mathcal{B}(N))\) such that for any set \( A \in \mathcal{B}(N) \), \( \hat{\Pi}\{A\} = \mathbb{P}\{\hat{M} \in A\} \). The standard Poisson random measure can be generalized by scaling time according to an intensity measure, \( \mu \). By letting \( \hat{M} \circ \mu = \hat{M}(\mu) : N \times M \to N \) be a product measure with distribution \( \Pi_\mu \), we obtain a random measure model for a Poisson process with intensity \( \mu \). Note if we let \( \mu(t) = \lambda t \) then we get the traditional Poisson process with rate \( \lambda \). In our case, however, we are interested in allowing the intensity itself to also be a realization of a random measure.

In particular, we want the intensity to represent the arrivals to the various wards of the hospital over time under a total arrival process that is non-homogeneous Poisson. We will specify the PATTERN PALM intensity measure shortly, but first it is important to finish
the characterization of the product Poisson random measure, or what is also called a *doubly stochastic Poisson process* \[15\].

Let \( M = \hat{M} \circ \Lambda \) be a Poisson random measure on a product space \( \mathbb{N} \times \mathcal{M} \), where the intensity random measure \( \Lambda \) has distribution \( \Pi \). As before the standard Poisson random measure, \( \hat{M} \), has distribution \( \hat{\Pi} \). The distribution of \( M \) is then specified by

\[
P_\Pi(B) = \hat{\Pi} \times \Pi(\{(\nu, \mu) \in \mathbb{N} \times \mathcal{M} : \nu \circ \mu \in B\})
\]

\[
= \int_{\mathcal{M}} \hat{\Pi}(\{\nu \in \mathbb{N} : \nu \circ \mu \in B\}) \Pi(d\mu)
\]

\[
= \int_{\mathcal{M}} \Pi_\mu(B) \Pi(d\mu) \quad \text{for all } B \in \mathcal{B}(\mathbb{N}),
\]

which follows from Fubini’s theorem \[11\]. It is important to note that \( \Pi_\mu(B) \) is a \( \mathcal{B}(\mathcal{M}) \)-measurable function in \( \mu \), so the preceding integral is valid \[15\]. All that remains now is to define our PATTERN PALM intensity measure and verify its properties.

To specify the intensity of the Poisson random measure, we begin by specifying the random measure on the location function space \( \Sigma \) in a similar manner to the definition of the standard Poisson random measure. That is we define a mapping from the probability space \((\Omega, \mathcal{B}, \mathbb{P})\) into the measure space \((\mathcal{M}, \mathcal{B}(\mathcal{M}))\). Here we let \( \Omega = \Sigma \), which we defined in Section 3.2. For patient type \( i \) let

\[
\mathbb{P}_i(\sigma \in \Sigma : \sigma(t) = j, \sigma(s) \in S^0, \text{ and } \sigma(t') = \Delta_* \text{ for } t' < s) \equiv P_{s,i}(\sigma \in \Sigma : \sigma(t) = j) = \begin{cases} 0 & \text{if } t < s \\ p_{s,i,j}(t - s) & \text{if } t \geq s \end{cases}
\]

(3.7)

The random location measure for a patient of type \( i \), called \( \Lambda_i \), is then specified by

\[
\Lambda_i(t, j, \sigma) = \begin{cases} 1 & \text{if } \sigma(t) = j \\ 0 & \text{otherwise} \end{cases}
\]

(3.8)

Now we can specify the intensity measure for patients of type \( i \) by combining the non-homogeneous Poisson arrival process with nonnegative deterministic external-arrival-rate function \( \alpha_i(\cdot) \) with the location random measure from Equations 3.7 and 3.8. The arrival rate function, \( \alpha_i(\cdot) \), drives the type \( i \) 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, \( L_{s,i}(t) \) with dynamics driven by the probability measure \( P_{s,i}(\cdot) \). Therefore, the for any \( B_i = (a_i, b_i] \times \Gamma_i \), where \( \Gamma_i \subseteq \Sigma \) the set of location functions and \( (a_i, b_i] \) is arrival interval, the intensity measure is given by
\[
\mu((a_i, b_i] \times \Gamma_i) = \int_{a_i}^{b_i} \alpha_i(s) P_{s,i}(\Gamma_i) ds.
\] (3.9)

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 split by the probability that the patient is in some location some time in the future. Combining the intensity measure with the standard Poisson random measure we can obtain the PATTERN Poisson random measure \(M\), which can be shown to have distribution:

\[
P(M(B_1) = n_1, M(B_2) = n_2, \ldots, M(B_k) = n_k) = \prod_{i=1}^{k} \frac{e^{-\gamma_i} \gamma_i^{n_i}}{n_i!}
\] (3.10)

where \(B_i = (a_i, b_i] \times \Gamma_i\) and

\[
\gamma_i \equiv E[M(B_i)] = \int_{B_i} \mu(ds,d\sigma) = \mu((a_i, b_i] \times \Gamma_i) = \int_{a_i}^{b_i} \alpha_i(s) P_{s,i}(\Gamma) ds
\] (3.11)

In words, the set \(B_i\) represents the patients arriving to the hospital on interval \([a, b)\) and the location of those patients at some time in the future. For purposes of the integration in equation 3.11, it is important to note that \(P_{s,i}\) is a measure on \(\Sigma\) for each \(s\) and is a measurable function of \(s\) for each measurable subset \(\Gamma_i \subseteq \Sigma\). This is the general formulation for any arrival, location model. In our model, we are specifically interested in the distribution on the number of emergency patients in a given ward at a given time. The following example illustrates how the approach would work in practice.

Suppose we want to understand the distribution on the number of emergency patients in ward \(i\) on a Monday. In this case we want to consider the cyclostationary system in steady state, where the arrival pattern is repeated on a weekly basis. If we let \(N_i\) be the maximum length of stay for a patient of type \(i\) then

**Theorem 3.1.** The number of emergency patients in each ward, denoted by \(Q_1(t), Q_2(t), \ldots, Q_M(t)\), are independent Poisson random variables for each \(d\) with finite mean given by

\[
m_j(t) = \sum_{i=1}^{n} \int_{k=t-N_i}^{t} \alpha_i(s) p_{s,i,j}(t-s)
\] (3.12)

where \(B_{d,w} = (\infty, d + 1) \times \Gamma_{d,w}\) represents all arrivals prior to time \(d\) that are in ward \(w\) at time \(d\) and \(\Gamma_{d,w}\) is as defined in 3.1.
Proof. If we endow the function space $\Sigma$ with the Skorohod $J_1$ topology (as in [25]) then $\Sigma$ 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 can be clearly bounded. The location functions are easily seen to be right continuous, with limits from the left. The result then follows directly from the properties of the Poisson random measure and the fact that we are considering disjoint subsets of $\mathbb{R} \times \Sigma$.

3.5 The Elective Census Process: $X(\Theta)_t$

The approach for the elective census model represents an extension of the PALM methodology to processes with deterministic arrivals, or what we called deterministic-arrival-location model (DALM). In this approach arrivals occur at specific times (possibly in batches), rather than according to a Poisson distribution. Once a patient of type $i$ has arrived at time $s$, they still flow through the hospital according to their PATTERN stochastic location process $L_{s,i}$ as in Section 3.2.

3.5.1 Elective Census Model

Combining the PATTERN model for individual patients with the elective admission schedule, $\Theta$, 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 3.4. For patients of type $k$ Let $\{(t_{k,1}, \Theta_{k,t_{k,1}}), (t_{k,2}, \Theta_{k,t_{k,2}}), \ldots\}$ represent the deterministic arrival times, $t_{k,i}$ and number of patients scheduled $\Theta_{k,t_{k,i}}$. Let $\Omega = \Sigma^\infty$ so that $\omega_k = \{\sigma_{k,t_{k,1},1}, \sigma_{k,t_{k,1},2}, \ldots, \sigma_{k,t_{k,1},\Theta_{k,t_{k,1}}}, \sigma_{k,t_{k,2},1}, \ldots\} \in \Omega$ represents the set of location functions for the scheduled arrivals. We define probability measure for patients of type $k$ as

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

(3.13)

Then we can define the DALM point process as

$$N_{k,i,\Theta}(t, \omega) = \begin{cases} \sum_{s \in \{t_{k,i}: t_{k,i} < t\}} \sum_{n=1}^{\Theta_{k,s,n}} \Lambda_j(s, t, \sigma_{k,s,n}) & \text{if } t_1 < t \\ 0 & \text{if } t_1 > t \end{cases}$$

(3.14)

where $\Lambda_j(\cdot)$ is the random measure defined for the stochastic location process in Section 3.4. It is easy to see that this point process can be written instead as the process
\[
N_{k,w,\Theta}(t) = \sum_{s \in \{t_{k,i} : t_{k,i} < t\}} \sum_{j=1}^{\Theta_{k,s}} \mathbb{1}\{L_{s,k}(t) = w\},
\]

(3.15)

where \(X_{k,w,\Theta}(t)\) is the number of elective patients of type \(k\) in ward \(w\) at time \(t\) under schedule \(\Theta\). We will work with this more convenient form to analyze the DALM process. The ward level workload can be calculated by summing over patient types and the hospital workload can be calculated by summing over all the wards. However, now consider the system design assumption that the hospital follow a cyclically repeating elective admission schedule. Further, consider the case where the hospital is concerned with daily measures of admissions and census. We analyze this case (though the approach will work in general) as it seems particularly useful for managerial insight and operational planning.

Using Equation 3.15 the total system workload (census), \(L_{d_1}\), can be calculated on a given day \(d_1\) of the planning horizon. First define \(\mathfrak{W}\) as the set of hospital wards and \(\mathfrak{D}\) as the set of patient types (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 by for a finite horizon of length \(t\) weeks (\(C_{d_1}^t\) from equation 3.16) or an infinite horizon (\(C_{d_1}^\infty\) from equation 3.17),

\[
C_{d_1}^t = \sum_{u \in \mathfrak{W}} \sum_{d_2=1}^{7} \sum_{i \in \mathfrak{D}} \sum_{j=0}^{t} \sum_{n=0}^{\Theta_{i,d_2}} \mathbb{1}\{L_{d_2+7n,i}^{j,n}(d_1 + 7t)) = u\}
\]

(3.16)

\[
C_{d_1}^\infty = \lim_{t \to \infty} L_{d_1}^t,
\]

(3.17)

where \(L_{s,i}^{j,n}(\cdot)\) represents the \((i, j)\)th independent, identically distributed instance of the location process \(L_{s,i}(\cdot)\), one process for each admission. For completeness, we adopt the convention that a negative index of the random variable will return the value zero and let \(\mathbb{1}\{\cdot\}\) be the indicator random variable. In equations 3.16 and 3.17, the first sum represents the load presented by each hospital ward. The second sum refers to the day of the week that the patient was admitted. The third sum refers to the diagnosis of the patient and the fourth 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 iterates through weeks (or through cycles of the planning horizon).

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 \(\mathbb{1}\{L_{s,CAR}^{j,k}(t) = c\}\) represent
whether the $(j,k)$ indexed cardiology patient is in the cardiology ward $c$ on day $t$ given they were admitted on day $s$. As a thought experiment start the system on Monday by admitting two cardiology patients (call them patient 1.0 and 2.0). This leads to a workload for Tuesday of the first week ($n = 0$) of $\mathbb{I}\{L_{1,\text{CAR}}^{1.0}(2) = c\} + \mathbb{I}\{L_{1,\text{CAR}}^{2.0}(2) = c\}$. Note that although $\mathbb{I}\{L_{1,\text{CAR}}^{1.0}(2) = c\}$ and $\mathbb{I}\{L_{1,\text{CAR}}^{2.0}(2) = c\}$ are identically distributed random variables, they are assumed to be independent 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 a week prior may also still be in the hospital (and thus on day 8 of their length of stay) the workload for the Tuesday of the second week ($n = 1$) is $\mathbb{I}\{L_{1,\text{CAR}}^{1.0}(9) = c\} + \mathbb{I}\{L_{1,\text{CAR}}^{2.0}(9) = c\} + \mathbb{I}\{L_{8,\text{CAR}}^{1.1}(9) = c\} + \mathbb{I}\{L_{8,\text{CAR}}^{2.1}(9) = c\}$. If we let the system run for $t$ weeks, then the workload on the Tuesday of week $t$ is given by

$$
\sum_{n=0}^{t} \mathbb{I}\{L_{7n+1,\text{CAR}}^{1,n}(7t + 2) = c\} + \mathbb{I}\{L_{7n+1,\text{CAR}}^{2,n}(7t + 2) = c\}.
$$

This shows how we build up the workload profile for equations 3.16 and 3.17. We are primarily interested in the steady state behavior of the system, and thus rely mostly on the infinite horizon formulation 3.17 in the analysis that follows.

### 3.5.2 Moments of the Elective Census Process

A useful feature of the above formulation of the elective census process is that it is possible to calculate the first and second moments of the process analytically. The mean can be calculated using the linearity of expectation. If we 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 by taking the expectation inside the sum,

$$
\mu_{d_1,u}(\Theta) = \mathbb{E}\left[\sum_{d_2=1}^{7} \sum_{i \in \mathcal{D}} \sum_{j=0}^{\Theta_{i,d_2}} \lim_{t \to \infty} \sum_{n=0}^{t} \mathbb{I}\{L_{d_2+7n,i}^{j,n}(d_1 + 7t) = u\}\right] = \sum_{d_2=1}^{7} \sum_{i \in \mathcal{D}} \Theta_{i,d_2} \cdot \lim_{t \to \infty} \sum_{n=0}^{t} p_{d_2+7n,i,u}(d_1 - d_2 + 7(t - n)). \tag{3.18}
$$

The equality follows from the fact that $\mathbb{I}\{X = x_k\}$ follows a Bernoulli distribution and thus $\mathbb{E}[\mathbb{I}\{X = x_k\}] = p_k$. The mean census level in the hospital can be calculated by summing equation 3.18 over all $u$ as in $\sum_{u \in \mathcal{W}} \mu_{d_1,u}$.

Next consider the variance of the elective census process. The two types of variance we consider are (1) the variance in ward census and (2) the variance in total hospital census,
which differ primarily in their covariance matrix. First, consider the covariance of any two indicator variables in the sum in equation 3.18 (the equation for ward census). It can be shown that the covariance be

\[
Cov(\mathbb{I}\{L_{s_1,k_1}^{i_1,j_1}(t) = u_1\}, \mathbb{I}\{L_{s_2,k_2}^{i_2,j_2}(t) = u_2\}) =
\]

\[
E[\mathbb{I}\{L_{s_1,k_1}^{i_1,j_1}(t) = u_1\} \mathbb{I}\{L_{s_2,k_2}^{i_2,j_2}(t) = u_2\}] - E[\mathbb{I}\{L_{s_1,k_1}^{i_1,j_1}(t) = u_1\}] E[\mathbb{I}\{L_{s_2,k_2}^{i_2,j_2}(t) = u_2\}]
\]

\[
P(L_{s_1,k_1}^{i_1,j_1}(t) = u_1, L_{s_2,k_2}^{i_2,j_2}(t) = u_2) - P(L_{s_1,k_1}^{i_1,j_1}(t) = u_1) P(L_{s_2,k_2}^{i_2,j_2}(t) = u_2) = 0. \quad (3.19)
\]

The final equality follows because we are considering bernoulli random variables for two different (and therefore independent) patients. It is natural to assume that the condition of one patient in the hospital should not affect the condition of another. 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, \(j\), and their admission number on the day they are admitted, \(i\). Because the sum in equation 3.17 contains each combination of \((i, j, k)\) at most once, each pair of terms in the sum must necessarily belong to the stochastic process from different patients. Therefore the covariance is zero and so the variance of the sum is the sum of the variances:

\[
\sigma^2_{d_1,n}(\Theta) = \text{Var} \left[ \sum_{d_2=1}^{7} \sum_{i \in \mathcal{D}} \sum_{j=0}^{\theta_i,d_2} \lim_{t \to \infty} \sum_{n=0}^{t} \mathbb{I}\{L_{d_2+7n,i}^{i_1,j_1}(d_1 + 7t) = u\} \right]
\]

\[
= \sum_{d_2=1}^{7} \sum_{i \in \mathcal{D}} \Theta_{i,d_2}.
\]

\[
\lim_{t \to \infty} \sum_{n=0}^{t} p_{d_2+7n,i,u}(d_1 - d_2 + 7(t - n))(1 - p_{d_2+7n,i,u}(d_1 - d_2 + 7(t - n))). \quad (3.20)
\]

The second equality follows by taking the variance inside the sum and from the fact that \(\mathbb{I}\{\cdot\}\) is a Bernoulli random variable with variance \(p(1-p)\). Due to the linearity of the variance calculation (covariance of zero) \(\sigma^2_{d_1}(\Theta)\) can be written as a linear function of the admission plan (decision) \(\Theta\). This is important if we want to include the variance of the process in an integer programming framework for determining optimal schedules.

Now consider the variance around total hospital census. For all pairs of wards \((i_1, j_1, k_1, s_1) \neq (i_2, j_2, k_2, s_2)\), the indicators \(\mathbb{I}\{L_{s_1,k_1}^{i_1,j_1}(t) = u_1\}\) and \(\mathbb{I}\{L_{s_2,k_2}^{i_2,j_2}(t) = u_2\}\) are independent as in equation 3.19. 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 the same patient is contributing to ward \(u_1\) or to ward \(u_2\) on a given day). In this situation, the two
indicators are obviously not independent, since if a patient is in ward \( u_1 \) they are clearly not in ward \( u_2 \neq u_1 \). For such pairs the covariance becomes

\[
\text{Cov}(\mathbb{1}\{L_{s,k}^{t,j}(t) = u_1\}, \mathbb{1}\{L_{s,k}^{t,j}(t) = u_2\}) =
\]

\[
\mathbb{E}[\mathbb{1}\{L_{s,k}^{t,j}(t) = u_1\}, \mathbb{1}\{L_{s,k}^{t,j}(t) = u_2\}] - \mathbb{E}[\mathbb{1}\{L_{s,k}^{t,j}(t) = u_1\}] \mathbb{E}[\mathbb{1}\{L_{s,k}^{t,j}(t) = u_2\}]
\]

\[
P(L_{s,k}^{t,j}(t) = u_1, L_{s,k}^{t,j}(t) = u_2) - P(L_{s,k}^{t,j}(t) = u_1)P(L_{s,k}^{t,j}(t) = u_2) = -p_{s,k,u_1}(t)p_{s,k,u_2}(t).
\]

(3.21)

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. Therefore the total hospital census variance is

\[
\sigma_d^2_1(\Theta) = \sum_{u \in \mathcal{W}} \sigma_d^2_{1,u}(\Theta) - \sum_{d_2=1}^7 \sum_{i \in \mathcal{D}} \lim_{t \to \infty} \sum_{n=0}^t \sum_{u_1 \neq u_2} p_{s,k,u_1}(t)p_{s,k,u_2}(t).
\]

(3.22)

This follows directly from the covariance equation 3.21. Fortunately, the variance and covariance terms can be calculated offline and enter the optimization as data and since equation 3.22 is still linear in terms of our decision variables \( \Theta_{i,d} \), the model remains solvable by traditional MIP solution approaches.

3.6 How Good is the Approximation?

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 from Sections 3.5 and 3.3. 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. Figure 8 and Table 3 show the mean census levels by day of week for the entire hospital for both the approximation and for the historical census levels.

Figure 8: Comparison of the Mean Census Approximation vs Historical Mean Census
The slight deviations between the approximation and the actual census, overestimating the census from Friday to Sunday and underestimating the census during the week, can be explained by day-specific hospital practices that are not captured in our aggregate estimates of patient behavior. Most hospitals operate at reduced resource levels on the weekend so the discharge rate on Fridays is often higher than during the rest of the week. The reduced weekend resource levels also mean fewer discharges on the weekends. While it is possible to allocate a separate PATTERN stochastic process for each day of the week to account for this phenomenon, the quantity of data available for this study was not deemed sufficient for such an analysis. Thus our census approximation does not specifically account for this change in discharge rate by day of week, instead assuming a time-stationary discharge process. This explains underestimating the weekend census and overestimating the weekday census. The deviations, however, are relatively small and have little effect on accurately approximating system metrics such as cancelations and blockages as shown in Section 4.3.

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. The approximation developed in this and previous sections can be easily incorporated into an integer programming optimization model as demonstrated in section 4.

4 Elective Admissions Mix and Volume Optimization

In Section 3 we built and analyzed a model for forecasting patient demand under a given admission plan. In this section we design an integer programming model to determine the optimal schedule given a set of criteria. For our criteria, we consider two opposing forces 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 amount of emergency and elective patient blockages. The stochastic process from Section 3 characterizes the raw
demand for beds, so to quantify the blockages we need to superimpose the hospital capacity on this model. Section 4.1 presents a method for calculating various blockage criteria in a manner that can be incorporated into an integer programming formulation. Section 4.2 presents two different formulations for the elective admissions mix and volume that could be useful to hospitals. Section 4.3 compares the forecasted census from the optimization model with a high fidelity simulation of hospital operations to quantify the accuracy of the approximations we develop for tractability of the optimization model.

4.1 System Effectiveness Criteria

This section presents a method for calculating two criteria for system effectiveness associated with hospital congestion. First we present a method for calculating expected blockages under a given admission plan. Then we extend this method to calculate expected off-unit census. Limiting both patient blockage and off-unit census is critical to operating a well functioning hospital as detailed in Section 1.1 and Section 2.

To calculate expected blockages it is important to consider both the elective census and the emergency census in the hospital. From Section 3.3, it is possible to approximate the entire distribution on the number of emergency patients in each ward (and for the entire hospital) using the PATTERN PALM model, but characterizing the elective census is more involved. Thus we break the total census process, into its elective and emergency process components. 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 function. The census is then calculated by summing indicators the PATTERN process. In particular, the census on a given day, \(d_1\), can be rewritten as:

\[
C_{d_1,u}^\infty = \sum_{d_2=1}^{7} \sum_{i \in \mathcal{D}} \lim_{t \to \infty} \sum_{n=0}^{\Theta_i,d_2} \sum_{j=0}^{1} \mathbb{1}\{L_{d_2+\gamma_{n,i}}^j(d_1 + 7t)) = u\} \tag{4.23}
\]

To investigate the distribution of \(C_{d_1,u}^\infty\), first consider that for fixed \(i, d_2, u, \) and \(n\), \(\sum_{j=0}^{\Theta_i,d_2} \mathbb{1}\{L_{d_2+\gamma_{n,i}}^j(d_1 + 7t)) = u\}\) is a sum of \(\Theta_i,d_2\) i.i.d. Bernoulli random variables, which is Binomial\((\Theta_i,d_2, p_{d_2+\gamma_{n,i},u}(d_1 - d_2 + 7(t - n)))\). 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\((\Theta_i,d_2, p_{d_2+\gamma_{n,i},u}(d_1 - d_2 + 7(t - n)))\) where

\[
p_{d_2+\gamma_{n,i}}(d_1 - d_2 + 7(t - n)) = \left[p_{d_2+\gamma_{n,i,1}}(d_1 - d_2 + 7(t - n)), \ldots, p_{d_2+\gamma_{n,i,W}}(d_1 - d_2 + 7(t - n))\right].
\]
is the probability vector of the categorical random variable defining the whether or not the patient will reside in the various hospital wards at a given time [21]. So the census random variable, $C_{d_i}^{\infty}$, defined in equation 4.23 is just the sum of many different Binomial (or Multinomial if we consider the joint process) random variables. 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 $\Theta$ takes the factorial operator, thus linearity is not preserved. In addition, the sums of various Multinomial or Binomial random variables with different distributions is not likely to have a nice form.

Another approach is to consider 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 3.5.2. The Poisson approach also suffers from the fact that it would employ the decision variable $\Theta$ in a non-linear manner. A more direct approach would be to use the fact that a sum of independent (but not identically distributed) Bernoulli random variables follows a Poisson-Binomial distribution as follows. If $Z_1, \ldots, Z_n$ are independent Bernoulli trials with probability $p_i$ of success, then $S_Z = Z_1 + \cdots + Z_N$ is a Poisson-Binomial random variable with distribution:

$$P(S_Z = n) = \left\{ \prod_{i=1}^{N} (1 - p_i) \right\} \sum_{w_{i_1}, \ldots, w_{i_n}} w_{i_1} \cdots w_{i_n}$$

where $w_i = p_i/(1 - p_i)$ and the sum is over all possible combinations (and orderings) of successful trials [6]. Again notice that this would require significant non-linearities in the decision variables of the MIP.

Due to the complications involved in working with distributions or variances of the elective census in an optimization framework, we propose the following heuristic approximation to obtain estimates of the expected blockages and off-unit census, which is illustrated in Figure 9. The approach begins by calculating the mean elective census by day of week, indicated by the solid bar in Figure 9 (which we justify below). Starting with the mean census as a baseline, we add in 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. Blockages are tallied when the number of emergency patients plus the mean number of elective patients
exceeds the hospital capacity, $K$. 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 4.2. Further, the design of the system as described in Section 3.1 ensures that we minimize the elective census variability by stabilizing the elective admissions. It is reasonable to assume, therefore, that the majority of the census variability is now coming from the emergency patients, which we capture fully with the emergency census distribution. In Section 4.3, we demonstrate the accuracy of this approximation on data from a partner hospital. The reasonable accuracy suggested by the testing described in section 4.3 indicates that more complicated approaches incorporating variance in elective census may not be necessary in light of the tradeoff between tractability/solution speed in using such approaches.

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 for a given ward that exceeds the ward’s capacity must necessarily be considered off-unit.

### 4.2 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.
Sets
\[ \mathcal{D} \quad \text{set of all patient diagnosis types} \]
\[ \mathcal{W} \quad \text{set of hospital wards} \]

Parameters
- \( B_i \): Ward \( i \) capacity in terms of beds
- \( \beta \): limit on the average number of blockages per week
- \( \alpha_u \): the 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 \( u \)
- \( p_{d_1}^{i,u} \): Probability that an elective patient of type \( i \) is in ward \( u \) \( d_1 \) days after admission
- \( \hat{p}_n^u \): Probability there are \( n \) emergency patients in ward \( u \) from the PATTERN PALM model
- \( \hat{\bar{p}}_n \): Probability there are \( n \) emergency patients in the hospital from the PATTERN PALM model
- \( \theta_{i,d} \): Represents the current elective admission mix and volume by day.
- \( \hat{\theta}_{i,d} \): The maximum number of elective admissions of type \( i \) allowed on day \( d \).

Decision Variables
- \( \Theta_{i,d} \): Number of type \( i \) patients scheduled on day \( d \)
- \( \delta_{n,d} \): the number by which \( n \) emergency patients will exceed the expected reserved capacity on day \( d \)
- \( \hat{\delta}_n^u \): the number by which \( n \) emergency patients will exceed the capacity of ward \( u \) on day \( d \), thereby triggering the use of off-unit capacity

It is important to note here that the probabilities, \( p_{d_1}^{i,u}, \hat{p}_n^u, \) and \( \hat{\bar{p}}_n \) are all calculated offline per the analysis in Sections 3.5 and 3.3 and then become data inputs to the mixed integer program that follows.

Maximum Elective Admissions Formulation

First we present a formulation that maximizes the number of elective admissions subject to constraints on bed blockage.
\[
\max \Theta \cdot R \cdot \mathbf{1} \quad (4.25)
\]

s.t.
\[
\delta_{j,d_1} \geq j - \sum_{u \in W} \left( B_u - \sum_{d_2=1}^7 \Theta_{i,d_2} \cdot \sum_{n=0}^\infty p_{i,u}^{d_1-d_2} \right) + 1 \quad \forall j, d_1 = 1, \ldots, 7 \quad (4.26)
\]
\[
\sum_{d=0}^7 \sum_{n=0}^\infty \tilde{p}_n \delta_{n,d} \leq \beta \quad (4.27)
\]
\[
\delta_{n+1,d} \geq \delta_{n,d} \quad \forall d \in D, \forall n = 1, 2, \ldots \quad (4.28)
\]
\[
\hat{\delta}_{j,d_1} \geq j + \sum_{d_2=1}^7 \sum_{i \in D} \Theta_{i,d_2} \cdot \sum_{n=0}^\infty p_{i,u}^{d_1-d_2} - B_u - \alpha_u \sum_{d=0}^7 \sum_{n=0}^\infty \delta_{n,d} \cdot \tilde{p}_n \quad (4.29)
\]
\[
\sum_{n=0}^\infty \tilde{p}_n \delta_{n,d} \leq \hat{\beta}_u \quad \forall u \in W, d_1 = 1, \ldots, 7 \quad (4.30)
\]
\[
\hat{\delta}_{n+1,d} \geq \hat{\delta}_{n,d} \quad \forall d \in D, \forall n = 1, 2, \ldots \quad (4.31)
\]
\[
\sum_{d=1}^7 \Theta_{i,d} \geq \sum_{d=1}^7 \theta_{i,d} \quad \forall i \in D \quad (4.32)
\]
\[
\Theta_{i,d} \leq \hat{\theta}_{i,d} \quad \forall i \in D, d = 1, \ldots, 7 \quad (4.33)
\]
\[
\Theta_{i,d}, \delta_{i,d}, \hat{\delta}_{i,d} \in \mathbb{Z}^+ \quad (4.34)
\]

The objective function (4.25) maximizes the weighted throughput of elective patients. Constraint 4.26 manages the 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 \( \sum_{u \in W} B_u \), which represents the capacity of the hospital. Next, \( \sum_{n=0}^\infty p_{i,u}^{d_1-d_2} \) represents the steady state expected bedload contribution of a single patient of type \( i \) to ward \( u \) on day \( d_1 \) of the cycle given that they were admitted on day \( d_2 \) of the cycle. Recall that \( \mathbf{E} \left[ \sum_{n=0}^\infty X_{d_1-d_2}^{i,u} \right] = \sum_{n=0}^\infty p_{i,u}^{d_1-d_2} \). Since expectation is additive we get that the total mean bedload contributed by elective admissions to the hospital on day \( d_1 \) is simply:

\[
\mathbf{E} \left[ \sum_{u \in W} \sum_{d_2=1}^7 \sum_{i \in D} \sum_{j=0}^\infty \sum_{n=0}^\infty X_{d_1-d_2}^{i,u} \right] = \sum_{u \in W} \sum_{d_2=1}^7 \sum_{i \in D} \sum_{n=0}^\infty \Theta_{i,d_2} \cdot p_{i,u}^{d_1-d_2} \quad (4.35)
\]

where \( X_{d,j}^{i,u} \) is the \( j \)th instance of an i.i.d. process describing the bed requirements of a
patient of type $i$ in ward $u$. The equality holds because expectation is additive so we can instead multiply by the number of patients of type $i$ admitted on day $d_2$, $\Theta_{i,d_2}$. Equation 4.35 essentially describes the expected bedload contribution of each day of the cycle to day $d_1$ and then sums over all wards to get total expected hospital bedload on day $d_1$. When we subtract Equation 4.35 (the expected elective bedload on the hospital on day $d_1$) from $\sum_{u \in W} B_u$ (total hospital capacity), the balance is the expected remaining capacity that can be used by emergency patients. Call this quantity $E[RC]$. 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 capacity reserved for those patients. 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 $\delta_{j,d_1}$ 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 $\delta_{j,d_1}$ can be set to 0.

Constraint 4.27 is the constraint that approximates the expected number of weekly blockages for given schedule $\Theta$ and limits it to at most $\beta$. The method to obtain this approximation is detailed in Section 4.1.

Constraint 4.28 is a cut that was added to the model to increase the speed of the CPLEX branch and bound algorithm. Because of the large number of $\delta$ decision variables, this cut greatly reduces the number of combinations that must be considered by branch and bound but doesn’t eliminate any feasible solutions to the original problem we want to consider. Prior to adding this constraint, a model with three wards and three patient types failed to solve in under 24 hours. After adding this constraint, the same model solved in under 30 seconds.

Constraints 4.29 - 4.31 serve the same function for measuring and limiting expected off-unit census as Constraints 4.26 - 4.28 do for expected blockages. The one piece that is different is the addition of the term $\alpha_u \sum_{d=0}^7 \sum_{n=0}^{\infty} \delta_{n,d} \cdot \tilde{p}_n$ in Equation 4.26. 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. In this equation the parameter $\alpha_u$ 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 4.32 and 4.33, represent the reality that the model
should not change the elective admission schedule in ways incommensurate with historical hospital practice. Specifically Equation 4.32 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. Equation 4.33 ensures that the model respects capacity constraints outside of 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. \( \Theta_{i,\text{Sunday}} \leq 0 \) \( \forall i \)).

**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.

\[
\min_{\Theta} \sum_{d=0}^{7} \sum_{n=0}^{\infty} \tilde{p}_n \delta_{n,d} \leq \beta \tag{4.36}
\]

s.t.

\[
\delta_{j,d_1} \geq j - \sum_{u \in \mathcal{W}} \left( B_u - \sum_{d_2=1}^{7} \sum_{i \in \mathcal{D}} \Theta_{i,d_2} \cdot \sum_{n=0}^{\infty} \tilde{p}_{i,u} \left( 7n + d_1 - d_2 \right) \right) + 1 \quad \forall j, d_1 = 1, \ldots, 7 \tag{4.37}
\]

\[
\delta_{n+1,d} \geq \delta_{n,d} \quad \forall d \in \mathcal{D}, \forall n = 1, 2, \ldots \tag{4.38}
\]

\[
\hat{\delta}_{j,d_1} \geq j + \sum_{d_2=1}^{7} \sum_{i \in \mathcal{D}} \Theta_{i,d_2} \cdot \sum_{n=0}^{\infty} \tilde{p}_{i,u} \left( 7n + d_1 - d_2 \right) - B_u - \alpha_u \sum_{d=0}^{7} \sum_{n=0}^{\infty} \delta_{n,d} \cdot \tilde{p}_n \quad \forall j, u \in \mathcal{W}, d_1 = 1, \ldots, 7 \tag{4.39}
\]

\[
\sum_{n=0}^{\infty} \tilde{p}_n \delta_{n,d} \leq \hat{\beta}_u \quad \forall u \in \mathcal{W}, d = 1, \ldots, 7 \tag{4.40}
\]

\[
\hat{\delta}_{n+1,d} \geq \hat{\delta}_{n,d} \quad \forall d \in \mathcal{D}, \forall n = 1, 2, \ldots \tag{4.41}
\]

\[
\sum_{d=1}^{7} \Theta_{i,d} = \sum_{d=1}^{7} \theta_{i,d} \quad \forall i \in \mathcal{D} \tag{4.42}
\]

\[
\Theta_{i,d} \leq \hat{\theta}_{i,d} \quad \forall i \in \mathcal{D}, d = 1, \ldots, 7 \tag{4.43}
\]

\[
\Theta_{i,d}, \delta_{i,d}, \hat{\delta}_{i,d} \in \mathbb{Z}^+ \tag{4.44}
\]

The main difference in this formulation is the objective function and the fact that we set the weekly volume strictly equal to the current weekly volume in Equation 4.42.
4.3 How Good is the Approximation?

As in Section 3.6, 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.

A year’s worth of historical hospital data was used to calibrate both the optimization and simulation models for a core subset of hospital wards. The nine (out of 22 total) modeled wards included medicine, surgical and ICU/CCU wards. The wards were selected as a representative subset that included significant network flow interactions, which are a critical feature of our modeling innovation.

Figure 10: Expected Blockage Constraint Illustration.

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

Table 4: Simulation output vs stochastic model output for characteristic hospital measures

Figure 10 and Table 4 confirm that the stochastic census model is a robust approximation of actual census levels and that the blockage formulation in the optimization model is also an accurate approximation of actual patient blockages. The small bias toward higher census levels that can be seen in the approximation 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. Although blockages are calculated, the blocked patients are not removed from the demand calculations. Thus the depression of census in the simulation versus the census approximation. 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% and the census differs on average by only 1%.

Because the stochastic census model is such an accurate and robust approximation, the detailed (and therefore slow) simulation is no longer needed to express the tradeoffs between census and blockages that we want to capture to design effective admission schedules.

5 Case Study and Practical Results

To demonstrate the effectiveness and potential uses of our approach to elective admissions scheduling, we present a case study based on the historical hospital data used in other sections for model validation. The hospital is a medium sized, non-teaching hospital and as in Section 4.3 we model the primary interacting 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 (midnight) 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 wards of the hospital 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 the major patient blockage, both elective cancelations and emergency patient bed block, without reducing the number of patients served. The wards chosen for modeling admitted 90 elective inpatients per week on average. To achieve this goal the minimum blockage formulation is employed, constraining the weekly elective volume to equal 90 and also constraining the volumes on each admitting service to match the current level. The optimization generated an optimal schedule matching these criteria, which we then simulated (for completeness) to compare
with the current schedule. The result was a 32% reduction in average cancelations per week as shown in “Minimum Blockages” of Table 5.

<table>
<thead>
<tr>
<th>Schedule</th>
<th>Elective Adm per Wk</th>
<th>Blockages per Wk</th>
</tr>
</thead>
<tbody>
<tr>
<td>Current</td>
<td>90</td>
<td>3.29</td>
</tr>
<tr>
<td>Minimum Blockages</td>
<td>90</td>
<td>2.34</td>
</tr>
<tr>
<td>Maximum Admissions</td>
<td>96</td>
<td>3.27</td>
</tr>
</tbody>
</table>

Table 5: Comparison of Elective Admissions and Blockages

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 access and quality of care. 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. We also add a constraint to ensure that each service is given at least as many elective admissions as in the current schedule. The result, “Maximum Admissions” in Table 5, is an additional 310 elective admission per year (six per week on average) with a slightly better access (3.27 blockages per week).

These prescriptive models are useful in exploring the boundaries of hospital efficiency, but the “extreme” solutions may not suit all hospitals from a practical standpoint. Certainly there are scheduling solutions that can achieve a balance between volume and blockage and our model can find them. The Pareto curve in Figure 11 presents the tradeoff between elective admission volume and blockages. Notice that current schedule is above the Pareto curve, so it can be improved by increasing admissions, decreasing blockages, or both.

![Elective Admissions vs Blockages](image)

Figure 11: Pareto tradeoff of elective volume and patient blockages.

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 the area of admission scheduling. It enables hospital administrators to understand the key tradeoffs involved in scheduling their admissions and gives them the freedom to choose where they want to be on the curve. Providing hospitals information and allowing them to make the decision is a far more useful decision making tool than a model that simply prescribes a solution. This enables the hospitals to make the decisions that are in their domain while giving them the best information and capability to implement those decisions.

This Pareto curve also represents an advance in the basic science of admission scheduling. The simulation work in this area that included the general network effects would likely struggle to produce such a figure, while the optimization models were able to generate the schedules automatically in a matter of minutes. Further, existing optimization models in this area lack the scope to capture the general network effects and accurately model the entire hospital.

6 Conclusions and Future Work

The hospital admission scheduling problem has been approached in many ways; however, previous approaches have not been able to efficiently generate optimal schedules for the entire hospital. The simulation approaches, on the one hand, capture the critical general network effects for hospital wards but lack a clear optimization method for generating improved schedules. 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 to true load on hospital wards. Our modeling approach has bridged this gap by accurately capturing the census and blockage dynamics in an analytical manner, eliminating the need for simulation and enabling the use of integer programming optimization. To do so we formulated a PATTERN PALM arrival location model to show that the emergency demand for ward beds (and total hospital beds) can be characterized as independent Poisson random variables. Secondly, we extended the PALM approach to the class of deterministic-arrival-location model (DALM) and analyzed it’s properties.

This new optimization method not only enables us to efficiently generate schedules that reduce blockages and/or increase elective volumes, but also enables us to generate a Pareto
tradeoff curve between blockages and admission volume. This curve represents what seems to be a far more effective decision making tool for hospital administrators, as it enables flexibility and choice rather than prescribing a fixed solution. This approach is likely to increase acceptance and enable hospitals to make important tactical decisions with the benefit of more information.

For future work, it would be useful to investigate the incorporation of discharge rates that vary by day of the week, because this was the major cause for error in our census approximation models. A major extension of the work would include linking the ward census models to an operating room schedule model. Currently our model develops a scheduling system based on downstream bed availability.

In conclusion, this new model 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 their admission schedule as a tool to create a healthcare delivery system that is more profitable while providing better quality service to patients.
References


