Operations Research

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

Design and Optimization Methods for Elective Hospital Admissions

Jonathan E. Helm, Mark P. Van Oyen

To cite this article:
Jonathan E. Helm, Mark P. Van Oyen (2014) Design and Optimization Methods for Elective Hospital Admissions. Operations Research 62(6):1265-1282. http://dx.doi.org/10.1287/opre.2014.1317

OPERATIONS RESEARCH
Vol. 62, No. 6, November–December 2014, pp. 1265–1282
ISSN 0030-364X (print) | ISSN 1526-5463 (online)

http://dx.doi.org/10.1287/opre.2014.1317

Design and Optimization Methods for Elective Hospital Admissions

Jonathan E. Helm
Department of Operations and Decision Technologies, Kelley School of Business, Indiana University, Bloomington, Indiana 47405, helmj@indiana.edu

Mark P. Van Oyen
Department of Industrial and Operations Engineering, University of Michigan, Ann Arbor, Michigan 48109, vanoyen@umich.edu

Hospitals typically lack effective enterprise level strategic planning of bed and care resources, contributing to bed census levels that are statistically "out of control." This system dysfunction manifests itself in bed block, surgical cancelation, ambulance diversions, and operational chaos. This is the classic *hospital admission scheduling and control* (HASC) problem, which has been addressed in its entirety only through inexact simulation-based search heuristics. This paper 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 *strategic planning/scheduling* portion of the HASC. Our new solution method coordinates elective 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 needs model. We further extend the PALM approach to the class of deterministic controlled-arrival-location models (d-CALM) and develop linearizing approximations to stochastic blocking metrics. This work provides the theoretical foundations for an efficient scheduled admissions planning system as well as a practical decision support methodology to stabilize hospital census.

Subject classifications: hospital admissions and bed management; stochastic patient flow modeling; mixed-integer programming; census smoothing; stochastic arrival-location models.
Area of review: Policy Modeling and Public Sector OR.
History: Received April 2011; revisions received September 2012, September 2013, April 2014; accepted July 2014. Published online in *Articles in Advance* October 23, 2014.

1. Introduction

The classical *hospital admission scheduling and control* (HASC) problem identified in the late 1970s addresses one of the major systemic failures in hospital care delivery, *census variability*, through better strategic planning of inpatient admissions. Solution methodologies to reduce this variability are often referred to as *census smoothing*. In this work we solve the elective inpatient (by which we mean all admissions that are scheduled in advance rather than emergency) strategic scheduling portion of the HASC problem to optimality. Our collaborations with multiple hospitals across three continents enable a broad validation of our approach, models, and results. Our HASC optimization creates a strategic plan, analogous to block scheduling, that allocates a specific number of slots each day for each patient type (similar to allocating a certain number of hours of operating room (OR) time to each service each day) over a planning horizon to be filled by admissions personnel scheduling according to the plan.

From Practice to Theory: *A Scientific Approach to the HASC Planning Problem.* The work in this paper was developed during more than four years of collaborative research with hospitals around the world. We have worked with both large and medium sized hospitals and teaching and nonteaching hospitals in the United States, the Netherlands, Singapore, and Canada. The causes and consequences of census variability detailed below, along with the classic census patterns that lead to systemic hospital congestion, were observed to be similar in every case. This suggests that the problem we address is a global one that, despite the many differences across hospitals and healthcare systems, occurs with remarkable consistency. For the purposes of cohesive exposition, we draw our examples from and develop a complete analysis for one hospital in particular, though the model is validated across all four hospitals. Our partner hospitals have agreed that this approach merits its development and implementation as a path toward the hospital of tomorrow.

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 overcrowding of the emergency department (ED), intensive care unit (ICU), and post anesthesia care unit (PACU) resulting in increased mortalities, compromised quality of care, emergency patient diversions, and significant excess cost (see Mirel and Carper 2013, Sprivulis et al. 2006, Richardson 2006, Derlet et al.

2001, Harrison et al. 2005, McManus et al. 2003, Proudlove et al. 2003, Richardson 2006, Fatovich et al. 2005, Hoot and Aronsky 2008, and Forster et al. 2003). Census variability also contributes to overloaded nurse staff, which is linked to patient mortality, nurse burnout, and job dissatisfaction (see Aiken et al. 2002). High levels of congestion also cause hospitals to divert overflow patients into non-preferred "off-ward" beds, which we call type 1 blocking (see §4.1). Patient safety, however, is not well served by placing patients off-ward and the practice increases nurse stress, job dissatisfaction, and turnover (see Anderson et al. 1988, Brownson and Dowd 1997, Needleman et al. 2002).

Figure 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 midweek census spike followed by a sharp drop in census on Saturday and Sunday (see Figure 1(b)). This weekly census "hump" contributes to hospital overcrowding despite a modest average census (the dotted line in Figure 1(b)).

Causes of Census Variability. It is well known that both a weekly pattern in elective admissions (see Figure 2(a)) and the week-to-week variation in number of elective admissions on a given day (see Figure 2(b)) significantly contribute to both the weekly census hump and the week-to-week variation in census, leading to hospital congestion and patient blockages (see Bekker and Koeleman 2011). Table 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 admissions actually exhibit *higher* coefficient of variation (i.e., standard deviation divided by mean) on many days than do emergency admissions. The scheduling portion of HASC can be defined in terms of the objective, *stabilizing hospital census*, and system controls, *elective admission scheduling*. In the following sections we describe an analytical model of the system dynamics that links the control decisions to the objective and allows for mixed-integer programming (MIP) optimization methods.

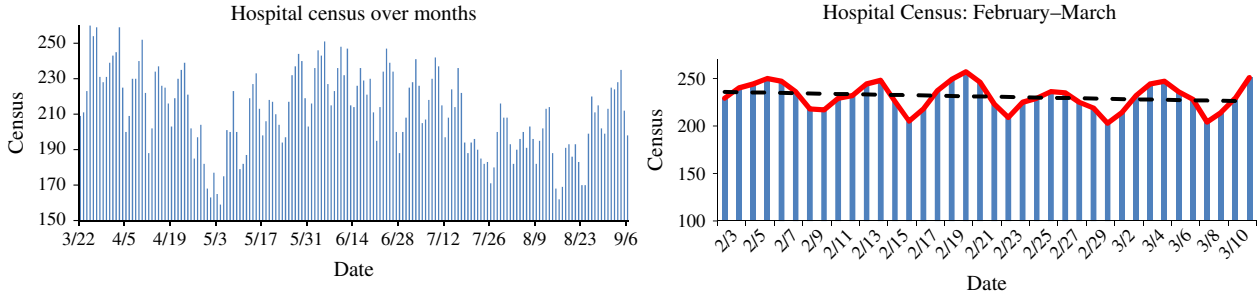
The paper is outlined as follows. Section 2 describes the current state of census smoothing research, §3 develops a stochastic model for hospital workloads across a network of resources, and §4 transforms the stochastic model into a linear MIP to generate optimal admission schedules.

2. Smoothing Hospital Census

Hospitals that increase throughput (achieved through better resource usage) can provide better access to their community at a lower cost, provided they limit patient blockages. To smooth hospital census, the hospital must address both (1) the census midweek "hump" and (2) weekly variability in admissions. The key to smoothing hospital census lies in modeling the downstream ward/bed requirements for admitted patients building toward a stochastic process capturing ward census levels over time for any particular mix and volume of patient admissions.

The importance of census levels and census variability to admission decision making has been studied in several contexts. Connors (1970) uses stochastic patient flow models to link admissions decisions with hospital census.

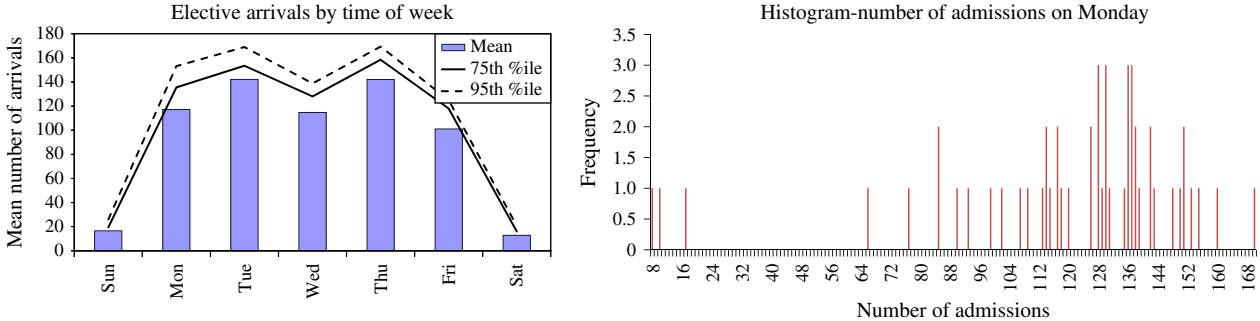
Figure 1. (Color online) Census variability shown in hospitals over (a) ∼ six months and (b) ∼ one month



(a) Census variation over time

(b) Weekly occupancy pattern

Figure 2. (Color online) Variability in elective admissions over the course of one year.



(a) Weekly elective admission pattern

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

Table 1. Variation in numbers of total elective and emergency admissions by day of week (DOW).

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

Harrison et al. (2005) uses simulation to show that census variability in combination with high census levels increases the risk of hospital overcrowding. Jun et al. (1999) argues that effective patient flow management can benefit the hospital through high patient throughput, low patient wait times, short length of stay (LOS), and low clinic overtime.

To effectively solve the HASC problem, models must incorporate control/scheduling decisions into census forecast models. Early work in this area began in the late 1970s with Hancock and Walter (1979, 1983) and Griffith et al. (1978). These early approaches took a comprehensive simulation modeling approach to capture 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 of Hancock and Walter (1983) 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. Gallivan and Utley (2005), Gallivan et al. (2002), Chow et al. (2011), Adan et al. (2009), Bekker and Koeleman (2011) have all studied the impact of elective admissions on census levels in various wards, optimizing schedules with MIP models. Recently, Harper (2002) and Helm et al. (2009) used simulation frameworks to improve scheduling decisions for better hospital resource usage.

Helm et al. (2011) 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 significant impact that elective scheduling has on system performance, this paper makes a contribution by (1) developing *analytical* census modeling methods, rather than simulation-based methods and (2) embedding them in a nonheuristic optimization to solve a model of the *scheduling* side of the HASC problem and to yield important 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 feed-forward 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 more realistic generalized network dynamics of the hospital wards and the use of flexible wards to serve patients off-ward. In short, we are able to solve a *scheduling* model of the complete HASC problem using nonheuristic optimization methods. 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) off-ward servicing when a patient's preferred ward is full.

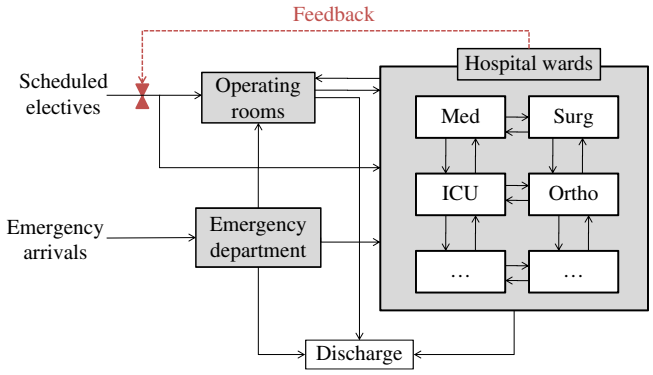
Ignoring the off-ward and interward transfer mechanisms omits critical dynamics of hospital system functioning. In one of our partner hospitals 56% of patients transfer wards at least once during their hospital stay, and among patients who transfer, the average is 1.6 transfers per visit. Considering only the first ward, or a feed-forward subset of wards, ignores a significant load that patients place on other hospital resources. Additionally, the percent of off-ward patients is often quite significant; even in one of the better managed hospitals we worked with, around 17% of patients were located off-ward.

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. Our main purpose is to develop a medium- or long-term plan for what mix and volume of patients to admit over time.

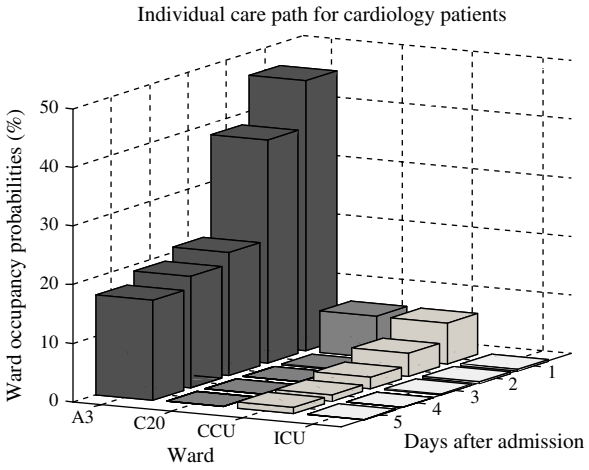
3. Characterization of the Stochastic Census Process

Figure 3(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. We show that it is possible to determine the volume and mix of elective patients that will generate a stable workload and minimize blockages and cancelations while maintaining or increasing patient throughput.

Figure 3. (Color online) Models of patient flow through a network of hospital wards.



(a) Conceptual patient flow model



(b) Cardiology patient flow path

This section develops offered-load models of hospital census that will incorporate corrections for cases where a patient is denied admission to a full hospital (type 2 blocking) or where they are placed "off-ward" because the preferred ward is full (type 1 blocking). As a foundation, we first characterize the patient trajectories for each patient type. Although many different definitions of patient types may be used, we consider patient type to be the patient's admitting service (e.g., cardiac, gastrointestinal, neurology, etc.), because this fits with typical hospital scheduling structures. 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 3(b) (corresponding also to Table 2) 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.

Using these trajectories we can characterize both the elective census process and the emergency census process to model the total census levels in each ward for a given elective admission schedule by day of week. In §4, these census processes are 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, including bed block.

The remainder of this section proceeds as follows. In §3.1 we discuss the design of the proposed elective admission scheduling system as well as modeling assumptions. Section 3.2 introduces our stochastic model regarding how patients move through the network of hospital wards over the course of their treatment and presents a method for extracting a patient's preferred ward from the data when patients are placed off-ward. Section 3.3 combines the stochastic model of patient flow from §3.2 with a Poisson arrival stream to create a stochastic model of hospital and ward workload. Section 3.4 extends this analysis to a broader class of arrival streams, that includes deterministic arrival streams, to model the more controlled arrival of elective patients. Finally, combining the elective and emergency workload models yields a model for the total ward and hospital census, which is validated for accuracy using historical data from four different hospitals in §3.5.

3.1. System Design and Assumptions

The decision variables (the number of patients of each type to admit on each day of the admission cycle) provide

Table 2. Patient Temporal Resource Needs (PATTERN) matrix of the percent of patients that require a bed on days following admission for a cardiology patient for (a) the congestion contaminated path and (b) the estimated true demand path.

(a)						(b)					
	Time (days)						Time (days)				
Ward	0	1	2	3	4	Ward	0	1	2	3	4
A3	45.6%	37.8%	21.3%	19.1%	17.0%	A3	45.8%	38.0%	21.6%	19.4%	17.3%
C2O	6.2%	0.1%	0.0%	0.0%	0.0%	C2O	6.2%	0.1%	0.0%	0.0%	0.0%
CCU	7.1%	3.5%	2.4%	1.3%	1.0%	CCU	7.4%	3.6%	2.5%	1.5%	1.3%
ICU	0.1%	0.1%	0.1%	0.1%	0.1%	ICU	0.1%	0.1%	0.2%	0.1%	0.1%

admission targets (number of slots) for each service, similar to the way that allocating surgical block time manages OR case mix. These slots will be filled by the various admissions personnel until the maximum number of patients of that type for a given day is reached. Patients that cannot be scheduled on a given day will be scheduled into empty slots on subsequent days. This mechanism coordinates a previously decentralized admissions scheduling process.

In high demand hospitals, like the four that we study in this paper, elective inpatient services have substantial waiting lists and thus it is reasonable to assume that these planned slots could be filled every week. A management challenge lies in changing the culture to stabilize the number of admissions from week to week. The operational mechanism of assigning admission maximums by day of week, if followed, will inherently reduce the fluctuation in admissions over time.

By using a deterministic arrival stream model, we are capturing the results if the plan is followed accurately and the admission plan is filled each week. It is important to note that, as this is a strategic planning model, the elective decision variables represent the planned admissions and *neither* (1) an uncontrolled arrival of requests for elective admissions or procedures, nor (2) how the admission slots dynamically get filled. The dynamic management of the admissions system is left to the control portion of HASC (see for example Hancock and Walter 1983, Helm et al. 2011). Instead, we allocate slots that set an optimal mix and volume of patient admissions and allow admissions personnel the flexibility of filling those slots with any patient that matches the criteria for the slot (much like OR block scheduling). Using our modeling framework it is also possible to capture deviations from the plan (see §3.4.3) modeled using various nondeterministic arrival streams.

In our admission plan design, we model a repeating admission cycle (e.g., a week) and, when the system goes beyond the cycle length, the admission plan is repeated exactly as before. It should be noted that the modeling framework is general and can work in a variety of contexts including situations with seasonality and scenarios that are not cyclic. Although noncyclic systems and systems with seasonality can be modeled as well, a weekly cyclo-stationary model matches the natural weekly cycle of almost all hospitals (e.g., planned clinic times, OR time, research time).

In traditional queueing network models, customers that are blocked stop receiving service while they wait for a server at the next station to free up. In contrast, when patients are blocked from their preferred ward, they continue to receive service while residing off-ward with only a small increase (~ 0.5 days) in LOS. The infinite capacity offered load model we develop captures this phenomenon of receiving service continuously while in the hospital. In §4 we further superimpose capacity constraints on the offered load model to capture the volume of off-service patients and the rate at which patients are blocked from entering the hospital altogether (e.g., ambulance diversion). This technique has been used successfully in other applications (e.g., the modified offered load approach of Massey and Whitt 1994b has been used extensively). The workload from patients that are blocked from the hospital entirely is removed from the workload estimates. Thus we only capture patients that are in the hospital and once in the hospital, the patient continues to be "served" until they leave. This approach models true patient flow dynamics far better than traditional queueing network blocking models or loss models.

Our model uses one day as the time step (though any time step can work), because we anticipate our elective admissions system being used most often at a daily granularity to give flexibility and decision-making power to admissions personnel, increasing the likelihood of acceptance of our coordinated strategic plan.

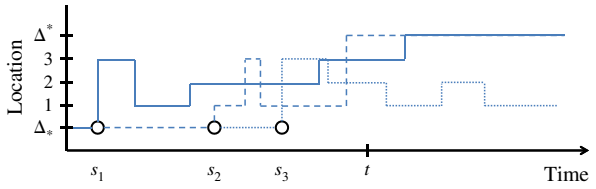
3.2. Development of the PATTERN Stochastic Location Process Model

To understand the effects of scheduling decisions and emergency arrivals on census levels across the network of hospital wards, consider as a foundational model the resource (bed) requirements of a single patient over the course of their treatment, which we call *Pat*ient *Te*mporal *R*esource *N*eeds (PATTERN). To describe the flow of patients through hospital wards, we develop a stochastic location process model in the spirit of Massey and Whitt (1993, 1994a). Some applications and extensions of this approach include Leung et al. (1994) and Liu and Whitt (2011). Let $\mathfrak{W}$ be the set of wards and $\mathfrak{D}$ be the set of patient types. The state space for the location functions can be defined as $\mathscr{S} = \mathfrak{W} \cup \{\Delta^*, \Delta_*\}$, where state u represents a patient being in ward u, 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 $\mathscr{S}$-valued stochastic location process $\{L_{s,k}(t)\colon s \in \mathbb{R},\ k \in \mathfrak{D}\}$, where k is the patient type, s is the arrival time, and $t > s$ is the time of interest. For notational convenience we let $\mathscr{S} = \mathscr{S}^0 \cup \{\Delta^*, \Delta_*\}$, so that $\mathscr{S}^0$ represents the locations within the hospital. Thus $L_{s,k}(t)$ denotes the location of a patient at time t given that the patient was admitted at time s.

Remark 1. Bed capacity is not explicitly represented in this stochastic location process; however, the location process is constructed from data that reflects actual flows observing the bed capacity constraints (even at the ward level). For hospitals with significant capacity constraints we view the data as congestion contaminated, and Theorem 1 will provide a tool to estimate the ideal flows, free of off-ward placement or rejection from the hospital (types 1 and 2 blocking).

Remark 2. The fact that $L_{s,k}(t)$ can depend on s enables the modeling of the key hospital feature that the length of stay and care path can depend on the time of admission.

Figure 4. (Color online) Illustration of $L_{s,k}(t)$ for three sample care paths.



To characterize the stochastic location process, let Σ_s be the set of right-continuous functions with left limits for patients 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 $L_{s,k}(t)$. An element $\sigma_s \in \Sigma_s$ is a (deterministic) mapping $\sigma_s\colon \mathbb{R} \to \mathcal{S}$ such that $\sigma_s(t)$ represents the location of the patient at time t. Figure 4 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 3 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 $t \geqslant T_s$.

We let the entire function space Σ be the collection of all Σ_s. For any subset $\Gamma \subseteq \Sigma$, let the associated probability measure, $P_s(\Gamma)$, represent the probability of realizing one of the location functions in Γ, assigning 0 measure to any location functions in Γ that did not begin at time s (the time of the patient's arrival). Thus $P_s(\Sigma_s) = 1$ and $P_s(\Sigma_t) = 0$ for $t \neq s$. $P_s(\cdot)$ characterizes the dynamics of the stochastic location process, $L_{s,k}(t)$. For our model, 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. To do so, we define a set of location functions and then the measure on that set as follows. The measure of the set of location functions that indicate that the patient is in ward u at time t is the probability of the patient being in ward u at time t. This set can be written as

$$\Gamma_{t,u} = \{\sigma_s \in \Sigma_s\colon s \leqslant t \text{ and } \sigma_s(t) = u\}, \tag{1}$$

which captures 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. Moreover, we require that the patient not remain in the hospital forever (consistent with Massey and Whitt 1993). As mentioned, the specific measure of this set is defined by the dynamics of the stochastic location process $L_{s,k}(t)$. We avoid the semi-Markov process because the solution to such processes for general distributions and general transition functions is often intractable, requiring further approximations. Rather, we define for each patient type, k, a specific stochastic location model with probability measure $P_{s,k}$ as

$$P_{s,k}(\Gamma_{t,u}) = p_s^{k,u}(t-s), \tag{2}$$

$$P_{s,k}(\Gamma_{t,\Delta^*}) = p_s^{k,\Delta^*}(t-s) = 1 - \sum_{u \in \mathcal{W}} p_s^{k,u}(t-s), \tag{3}$$

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 periods after their arrival.

For each patient type, we calculate the proportion of the total population present in each ward (or discharged) for each discrete time step (see for example Table 2). It is often the case that some of the historical hospital data captures the hospital's reaction to congestion (e.g., "off-ward" placement) rather that the patients' ideal flows. The simple solution is to take patient data only from periods when the hospital is not congested so that blocking and off-ward placement is minimal. Even high demand hospitals have some periods of low congestion; however, this approach requires a longer period of data. In general, we propose the following approach to correct for data contamination.

Congestion impacts two aspects of patient flow post-admission: length of stay and ward placement. Under standard procedures, no patients are forced out of the hospital by blockage once they have been admitted. Off-ward servicing during the entire hospital stay only increases LOS by around half a day (see Anderson et al. 1988), so this perturbation is not significant at the strategic planning level we are investigating. Thus, off-ward placement is the congestion effect we correct for to extract true patient flows from congestion-contaminated hospital data.

We model both observed flows and true flows, distinguished notationally by adding a hat "$\hat{}$" on top of the parameter to associate it with the true quantity. The four hospitals we worked with used overflow wards and overflow/blocking routing policies to manage congestion. Therefore, the set of wards, $\mathcal{S}^0$, can be partitioned into specialized wards, $\xi \subseteq \mathcal{S}^0$, and overflow wards $\psi = \mathcal{S}^0 \backslash \xi$. Let a_i represent the arrival rate to ward i and $p_{i,j}$ be the probability that a patient who completes service in ward i transitions next to ward j. Let $\hat{\beta}_i$ represent the blocking probability in ward i.

To discover the true flows, we relate the traffic equations of the true flows in the underlying flow system to the flow model observed in the hospital data. Each patient type has its own arrival rates and transfer probabilities, but for the sake of exposition, we initially suppress the dependency of these parameters on patient type. The system of equations below can be easily extended to include arbitrarily many patient types:

$$a_i = \hat{a}_i(1 - \hat{\beta}_i) \quad \text{for } i \in \xi, \tag{4}$$

$$p_{i,j} = \hat{p}_{i,j}(1 - \hat{\beta}_j) \quad \text{for } i, j \in \xi, \tag{5}$$

$$\sum_{k \in \psi} p_{i,k} = \sum_{j \in \xi} \hat{p}_{i,j}\hat{\beta}_j \quad \text{for } i \in \xi. \tag{6}$$

In Equation (5), the observed transfer probability from i to j is the true transfer probability multiplied by the proportion of time the patient is able to enter the ward (i.e., unblocked). Equation (4) follows similar logic. Equation (6) means that the total probability of transitioning into the overflow ward is the sum of all the between-ward transitions that were blocked. Equations (4)–(6) represent a nonlinear system of equations with n variables and n unknowns after patient types are included. Let $P = [p_{i,j}]$ be the observed ward transition matrix and let P^{ξ} be the its submatrix that represents the transitions between specialized hospital wards, i.e., $\{p_{i,j}\colon i, j \in \xi\}$. Let $P^{\psi} = \sum_{i\in\psi} P_i$, where P_i is the ith column vector. In words, P^{ψ} represents the probability of transitioning from a specialized ward into any of the overflow wards, i.e., $\{\sum_{j\in\psi} p_{i,j}\colon i \in \xi\}$. By transforming the nonlinear system into an equivalent linear system, we show that the traffic equations have a unique solution obtained by matrix inversion (proved in the online appendix (available as supplemental material at http://dx.doi.org/10.1287/opre.2014.1317)).

THEOREM 1. *Given P^{ξ} has full rank, the traffic equations given by Equations* (4)–(6) *have a unique solution given by*

$$\gamma = (P^{\xi})^{-1}P^{\psi}, \qquad \hat{\beta}_i = \frac{\gamma_i}{1+\gamma_i},$$
$$\hat{a}_i = \frac{a_i}{1-\hat{\beta}_i}, \qquad \hat{p}_{i,j} = \frac{p_{i,j}}{1-\hat{\beta}_j}. \tag{7}$$

If the transition probability matrix, P^{ξ}, does not have full rank (i.e., if any rows of the matrix are linearly dependent) then we can break the matrix into submatrices of full rank by extracting rows that cause linear dependence and then applying Theorem 1 to each cluster separately.

To validate our method, we developed a simulation model of admissions, blocking, and off-ward placement to generate congestion contaminated census process realizations from one year's worth of data. This enabled us to compare the results from our decontamination method with a known true demand distribution. We simulated three ward hospitals because three wards are considered sufficient to capture the rich network structure of interest, and this structure is often used in the patient flow literature. We designed a test suite of 1,000 cases with the hospital parameters generated randomly. We used the following parameterizations. Patient LOS for each patient type was log-normally distributed with mean (in days) and variance parameters randomly chosen from a uniform(2,8). The ward sizes were chosen from a uniform(8,60) distribution. Transfer probabilities were uniform(0,1) for each pair of wards. Random arrival rates were generated also according to a uniform, but with attention to creating a stable queueing system. Using the simulation, we generated congestion-contaminated "observed" data and then used Theorem 1 to estimate the underlying flow parameters and compared our estimate with the true parameters used to design the simulation.

The primary quantity needed to estimate a patient's true care pathway is the transfer probability, which had a small 0.1% *average absolute error* across the test suite. The blocking probability and arrival rate estimates, also had small errors 1.0% *average absolute error* and 1.5% *average absolute percent error*, respectively. As three ward structures are taken to be representative in the patient flow literature, we expect the results to be similar for other systems. The high level of accuracy demonstrated with our large test suite supports the claim that our method is capable of identifying true care needs from congested historical data, enabling contamination-free parameterization of the location processes.

Once the true system dynamics have been calculated, Theorem 1 can be used in a preprocessing step (prior to calculating patient path probabilities) to adjust the patient pathways for each individual patient type to reflect each patient's true demand by removing congestion contamination. We begin by noting that the hospital's raw data accurately represents each patient's desired ward (true demand) except when a patient enters an overflow ward as an off-unit patient. This is where Theorem 1 is needed to correct for the congestion that forced the patient into overflow wards by capturing which ward the patient was trying to enter (their preferred ward for that segment of treatment) when they were forced into the overflow ward because of lack of capacity. To estimate the patient's preferred ward in such a situation, we replace the line in the raw data that has one in the overflow ward and zero in all other wards (e.g., row 2 of "congestion contaminated data" in Figure 5) with a set of probabilities across all specialized wards: $\mathbb{P}$(ward u is the patient's preferred ward) for each $u \in \xi$ (e.g., row 2 of "preferred ward transformation" in Figure 5).

If a patient moves to the overflow ward as a result of a transfer from ward i, then $\mathbb{P}$(ward u is the patient's preferred ward | the patient was blocked) $= \hat{p}_{i,u}\hat{\beta}_u / \sum_{j\in\xi} \hat{p}_{i,j}\hat{\beta}_j$, where $\hat{p}_{i,j}$ is the true probability of transferring from ward i to ward j and $\hat{\beta}_u$ is the blocking probability in ward u, both obtained from Theorem 1. If, instead, the patient begins their stay in the overflow ward, the probability that the patient's preferred ward is u is given by $\hat{a}_u\hat{\beta}_u / \sum_{j\in\xi} \hat{a}_j\hat{\beta}_j$, where $\hat{a}_j$ is the true arrival rate to ward j obtained from Theorem 1. Thus to calculate the patient's true demand for service, the workload is shifted from the observed load in the overflow ward to the specialized ward(s) that represent the patient's preferred ward as shown in Figure 5. Doing this for each patient, we can transform the hospital's congestion-contaminated raw data into congestion free data that can be used to more accurately compute the location function probabilities.

An example of a fully parameterized location process for cardiology patients is shown in Table 2, illustrating before and after the transformation to extract the true demand over five days. 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

Figure 5. Transformation of congested data into true demand data for an individual patient who stayed in ward 4 on day 2 of their hospital stay and then was transferred to an overflow unit on day 3 because the preferred ward that they wanted to transfer to was full.

Congestion contaminated data					
	Ward 1	W2	W3	W4	Ovrflw
Day 2	0	0	0	1	0
Day 3	0	0	0	0	1
Day 4	0	0	0	0	1

Preferred ward transformation					
	Ward 1	W2	W3	W4	Overflow
Day 2	0	0	0	1	0
Day 3	$\hat{p}_{4,1}\hat{\beta}_1 \Big/ \sum_{j\in\xi}\hat{p}_{4,j}\hat{\beta}_j$	$\hat{p}_{4,2}\hat{\beta}_2 \Big/ \sum_{j\in\xi}\hat{p}_{4,j}\hat{\beta}_j$	$\hat{p}_{4,3}\hat{\beta}_3 \Big/ \sum_{j\in\xi}\hat{p}_{4,j}\hat{\beta}_j$	0	0
Day 4	$\hat{p}_{4,1}\hat{\beta}_1 \Big/ \sum_{j\in\xi}\hat{p}_{4,j}\hat{\beta}_j$	$\hat{p}_{4,2}\hat{\beta}_2 \Big/ \sum_{j\in\xi}\hat{p}_{4,j}\hat{\beta}_j$	$\hat{p}_{4,3}\hat{\beta}_3 \Big/ \sum_{j\in\xi}\hat{p}_{4,j}\hat{\beta}_j$	0	0

is a cardiology ward, CCU is the critical care unit, ICU is the intensive care unit, and C2O is a ward for short stay patients (usually less than two days). 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.

Remark 3. Through testing, the optimization model presented in §4 was shown to be robust to the modifications to the location process resulting from correcting for congestion-contaminated paths. The difference between the optimization results using congestion contaminated location functions versus the transformed "true demand" location functions was small. Over several different scenarios, the relative percent difference in objective function (blocking probability) was very small—between 3% and 5% in experiments run, or likewise a difference in expected blockages per week of 0.05 to 0.1—and the final schedules were very similar. To explain how this can occur, consider that the math model optimizes system level metrics (hospital level blockage, or total elective throughput) subject to constraints on off-ward census, which is the model component impacted by congestion contamination. If those constraints are not very tight, they will not have much impact. Furthermore, in most cases there are many possible solutions that achieve a similar objective value, so in many instances a change in the off-ward constraint brought about by correcting for congestion contamination will not have a large impact on the objective if another solution with similar performance can be achieved by shifting some of the elective workload. Logically, sufficiently high levels of data contamination will eventually have a strong impact on the optimal solution after correcting for data contamination.

3.3. PATTERN Poisson-Arrival-Location Model (PALM) of Emergency Census

We begin by modeling the demand for services with each ward modeled as a cluster of infinite server queues. There is one queue for each emergency patient type, with its own nonhomogeneous arrival rate, its own service distribution, and its own routing probabilities, denoted by Massey and Whitt (1993) as $(M_t/G_t/\infty)^N/G_t$. It has been shown that the nonstationary Poisson process is a good model for emergency patient arrivals (see Harrison et al. 2005), and we allow for general, nonstationary service time distributions as well as nonstationary routing probabilities that may also depend on the length of stay in a given ward. Our interest lies in the number of patients demanding a bed in each ward. Letting there be $|\mathfrak{W}| = M$ wards and n emergency patient types, the network of $M \cdot n$ queues has $Q_u(t) = Q_u^1(t) + Q_u^2(t) + \cdots + Q_u^n(t)$ emergency patients placing a service load on ward u at time t, where $Q_u^k(t)$ is the demand of type k patients for ward u. Let $\mathbf{Q}(t) = \sum_{u\in\mathfrak{W}}\sum_{k=1}^{n} Q_u^k(t)$ denote the total emergency patient load at time t. These two quantities are sufficient for our later analysis, in which we overlay capacities on the demand model to calculate blockages and off-ward census.

To specify the PATTERN PALM model for the emergency census process, we rely on the Poisson random measure approach proposed by Massey and Whitt (1993). In PATTERN PALM, patients arrive according to a nonhomogeneous Poisson process and then flow through the hospital according to our PATTERN stochastic location process $L_{s,k}(t)$ described in §3.2. Details of the standard Poisson random measure and its extension to a doubly stochastic Poisson process can be found in the online appendix EC.1, which also provides elaboration on this section. 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 $\mathcal{N} = \{\mu \in \mathcal{M}\colon \mu(t) \in \mathbb{Z}^+\}$, the set of measures $\mu \in \mathcal{M}$ that yield integer values.

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 §3.4. To specify the location random measure, we define a mapping from the probability space $(\Sigma, \mathcal{B}, \mathbb{P})$ into the measure space $(\mathcal{M}, \mathcal{B}(\mathcal{M}))$, with $\mathcal{B}(\cdot)$ as the Borel sigma algebra. Let the probability that a patient of type k arriving at time s is in ward u at time t be defined as

$$\mathbb{P}_k(\sigma_s \in \Sigma_s\colon \sigma_s(t) = u) \equiv P_{s,k}(\sigma \in \Sigma\colon \sigma(t) = u) = \begin{cases} 0 & \text{if } t < s \\ p_s^{k,u}(t-s) & \text{if } t \geqslant s. \end{cases} \tag{8}$$

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

$$\Lambda_{k,s}(t, \mathcal{J}, \sigma) = \begin{cases} 1 & \text{if } \sigma(t) \in \mathcal{J},\ \sigma \in \Sigma_s \\ 0 & \text{otherwise.} \end{cases} \tag{9}$$

Now we can specify the random intensity measure, $\mathbf{N}_k$, for patients of type k by combining the nonhomogeneous Poisson arrival process having nonnegative deterministic integrable external-arrival-rate function $\alpha_k(t) \in \mathbb{R}^+$ with the location random measure from Equations (8) and (9). 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 nonstationary 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 of type k arrivals to wards $\mathcal{J}$ at time t for patients that entered the hospital over the interval $(a, b]$:

$$\mathbf{N}_k((a, b], t, \mathcal{J}) = \int_a^b \alpha_k(s)\Lambda_{k,s}(t, \mathcal{J})\, ds. \tag{10}$$

Intuitively, this can be related to Poisson splitting of a nonhomogenous 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 $L_{s,k}(t)$. Therefore the external arrival intensity is distributed over time across the wards (or "departed"). 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 its properties are defined in the following lemma (proved in the online appendix EC.2):

Lemma 1. *For the deterministic average arrival intensity measure, μ_k, the following hold*:

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

(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 $B_i = (a_i, b_i] \times t_i \times \mathcal{J}_i$ 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 future time, t_i. Then $\mathbf{M}_k$ can be shown to have a product form Poisson distribution with rate γ_i:

$$P\big(\mathbf{M}_k(B_1) = m_1, \mathbf{M}_k(B_2) = m_2, \ldots, \mathbf{M}_k(B_n) = m_n\big) = \prod_{i=1}^{n} \frac{e^{-\gamma_i}\gamma_i^{m_i}}{m_i!} \tag{11}$$

$$\gamma_i \equiv \mathbf{E}[\mathbf{M}_k(B_i)] = \mu_k\big((a_i, b_i], t_i, \mathcal{J}_i\big) = \int_{a_i}^{b_i} \alpha_k(s) \sum_{u \in \mathcal{J}_i} p_s^{k,u}(t_i - s)\, ds. \tag{12}$$

Equation (12) follows from Lemma 1. We now quantify the distribution on the number of emergency patients in the cyclo-stationary system (mentioned in §3.1) 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 (in our case study of §4.4, $\max_k \tau_k = 215$) then we have the following result, which is proved in the online appendix EC.2.

Theorem 2. *The number of emergency patients in ward u, denoted by $Q_u(t)$ for $u \in \{1, \ldots, n\}$, are independent Poisson random variables for each time $t \in \mathbb{R}^+$ with finite mean given by*

$$m_u(t) = \sum_{k=1}^{n} \int_{t-\tau_k}^{t} \alpha_k(s) p_s^{k,u}(t - s)\, ds. \tag{13}$$

3.4. PATTERN Deterministic Controlled-Arrival-Location Model (d-CALM) of Elective Census

The approach for the elective census model represents an extension of the PALM methodology to processes with deterministic arrivals, which we term 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 §3.2. This makes explicit our condition that each patient type has a unique location process determined by their characteristics at the time of admission.

3.4.1. 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 to represent either historical or future behavior. We first present the formal analysis, then illustrate it with an example. Our approach is to formulate a point process as in §3.3. 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 sequence of deterministic arrivals with $t_{k,i}$ being the time of arrival of the ith 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

$$\omega_k = \big\{\sigma_{k,(t_{k,1}),1}, \sigma_{k,(t_{k,1}),2}, \ldots, \sigma_{k,(t_{k,1}),\Theta_{k,t_{k,1}}}, \sigma_{k,(t_{k,2}),1}, \sigma_{k,(t_{k,2}),2}, \ldots, \sigma_{k,(t_{k,2}),\Theta_{k,t_{k,2}}}, \ldots\big\} \in \Omega$$

represents the set of location functions for the scheduled arrivals. Under the infinite capacity model, we can define the d-CALM probability measure for patients of type k being in ward u as

$$\mathbb{P}_k\big(\{\omega \in \Sigma^\infty\colon \sigma_{k,(t_{k,n}),n}(t) = u\}\big) = \begin{cases} 0 & \text{if } t < t_{k,n}, \\ p^{k,u}_{(t_{k,n})}(t - t_{k,n}) & \text{if } t \geqslant t_{k,n}. \end{cases} \tag{14}$$

where $p^{k,u}_{(t_{k,n})}(t - t_{k,n})$ is as before in Equation (3). Then we can define the d-CALM point process, for a realization vector ω as

$$\mathbf{N}_{k,u,\Theta}(t,\omega) = \begin{cases} \sum_{s\in\{t_{k,i}\colon t_{k,i}<t\}} \sum_{n=1}^{\Theta_{k,s}} \Lambda_{k,s}(t,u,\sigma_{k,s,n}) & \\ \qquad\qquad \text{if } t_{k,1} < t, & \\ 0 \qquad\qquad \text{if } t_{k,1} > t, & \end{cases} \tag{15}$$

where $\Lambda_{k,s}(\cdot)$ is the patient type k random measure defined for the stochastic location process in Equations (8) and (9). It can be seen that this process describing the elective/scheduled workloads across the network of wards can be written instead as

$$\mathbf{N}_{k,u,\Theta}(t) = \sum_{s\in\{t_{k,i}\colon t_{k,i}<t\}} \sum_{j=1}^{\Theta_{k,s}} \mathbf{1}\{L^j_{s,k}(t) = u\}, \tag{16}$$

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 Equations (14) and (15). The ward level census can be calculated by summing over patient types. 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.

Using Equation (16) the census in ward u, C^t_{u,d_1}, can be calculated on week t on a given day d_1 of the admission cycle. If we take the length of the cycle to be one week for example, the census in ward u on a given day d_1 can be calculated for a t week horizon (C^t_{u,d_1} from Equation (17)) or an infinite horizon (C^∞_{u,d_1} from Equation (18)),

$$C^t_{u,d_1} = \sum_{d_2=1}^{7} \sum_{k\in\mathfrak{D}} \sum_{j=0}^{\Theta_{k,d_2}} \sum_{n=0}^{t} \mathbf{1}\big\{L^{j,n}_{d_2+7n,k}(d_1+7t) = u\big\}, \tag{17}$$

$$C^\infty_{u,d_1} = \lim_{t\to\infty} C^t_{u,d_1}, \tag{18}$$

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 $\mathbf{1}\{\cdot\}$ is the indicator function. In Equations (17) and (18), 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 admission cycle. The final sum over n iterates through weeks (or through repeating cycles). Total hospital census is found by adding up all wards.

These equations are best understood through a simple example. Consider a plan that admits two cardiology patients (patient type = CAR) every Monday. What is the load that this plan places on the cardiology ward (ward c) on Tuesdays? Let $\mathbf{1}\{L^{j,n}_{s,\mathrm{CAR}}(t) = c\}$ represent whether the (j,n) indexed cardiology patient (i.e., jth patient admitted on week n) 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 $(j=1, n=0)$ and $(j=2, n=0)$). This leads to a census for Tuesday of the first week ($n=0$) of $\mathbf{1}\{L^{1,0}_{1,\mathrm{CAR}}(2) = c\} + \mathbf{1}\{L^{2,0}_{1,\mathrm{CAR}}(2) = c\}$. Note that $\mathbf{1}\{L^{1,0}_{1,\mathrm{CAR}}(2) = c\}$ and $\mathbf{1}\{L^{2,0}_{1,\mathrm{CAR}}(2) = c\}$ are i.i.d. because they represent two different patients. In the second week we admit two more cardiology patients (call them patient $(j=1,\ n=1)$ and $(j=2,\ n=1)$). 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

$$\mathbf{1}\{L^{1,0}_{1,\mathrm{CAR}}(9) = c\} + \mathbf{1}\{L^{2,0}_{1,\mathrm{CAR}}(9) = c\} + \mathbf{1}\{L^{1,1}_{8,\mathrm{CAR}}(9) = c\} + \mathbf{1}\{L^{2,1}_{8,\mathrm{CAR}}(9) = c\}.$$

If we let the system run for t weeks, then the census on the Tuesday of week t is given by

$$\sum_{n=0}^{t} \mathbf{1}\{L^{1,n}_{7n+1,\mathrm{CAR}}(7t+2) = c\} + \mathbf{1}\{L^{2,n}_{7n+1,\mathrm{CAR}}(7t+2) = c\}.$$

This shows how we construct the census profile for Equations (17) and (18). In this paper we are primarily interested in the steady state behavior of the system, and thus rely mostly on the infinite horizon formulation of Equation (18) in the analysis that follows.

3.4.2. Moments of the PATTERN d-CALM Elective Census Process. An important feature of the d-CALM model is that the first and second moments of the elective census process can be calculated analytically, which facilitates the elective admissions optimization. Taking the cycle length to be one week, for example, the census mean for ward u on a given day d_1 can be calculated from Equations (17) and (18) by the monotone convergence theorem as

$$\mu_{d_1,u}(\Theta) = \mathbf{E}\left[\sum_{d_2=1}^{7} \sum_{k\in\mathfrak{D}} \sum_{j=0}^{\Theta_{k,d_2}} \lim_{t\to\infty} \sum_{n=0}^{t} \mathbf{1}\big\{L^{j,n}_{d_2+7n,k}(d_1+7t) = u\big\}\right] = \sum_{d_2=1}^{7} \sum_{k\in\mathfrak{D}} \Theta_{k,d_2} \sum_{n=0}^{\infty} p^{k,u}_{d_2-7n}(d_1 - d_2 + 7n). \tag{19}$$

The equality follows from the fact that $\mathbf{1}\{\mathbf{X} = x_k\}$ follows a Bernoulli distribution and thus $\mathbf{E}[\mathbf{1}\{\mathbf{X} = x_k\}] = p_k$. The mean census level in the hospital is $\sum_{u\in\mathfrak{W}} \mu_{d_1,u}(\Theta)$.

We compute the variance of the elective census process for (1) the variance in ward census and (2) the variance in total hospital census (with proof in the online appendix EC.2).

Lemma 2. *The covariance at day t of the cyclo-stationary location processes for two patients of types k_1 and k_2 arriving at times s_1 and s_2 being in ward u_1 and u_2 that were admitted as patient j_1 and j_2 of week n_1 and n_2 is*

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

(ii) $\mathrm{Cov}(\mathbf{1}\{L^{j,n}_{s,k}(t) = u_1\}, \mathbf{1}\{L^{j,n}_{s,k}(t) = u_2\}) = -p^{k,u_1}_s(t-s)p^{k,u_2}_s(t-s)$ *for* $u_1 \neq u_2$.

Theorem 3. *Letting $d(n) = d_1 - d_2 + 7n$, the variance of the cyclo-stationary ward and total census processes on day $d_1 \in \{1, 2, \ldots, 7\}$ considering an infinite horizon and admission plan Θ is*

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

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

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. Since, from Theorem 3, the variance is still linear in terms of our decision variables $\Theta_{k,d}$, the model remains solvable by standard MIP solution approaches.

3.4.3. CALM with Deviations from the Planned Admission Schedule. Although the purpose of this paper is to develop an optimal strategic admission plan and predict its benefits if followed correctly, we can also capture the effect of deviations from the plan by letting the decision variable (admission target) be the mean rate of a general stochastic arrival process. The variability around the mean represents deviation from the plan. The following theorems show that the mean of the CALM model can be calculated linearly for any arrival process and that the variance of the CALM model can be calculated linearly for a certain class of arrival processes (with proofs given in the online appendix EC.2).

Theorem 4. *Let $X_{k,d}$ be the random number admissions of type k and day d having mean $\mathbb{E}[X_{k,d}] = \Theta_{k,d}$. The mean workload in ward u on day d_1 is given by*

$$\mu_{d_1,u}(\Theta) = \sum_{d_2=1}^{7}\sum_{k\in\mathfrak{D}} \Theta_{k,d_2} \cdot \sum_{n=0}^{\infty} p^{k,u}_{d_2-7n}(d_1 - d_2 + 7n). \qquad (20)$$

Theorem 5. *If $\mathrm{Var}[X_{k,d}] = f(\Theta_{k,d})$, where f is a linear function, then the variance of the workload in ward u on day d_1 is also linear in $\Theta_{k,d}$ and is given by*

$$\text{(i) } \sigma^2_{d_1,u}(\Theta) = \sum_{d_2=1}^{7}\sum_{k\in\mathfrak{D}}\Bigg[\Theta_{k,d_2}\sum_{n=0}^{\infty} p^{k,u}_{d_2-7n}(d(n)) \cdot(1 - p^{k,u}_{d_2-7n}(d(n))) + f(\Theta_{k,d}) \cdot\Bigg(\sum_{n=0}^{\infty} p^{k,u}_{d_2-7n}(d(n))\Bigg)^2\Bigg],$$

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

Theorem 4 shows that the mean workload can be calculated linearly in the admission plan, Θ regardless of the arrival distribution. Theorem 5 asserts that, if the variance of the arrival distribution is a linear function of the mean, then the variance may also be calculated linearly in Θ. A number of distributions that would be good for modeling deviation from the admission plan also have the property that the variance is a linear function of the mean including uniform, Poisson, and normals of the form $N(\Theta_{k,d}, f(\Theta_{k,d}))$.

3.5. Validating the Hospital Census Model

The total census process (for wards and for the hospital) is approximated by summing the emergency and elective census processes (§3.3 and 3.4). In this section, we show that this approximation closely matches the actual census process for four hospitals from four countries and three continents. For each hospital, we divided the data equally into a training set and a test set. The training set (consisting of the first half of the data, in terms of time) was used to parameterize the location process models and estimate future emergency arrival rates. Our model then generated emergency census estimates for the test set (the second half of the data) using the PATTERN PALM model parameterized by the location processes and emergency arrival rate estimates from the training data. The d-CALM census model was built using location processes parameterized by the training data and a weekly cyclo-stationary, deterministic controlled arrival (admission) process. The deterministic arrival rate of each type on any day was based on the mean arrival rates for elective admission by day of week in the test data. This was done to emulate the idea that the strategic plan was followed on average, though in the test data there were deviations from the "plan" from week to week.

It is important to note that, despite the fact that we built a cyclo-stationary model with deterministic planned elective

Table 3. Accuracy of the hospital level occupancy model across four hospitals on three different continents as measured by the percent error in the mean (Err) and the percent error in the 95% quantile (95% Q Err).

DOW	All avg.	Hospital 1 (%)		Hospital 2 (%)		Hospital 3 (%)		Hospital 4 (%)	
		Err	95% Q Err	Err	95% Q Err	Err	95% Q Err	Err	95% Q Err
Sunday	1.0	0.1	1.0	−3.8	−2.6	1.0	2.4	2.8	3.2
Monday	−2.0	−0.1	−0.5	−1.8	−0.2	−4.1	−4.2	−0.6	−3.3
Tuesday	0.3	−0.8	−0.1	−1.0	1.2	1.0	2.5	−1.0	−2.6
Wednesday	−1.7	−1.5	1.1	−0.7	0.8	−5.2	−5.9	−2.8	−2.9
Thursday	0.3	−1.1	2.8	1.1	1.6	2.3	1.6	−1.8	−4.8
Friday	1.4	−1.2	1.9	1.5	2.1	5.5	2.5	1.3	−0.9
Saturday	2.6	−1.6	−0.2	−1.4	−0.6	5.9	7.4	3.7	3.7
MAPE	**2.0**	**0.9**	**1.1**	**1.6**	**1.3**	**3.6**	**3.8**	**2.0**	**3.0**

Note. The mean percent error across hospitals is the first column and the mean absolute percent error (MAPE) across days of the week is given in the last row in bold.

Table 4. Average absolute percent error of the model vs. true occupancy across 33 wards and a sample of the best and worst errors for six of the 33 wards.

DOW	Avg. of 33 wards (%)	Sample of the best and worst wards (%)					
		Ward 3	Ward 7	Ward 12	Ward 22	Ward 25	Ward 31
Sunday	3	1	1	8	6	0	1
Monday	4	0	4	7	11	0	0
Tuesday	3	0	1	6	8	0	1
Wednesday	3	2	1	5	6	2	1
Thursday	5	1	0	6	3	3	2
Friday	5	1	0	7	4	4	3
Saturday	4	3	2	6	7	3	2
Average	**4**	**1**	**1**	**7**	**6**	**2**	**1**

admission rates based on historical data, the model still performed quite well in predicting true future occupancy levels across a variety of measures (mean, quantile, and ward-level) in four significantly different hospitals. Tables 3 and 4 show the model accuracy at the hospital and at the ward level for 33 hospital wards. Hospital 3 was a small hospital (in terms of beds), hospitals 2 and 4 were medium sized, and hospital 1 was large. The average occupancy in the four hospitals ranged from 82% and 92%. We obtained an average absolute percent error of 2.0% across all hospitals.

The deviations are seen to be relatively small and have little effect on accurately approximating system metrics such as cancelations and blockages as will be calculated in §4.3. Prior efforts at solving this problem for the entire hospital have relied on simulations to achieve accurate census approximations, making optimization difficult (see Helm et al. 2011, Hancock and Walter 1979, Harper 2002). The validation in this section demonstrates that the analytical model, which is amenable to optimization, is sufficient for modeling hospital occupancies, thereby eliminating the need for computationally limiting simulations.

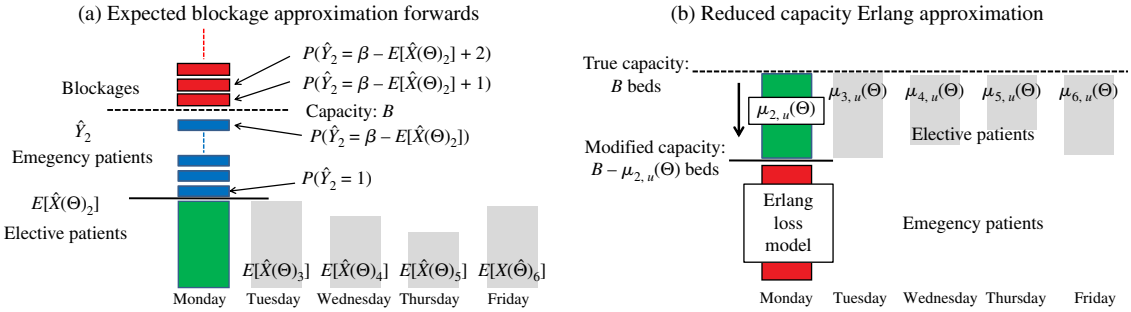
4. Optimization of Elective Admissions Mix and Volume

In §3 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. 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-ward boarding for both emergency and elective patients. The stochastic process from §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 methods for calculating various blockage metrics in a manner that can be incorporated into an integer programming formulation. Section 4.2 presents two different formulations for elective admissions mix and volume optimization that could be useful to hospitals. Section 4.3 validates the method by comparing the forecasted census from the optimization model with a high fidelity simulation of hospital operations.

4.1. 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-ward bed. 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. To capture the different blocking effects,

Figure 6. (Color online) Illustration of approximations for (a) expected off-ward census and (b) hospital level blocking probability.



we develop two linearizing approximations to capture type 1 (ward-level) and type 2 (hospital-level) blocking; illustrated in Figure 6. Both approaches begin by calculating the mean elective census by day of week (using Theorem 4), indicated by the solid bar (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).

Expected Blockages: *Ward-Level.* For type 1 blocking, shown in Figure 6(a), we start with the mean census, given by Theorem 4, and add the emergency patients, indicated by the individual bars on top of the solid bar, calculating the probability of each quantity of emergency patients present using Theorem 2. Blockages are tallied when the number of emergency patients plus the mean number of elective patients exceeds the ward capacity, B_u for ward u. This enables, for each ward, the expected amount by which demand will exceed capacity and therefore trigger off-ward servicing. This captures the dynamics in which patients are not "lost" and service continues even for blocked patients, which is a better representation of reality than a loss model or a traditional blocking model. This approximation is presented mathematically in Equations (27) and (28) of §4.2.

Reduced Capacity Erlang Approximation: *Hospital Level.* Demand from patients blocked from entering the hospital is lost. To capture this phenomenon we propose an approximation based on the Erlang loss model that we call the reduced capacity Erlang approximation, illustrated in Figure 6(b). Blocking is calculated using the Erlang loss formula on a system that has an offered load of $m_1(t)$, the emergency offered load, and number of servers $B - m_2(t)$, where $m_2(t)$ is the elective offered load:

$$\beta(s - m_2(t), m_1(t)) = \frac{m_1(t)^{s-m_2(t)}/(s - m_2(t))!}{\sum_{k=0}^{s-m_2(t)} m_1(t)^k/k!} = \frac{\mathbb{P}(C_{Em}(t) = s - m_2(t))}{\mathbb{P}(C_{Em}(t) \leqslant s - m_2(t))}. \quad (21)$$

This approximation can be linearized in the decision variable, Θ, the elective admission schedule. In §4.2, Equations (23)–(25) serve to linearly calculate to arbitrary precision $\mathbb{P}(C_{Em}(t) \leqslant s - m_2(t))$. To linearize the blocking probability constraint—limiting the blocking probability to a value less than α—simply multiply both sides of the inequality by the Erlang denominator:

$$\beta(s - m_2(t), m_1(t)) \leqslant \alpha \quad \Leftrightarrow \quad \mathbb{P}\big(C_{Em}(t) = s - m_2(t)\big) \leqslant \alpha \mathbb{P}\big(C_{Em}(t) \leqslant s - m_2(t)\big).$$

Since $\mathbb{P}(C_{Em}(t) \leqslant s - m_2(t))$ can be calculated linearly, the Erlang loss blocking constraint can be linearized. This constraint is represented by Equations (23)–(25) of §4.2.

Since the controlled cyclo-stationary system will optimize a deterministic number of elective admissions by day of week, the majority of the census variability will now come from the emergency patients, which we capture with the emergency census distribution. Hence, these approximations capture much of the stochastic dynamics that contribute to blocking. This likely explains the good accuracy of these approximations using real data that are exhibited in §4.3.

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. The cycle length we consider is days $1, \ldots, 7$ to match a typical weekly schedule.

Sets

$\mathfrak{D}$ set of elective patient diagnosis types,
$\mathfrak{W}$ set of hospital wards.

Parameters

B_u ward u capacity in terms of beds,
α limit on the blocking probability for arriving patients,
κ_u percent of total cancelations that are attributed to ward u,
$\hat{\alpha}_u$ limit on the average number of off-ward patients allowed for ward u,
$p_s^{k,u}(d)$ probability that an elective patient of type k admitted on day s is in ward u, d days after admission,

$\hat{p}^{u}_{z,d}$ probability there are z emergency patients in ward u on day d from the PATTERN PALM model,
$\tilde{p}_{z,d}$ probability there are z 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,
$\delta^{1}_{z,d}$ indicator of whether z emergency patients in the hospital on day d would exceed capacity,
$\delta^{2}_{z,d}$ indicator of whether z emergency patients in the hospital on day d would exceed capacity minus one,
$\hat{\delta}^{u}_{z,d}$ number of ward u off-ward patients on day d if there are z emergency patients in ward u.

It is important to note here that the probabilities $p^{k,u}_{s}(d)$, $\hat{p}^{u}_{z,d}$, and $\tilde{p}_{z,d}$ are all calculated offline per the analysis in §§3.3 and 3.4 and then become data inputs to the two mixed-integer programs that follow.

4.2.1. Maximum Elective Admissions Formulation. First we present a formulation that maximizes the weighted throughput of elective admissions subject to constraints on bed blockage; $\mathbf{1}$ denotes a column vector of all ones and M is a large number. For the sake of generality we include the "reward" row vector, $\mathbf{R}$, providing a relative value of admitting a patient of type k. Our validation sets $\mathbf{R}$ to be a row of all 1's (every patient type has the same value) and then manipulates the constraints if management's goal is to increase the volume of one particular service:

$$\max_{\Theta,\delta,\hat{\delta}} \ \mathbf{R}\cdot\Theta\cdot\mathbf{1} \tag{22}$$

$$\text{s.t.} \quad M\delta^{1}_{z,d_1} \geqslant z - \sum_{u\in\mathfrak{W}}\left(B_u - \sum_{d_2=1}^{7}\sum_{k\in\mathfrak{D}}\Theta_{k,d_2} \cdot \sum_{n=0}^{\infty} p^{k,u}_{d_2-7n}(d_1-d_2+7n)\right) \quad d_1=1,\ldots,7,\ z=1,2,\ldots, \tag{23}$$

$$M\delta^{2}_{z,d_1} \geqslant z - \sum_{u\in\mathfrak{W}}\left(B_u - \sum_{d_2=1}^{7}\sum_{k\in\mathfrak{D}}\Theta_{k,d_2} \cdot \sum_{n=0}^{\infty} p^{k,u}_{d_2-7n}(d_1-d_2+7n)\right)+1 \quad d_1=1,\ldots,7,\ z=1,2,\ldots, \tag{24}$$

$$\sum_{z=0}^{\infty}\tilde{p}_{z,d}\left(\delta^{1}_{z,d}-\delta^{2}_{z,d}\right) \leqslant \alpha\left(1-\sum_{z=0}^{\infty}\tilde{p}_{z,d}\delta^{1}_{z,d}\right) \quad d=1,\ldots,7, \tag{25}$$

$$\delta^{i}_{z+1,d} \geqslant \delta^{i}_{z,d} \quad i=1,2,\ d=1,\ldots,7,\ z=1,2,\ldots, \tag{26}$$

$$\hat{\delta}^{u}_{z,d_1} \geqslant z + \sum_{d_2=1}^{7}\sum_{k\in\mathfrak{D}}\Theta_{k,d_2}\cdot\sum_{n=0}^{\infty}p^{k,u}_{d_2-7n}(d_1-d_2+7n) - B_u - \kappa_u\sum_{d=1}^{7}\sum_{l=0}^{\infty}\tilde{p}_{l,d}\sum_{z=0}^{l}\delta^{1}_{z,d} \quad \forall u\in\mathfrak{W},\ d_1=1,\ldots,7,\ z=1,2,\ldots, \tag{27}$$

$$\sum_{z=0}^{\infty}\hat{p}^{u}_{z,d}\hat{\delta}^{u}_{z,d} \leqslant \hat{\alpha}_u \quad \forall u\in\mathfrak{W},\ d=1,\ldots,7, \tag{28}$$

$$\hat{\delta}^{u}_{z+1,d} \geqslant \hat{\delta}^{u}_{z,d} \quad d=1,\ldots,7,\ z=1,2,\ldots, \tag{29}$$

$$\sum_{d=1}^{7}\Theta_{k,d} \geqslant \sum_{d=1}^{7}\theta_{k,d} \quad \forall k\in\mathfrak{D}, \tag{30}$$

$$\Theta_{k,d} \leqslant \hat{\theta}_{k,d} \quad \forall k\in\mathfrak{D},\ d=1,\ldots,7, \tag{31}$$

$$\Theta_{k,d},\hat{\delta}^{u}_{z,d}\in\mathbb{Z}^{+},\delta^{1}_{z,d},\delta^{2}_{z,d}\in\{0,1\} \quad \forall k\in\mathfrak{D},\ u\in\mathfrak{W},\ z\in\mathbb{N},\ d=1,2,\ldots,7. \tag{32}$$

The objective function, Equation (22), maximizes the weighted throughput of elective patients. Constraints 23–25 calculate the reduced capacity Erlang approximation described in §4.1 by setting indicator decision variables: $\delta^{1}_{z,d}=\mathbf{1}\{z$ emergency patients exceeds hospital capacity on day $d\}$ in Equation (23) and $\delta^{2}_{z,d}$ in Equation (24), which is the same except that capacity is reduced by 1. Thus

$$\sum_{z=0}^{\infty}\tilde{p}_{z,d}\delta^{1}_{z,d} = \mathbb{P}(\text{Emerg Load} \geqslant \text{Capacity}) \quad \text{and}$$

$$\sum_{z=0}^{\infty}\tilde{p}_{z,d}\delta^{2}_{z,d} = \mathbb{P}(\text{Emerg Load} \geqslant \text{Capacity}-1).$$

That means that the LHS of Equation (25) is precisely $\mathbb{P}(\text{Emerg Load} = \text{Capacity})$. The RHS is therefore $\mathbb{P}(\text{Emerg Load} \leqslant \text{Capacity})$, completing the Erlang loss B formula described in §4.1. Note that

$$\sum_{d_2=1}^{7}\sum_{k\in\mathfrak{D}}\Theta_{k,d_2}\cdot\sum_{n=0}^{\infty}p^{k,u}_{d_2-7n}(d_1-d_2+7n)$$

is $\mu_{d_1,u}(\Theta)$ from the analysis of the CALM model in §3.4.2 (Theorem 4) and $\tilde{p}_{n,d}$ is the Poisson probabilities from Theorem 2 of §3.3. Constraints 27–28 mathematically capture the expected off-ward census approximation detailed in §4.1. In constraint 27, note that the term $\kappa_u\sum_{d=1}^{7}\sum_{l=0}^{\infty}\tilde{p}_{l,d}\sum_{z=0}^{l}\delta^{1}_{z,d}$ is subtracted from $\mu_{d_1,u}(\Theta)$, which accounts for the fact that, if patients are blocked from the hospital, they will not contribute to ward demand and off-ward census. The parameter κ_u refers to the historical trend and/or hospital protocols for what types of patients get canceled when a cancelation decision must be made. Note that, although we use infinite sums to represent expectations and blocking probabilities, truncating these to finite sums is necessary. One approach is to truncate based on the product

of the overage and the Poisson probabilities (which decay quickly two standard deviations above the mean) to limit the expected overage error. The same truncation scheme is used for both the z index (Equations (23), (24), (27)) and the n index (Equations (23)–(29)) since the two are linked through the δ and $\hat{\delta}$ helper variables.

Constraints 26 and 29 are cuts that were added to the model to increase solution speed by greatly reducing the number of combinations that must be considered by branch and bound. The meaning of Equation (26) (Equation (29)) is straightforward: if an n patient emergency load exceeds capacity (exceeds capacity by a certain amount) then an $n+1$ patient emergency load will also exceed capacity (exceed by at least as much). Without this constraint, a model with three wards and three patient types failed to solve in under 24 hours, whereas solving in under 30 seconds with the constraint.

The final two constraints, Equations (30) and (31), represent the reality that the model should not change the elective admission schedule in ways incommensurate with historical hospital practice. Specifically Equation (30) ensures that, under the improved plan, no service's planned volume is reduced from their historical levels (a feature easily controlled if a specialty OR service needs to target a different load). Equation (31) allows the model to enforce capacity constraints beyond hospital beds. For example, limiting the amount of operating room time, or enforcing the typical practice to admit few or no elective patients on the weekends (e.g., $\Theta_{k,\,\text{Sunday}} \leqslant 0\ \forall k$).

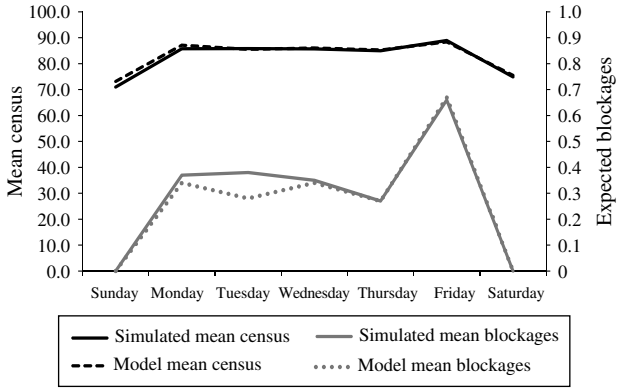
4.2.2. Minimum Blockages Formulation. Another useful formulation is to keep the weekly volume of elective admissions fixed and attempt to minimize the number of blockages. This model reshuffles the mix of elective admissions across the days of the week to eliminate unnecessary blockages caused by an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

$$\min_{\Theta,\,\delta,\,\hat{\delta}} \sum_{d=1}^{7} \sum_{l=0}^{\infty} \tilde{p}_{l,d} \sum_{z=0}^{l} \delta^{1}_{z,\,d}. \tag{33}$$

We no longer need the blocking constraints defined by Equations (23), (25), and (26) would only apply to $\delta^{1}_{z,\,d}$, since $\delta^{2}_{z,\,d}$ is no longer needed as a decision variable.

Another possible objective that our framework is capable of capturing is minimizing the day-to-day variability of bed census. For example, it would be possible to minimize the gap between the largest and smallest daily census within the MIP.

Figure 7. Simulation output vs. stochastic model output for characteristic hospital measures.



4.3. Validating the Hospital Census Optimization Model

As in §3.5, 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 Helm et al. 2011, 2010, 2009).

A year's worth of historical data from a medium-sized, nonteaching partner hospital 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 reduced the amount of data analysis and cleaning without degrading the value of the study.

Figure 7 and Table 5 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

Table 5. Simulation output vs. stochastic model output for characteristic hospital measures.

	Sunday	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday	Total
Simple census	71.0	85.8	85.8	85.7	85.0	88.9	74.9	N/A
Approximate census	73.2	87.1	85.5	86.1	85.3	88.4	75.5	N/A
Different 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
Approximate blockages	0.00	0.34	0.28	0.34	0.27	0.67	0.00	1.9
Different blockages	0.0	0.03	0.1	0.01	0.0	0.01	0.0	0.13

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; average weekly blockages only differ by 0.13 patients in absolute value and the census differs on average by only 1%. Because the stochastic census model is an accurate approximation, a detailed (and slow) simulation is *no longer needed* to express the trade-offs between census and blockages to design effective admission schedules.

4.4. Case Study, Proof of Concept, and Improving Management Practice

To demonstrate the effectiveness and potential uses of our approach to elective admissions scheduling, we validate our method for the partner hospital of §4.3 by comparing our optimized schedules with the current schedule for overnight patients. In 2008, 14,827 patients stayed at least one night, of which 7,016 were emergency patients and the remaining 7,811 were scheduled patients. Patients transferred within the hospital 20,462 times for an average of 1.4 transfers per patient, which underscores the importance of modeling ward network effects. The nine wards we model comprise about 60% of the total patient volume with similar characteristics to the total patient population.

One of our primary goals when developing our new modeling methodology was to address patient blockage (see §4.2.2), 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* of §4.2.2 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 a 32% *reduction in average cancelations per week*, shown as the data point at the end of the down arrow in Figure 8(b). This reduction results from reducing the midweek occupancy and smoothing the census as seen in Figure 8(a), which compares the original and optimal census curves.

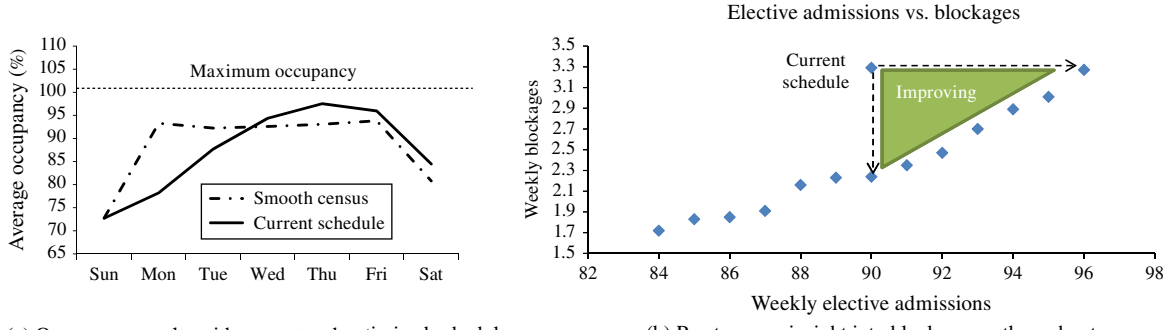
Some hospitals with growth potential will want to increase the volume of patients served while maintaining the same level of access. To achieve this, the *maximum admissions formulation* of §4.2.1 is employed, constraining the blockages to be less than or equal to the current (3.29 per week) and maximizing the number of admissions. This included Constraint 30 to ensure that each service is given at least as many elective admissions as in the current schedule. The data point at the end of the right arrow in Figure 8(b) shows an *additional* 310 *elective admissions per year* (six per week) with slightly better access (3.27 blockages per week).

Remark 4. Because of Little's Law (WIP = Throughput ∗ Cycle Time) and Constraint (30) that maintains (or increases) the historical average number of patients scheduled per unit time, the average delay to obtain an admission for scheduled patients will not increase.

Improving Management Practice. These prescriptive models are useful in exploring the boundaries of hospital efficiency, but hospitals may prefer a balance between volume and blockage, which our method can also provide. The Pareto curve in Figure 8(b) presents the trade-off 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 the Pareto 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. The computational speed with which the analytical model generates this curve represents an advance in decision support that enables hospital administrators to understand the key trade-offs involved in scheduling their admissions and gives them the freedom to choose their desired operating point.

Figure 8. (Color online) Controlling census and occupancy variability in hospitals and obtaining trade-off insights with our methodology.



(a) Occupancy results with current and optimized schedules

(b) Pareto curve insight into blockage vs. throughput

Our optimization models were able to generate the operating curve automatically in a matter of minutes, with each point taking about 30 seconds.

The key implication is a capability (which we believe is both highly important and not currently possible) to make strategic admission policies at the hospital level regarding the trade-off between throughput and access. This trade-off is the key to cost, quality, and access, and the Pareto curve provides hospital managers the ability to make informed decisions that affect this trade-off.

Implementing the Strategic Admission Plan. As mentioned in §3.1, the output of the optimization is a set of decision rules that guide the maximum number of admissions of each type of patient that can be admitted on a given day of the admission cycle. To avoid being too restrictive (and risk increased barriers to acceptance) an implementation should allow the admissions personnel to fill the slots with whatever patients they would like to schedule as long as they do not exceed the admission numbers by patient type. Just as with most other appointment systems, when all the slots fill up on a given day the patient must be scheduled for a different day with slots available.

It is also important to note that our model can go far beyond what has been presented in the case study. For example, the model can be used to design unique schedules for each season—many hospitals have a low season and a high season in terms of patient demand—by solving one optimization for each season modeled. Another way to incorporate projected changes in demand (e.g., hiring more surgeons) is to extend the cycle length (which can be adjusted to any desired cycle length). Demand forecasts for emergency patients can also be incorporated by modifying the arrival rates over an admission cycle that matches the demand forecast horizon.

5. Conclusions and Future Work

We have developed new models for a longstanding 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 less 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, mitigating 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-ward 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 d-CALM for elective admissions and analyzed its properties.

Through the validation process of the hospital census model in §3.5 and the analysis of optimization results, we have generated a number of managerial insights in addition to the methodological contributions. The results of our analysis and case study show that (1) smooth census levels do in fact improve blocking and throughput performance, and (2) elective admission scheduling can be used to smooth census while constraining the amount of planned off-ward placement across the network of hospital resources. Thus, from a managerial perspective, the goal should be to design admission plans that drive toward smooth census levels even in the absence of an optimization model. Further, our validation indicated that a cyclo-stationary (weekly) model with a deterministic elective admissions schedule can be a fairly accurate model of actual hospital flows even when tightly controlled elective admissions processes are not in place. Although hospital blocking and off-ward placement are complicating realities that contaminate historical data, we showed that one can often use a system of linear equations to parameterize a corrected flow model from historical data. Our results provide strong evidence that model-based analysis using historical data can predict and smooth census by day of week (or cycles of arbitrary lengths).

Methodologically, the novel analytical models developed are much more tractable and powerful than prior models, allowing the generation of trade-off curves for hospital blocking versus throughput. This curve represents an effective decision-making tool for hospital administrators, because it enables flexibility and choice rather than a fixed solution. This approach is likely to increase acceptance by administrators, enabling them to make important decisions based on deeper managerial insights and quantitative analysis.

Supplemental Material

Supplemental material to this paper is available at http://dx.doi.org/10.1287/opre.2014.1317.

Acknowledgments

The authors wish to thank the reviewers, the area editor, and the department editor for helpful comments that greatly improved the technical elements of this paper. This work was supported in part by the National Science Foundation [Grant CMMI-1068638].

References

Adan I, Bekkers J, Dellaert N, Vissers J, Yu X (2009) Patient mix optimisation and stochastic resource requirements: A case study in cardiothoracic surgery planning. *Health Care Management Sci.* 12(2):129–141.

Aiken LH, Clarke SP, Sloane DM, Sochalski J, Silber JH (2002) Hospital nurse staffing and patient mortality, nurse burnout, and job dissatisfaction. *JAMA* 288(16):1987–1993.

Anderson P, Meara J, Brodhurst S, Attwood S, Timbrell M, Gatherer A (1988) Use of hospital beds: A cohort study of admissions to a provincial teaching hospital. *BMJ* 297(6653):910–912.

Bekker R, Koeleman PM (2011) Scheduling admissions and reducing variability in bed demand. *Health Care Management Sci.* 14(3):237–249.

Brownson K, Dowd SB (1997) Floating: A nurse's nightmare? *The Health Care Management* 15(3):10–15.

Chow VS, Puterman ML, Salehirad N, Huang W, Atkins D (2011) Reducing surgical ward congestion through improved surgical scheduling and uncapacitated simulation. *Production Oper. Management* 20(3): 418–430.

Connors MM (1970) A stochastic elective admissions scheduling algorithm. *Health Serv. Res.* 5(4):308–319.

Derlet RW, Richards JR, Kravitz RL (2001) Frequent overcrowding in U.S. emergency departments. *Acad. Emergency Medicine* 8(2):151–155.

Fatovich DM, Nagree Y, Sprivulis P (2005) Access block causes emergency department overcrowding and ambulance diversion in Perth, Western Australia. *British Medical J.* 22(5):351–354.

Forster AJ, Stiell I, G. Wells, Lee AJ, Van Walraven C (2003) The effect of hospital occupancy on emergency department length of stay and patient disposition. *Acad. Emergency Medicine* 10(2):127–133.

Gallivan S, Utley M (2005) Modelling admissions booking of elective in-patients into a treatment centre. *IMA J. Management Math.* 16(3):305–315.

Gallivan S, Utley M, Treasure T, Valencia O (2002) Booked inpatient admissions and hospital capacity: Mathematical modelling study. *British Medical J.* 324(7332):280–282.

Griffith JR, Hancock WM, Munson FC (1978) *Cost Control in Hospitals* (Lippincott Williams & Wilkins, Philadelphia).

Hancock WM, Walter PF (1979) The use of computer simulation to develop hospital systems. *SIGSIM Simul. Dig.* 10(4):28–32.

Hancock WM, Walter PF (1983) *The "ASCS": Inpatient Admission Scheduling and Control System* (Health Administration Press, Ann Arbor, MI).

Harper PR (2002) A framework for operational modelling of hospital resources. *Health Care Management Sci.* 5(3):165–173.

Harrison GW, Shafer A, Mackay M (2005) Modelling variability in hospital bed occupancy. *Health Care Management Sci.* 8(4):325–334.

Helm JE, AhmadBeygi S, Van Oyen MP (2009) The flexible patient flow simulation framework. *Proc. 2009 IIE IERC Conf.*

Helm JE, AhmadBeygi S, Van Oyen MP (2011) Design and analysis of hospital admission control for operational effectiveness. *Production Oper. Management* 20(3):359–374.

Helm JE, Lapp M, See BD (2010) Characterizing an effective hopsital admission scheduling and control management system: A genetic algorithm approach. *Proc. 42nd Winter Sim. Conf* (IEEE, Piscataway, NJ), 2387–2398.

Hoot NR, Aronsky D (2008) Systematic review of emergency department crowding: causes, effects, and solutions. *Ann. Emergency Medicine* 55(2):126–136.

Jun JB, Jacobson SH, Swisher JR (1999) Application of discrete-event simulation in health care clinics: A survey. *J. Oper. Res. Soc.* 50(2): 109–123.

Leung KK, Massey WA, Whitt W (1994) Traffic models for wireless communication networks. *Selected Areas Comm., IEEE J.* 12(8): 1353–1364.

Liu Y, Whitt W (2011) A network of time-varying many-server fluid queues with customer abandonment. *Oper. Res.* 59(4):835–846.

Massey WA, Whitt W (1993) Networks of infinite-server queues with nonstationary Poisson input. *Queueing Syst.* 13(1):183–250.

Massey WA, Whitt W (1994a) A stochastic model to capture space and time dynamics in wireless communication systems. *Probability Engrg. Informational Sci.* 8(4):541–569.

Massey WA, Whitt W (1994b) An analysis of the modified offered-load approximation for the nonstationary erlang loss model. *Ann. Appl. Probability* 4(4):1145–1160.

McManus ML, Long MC, Cooper A, Mandell J, Berwick DM, Pagano M, Litvak E (2003) Variability in surgical caseload and access to intensive care services. *Anesthesiology* 98(6):1491–1496.

Mirel LB, Carper K (2013) Expenses for hospital inpatient stays, 2010. *AHRQ, Statist. Brief 401*. http://meps.ahrq.gov/mepsweb/data_files/publications/st401/stat401.shtml.

Needleman J, Buerhaus P, Mattke S, Stewart M, Zelevinsky K (2002) Nurse-staffing levels and the quality of care in hospitals. *New England J. Medicine* 346(22):1715–1722.

Proudlove NC, Gordon K, Boaden R (2003) Can good bed management solve the overcrowding in accident and emergency departments? *British Medical J.* 20(2):149–155.

Richardson DB (2006) Increase in patient mortality at 10 days associated with emergency department overcrowding. *Medicine J. Australia* 184(5):213–216.

Sprivulis PC, Da Silva J, Jacobs IG, et al. (2006) The association between hospital overcrowding and mortality among patients admitted via Western Australian emergency departments. *Med. J. Australia* 184:208–212.

Jonathan E. Helm is an assistant professor of operations and decision technologies at the Kelley School of Business, Indiana University. His research interests include capacity management in healthcare networks, patient flow modeling and optimization, patient monitoring for chronic disease, and distribution of medical aid in developing countries. Recently, he has been studying mechanisms to reduce the impact of hospital readmissions at both the operational and policy levels. His awards and honors include an invitation to give a Showcase Presentation at the 2014 POMS CHOM Mini Conference, he was finalist (2nd place) in the 2013 INFORMS Data Mining Best Student Paper Competition, he received first prize in the 2012 INFORMS "Doing Good with Good OR," and he received first prize in 2011 and he was the finalist in 2012 for the POMS CHOM best paper competition.

Mark Van Oyen is a professor of industrial and operations engineering (IOE) at the University of Michigan, which he joined in 2005. His research includes the analysis, design, control, and management of operational systems, with emphasis on performance, flexibility, and controlled queueing network models. Healthcare operations research and medical decision making are key application areas. His awards *with student advisees* include first prize in the 2012 MSOM Student Paper Competition, the 2012 INFORMS "Doing Good with Good OR" first prize, and two finalist papers with MSOM. He also coauthored papers that won the 2010 Pierskalla Award, second prize in the 2013 POMS CHOM best paper competition, and first and second prize winning papers in the 2011 POMS CHOM competition. He was the IOE Department Faculty of the Year for 2008–2009, an ALCOA Manufacturing Systems Faculty Fellow, and the 2003 Researcher of the Year at Loyola University Chicago's School of Business.