Dynamic Assignment of Patients to Primary and Secondary Inpatient Units: Is Patience a Virtue?

Derya Kilinc[1], Soroush Saghafian[2], Stephen J. Traub[3]
[1]School of Computing, Informatics, and Decision Systems Engineering, Arizona State University, Tempe, AZ 85287
[2]Harvard Kennedy School, Harvard University, Cambridge, MA 02138
[3]Department of Emergency Medicine, Mayo Clinic, Phoenix, AZ 85054

An important contributor to the well-known problem of Emergency Department (ED) overcrowding is prolonged boarding of patients who are admitted through the ED. Patients admitted through the ED constitute about 50% of all non-obstetrical hospital admissions, and may be boarded in the ED for long hours with the hope of finding an available bed in their primary inpatient unit. We study effective ways of reducing ED boarding times by considering the trade-off between keeping patients in the ED and assigning them to a secondary inpatient unit. The former can increase the risk of adverse events and cause congestion in the ED, whereas the latter may adversely impact the quality of care. Further complicating this calculus is the fact that a secondary inpatient unit for a current patient can be the primary unit for a future arriving patient; assignments, therefore, should be made in an orchestrated way. Developing a queueing-based Markov decision process, we demonstrate that patience in transferring patients is a virtue, but only up to a point. We also find that, contrary to the prevalent perception, idling inpatient beds can be beneficial. Since the optimal policy for dynamically assigning patients to their primary and secondary inpatient units is complex and hard to implement in hospitals, we develop a simple policy which we term penalty-adjusted Largest Expected Workload Cost (LEWC-p). Using simulation analyses calibrated with hospital data, we find that implementing this policy could significantly help hospitals to improve their patient safety by reducing boarding times while controlling the overflow of patients to secondary units. Using data analyses and various simulation experiments, we also help hospital administrators by generating insights into hospital conditions under which achievable improvements are significant.*

Key words: ED boarding; inpatient flow management; patient safety; healthcare operations; Markov decision process
History: Version: January 31, 2018

"Ah, all things come to those who wait." (Violet Fane (1843-1905))

1. Introduction

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

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

One important factor associated with ED overcrowding is the prolonged ED boarding of patients admitted to inpatient units (GAO (2003)). ED boarding occurs when an admitted patient waits for transfer to an inpatient unit due to bed unavailability in a downstream unit. Although this may cause congestion and may block ED resources from being assigned to newly arrived patients (see, e.g., Saghafian et al. (2012), and the references therein for the so-called "bed-block" effect), boarding

* This work was supported in part by the Mayo Clinic through grant XSS0133.

may be viewed in a positive light by some when it is done to ensure transfer to the most appropriate inpatient unit (rather than a secondary unit that has bed availability). This is due to a questionable yet prevalent belief that patience in transferring admitted patients is always a virtue. This belief deserves further scrutiny, especially because it is well understood that prolonged boarding times have several negative consequences for patients, including an increased risk of adverse events (ROAE).[1]

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

Overflowing is not a new concept in hospitals; however, in practice the overflow process is often controlled in a myopic manner without much attention to the needs of future patients. Instead, what is needed is a reasonable balance between the risk of keeping a patient in the ED (with the hope of a primary unit assignment) vs. that of assigning the patient to a secondary unit that has current bed availability. A careful consideration of these trade-offs might have a significant impact on both patient safety and operational efficiency of hospitals. Our goal in this paper is to develop a systematic approach to facilitate better decision making with respect to inpatient unit assignments.

To gain insights, we explore these issues in our partner hospital, Mayo Clinic Arizona (MCA). There are eight inpatient wards (IWs) in MCA from which a bed can be requested for an admitted patient. A detailed description of these eight IWs are shown in Table 1. The data we have collected from MCA shows that the average ED boarding time (the average time between bed request and occupancy) at MCA is 111 minutes, with boarding times up to 150 minutes for some patients. Moreover, we observe from our data that about 30% of the patients admitted through the MCA ED are boarded for at least two hours. An average delay of 111 minutes is significant, especially when we consider that the average ED Length of Stay (LOS) for admitted patients in MCA is

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

Table 1 IWs and their size in MCA

IW Name	Abbrev.	Definition	Number of Beds
2 West	2W	Intensive Care Unit (ICU)	30
3 East	3E	Orthopedics and Urology Surgical Services	40
3 West	3W	Medical/Surgical Organ Transplant	36
4 East	4E	Bone Marrow Transplant, Hematology, Oncology	30
4 West	4W	Cardiology and Cardiothoracic Surgery	36
5 West	5W	Neurosciences and E.N.T.	36
7 East	7E	Palliative Care, General Surgery	36
7 West	7W	Hematology and Oncology patients with medical-surgical overflow	24

about 5 hours. This suggests that, on average, almost 37% of ED LOS is caused by boarding.[2] Furthermore, as is shown in Figure 1(a), we find that boarding duration is highly time-dependent. Therefore, even if the average waiting time is not extremely long, patients admitted through the ED experience different levels of delay based on the hour in which their inpatient bed is requested.

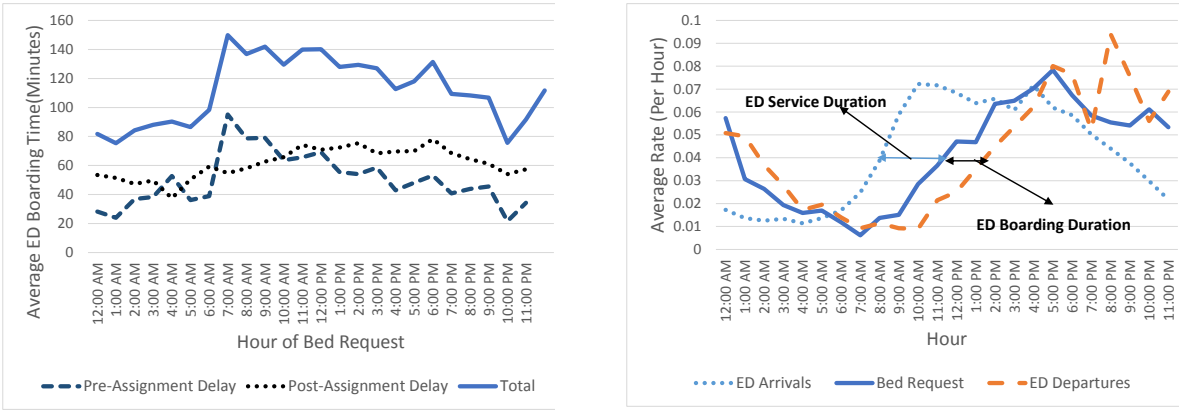
As we illustrate in Figure 1(a), boarding delays consists of two parts: Pre-Assignment and Post-Assignment. Pre-Assignment delay is the time between bed request and assignment of a suitable inpatient bed to the patient. Post-assignment delay is the time between bed assignment and bed occupancy. Our analyses of MCA data reveal that post-assignment delays are higher on average than pre-assignment delays (see Figure 1(a)). Additionally, as we show in Figure 1(b), there is a significant mismatch (i.e., time lag) between the hourly bed request pattern and the ED departure pattern (see, e.g., Shi et al. (2015), Armony et al. (2015), and Powell et al. (2012) for related results reported for other hospitals). The time between ED arrivals and ED departures in Figure 1(b) is defined as ED LOS, and the time between bed requests and ED departures represents the ED boarding time. As can be seen from this figure, the ratio of ED boarding time to ED LOS can be as high as 48% for some patients.

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

- *Structure:* When should a patient be kept in the ED until a bed becomes available in his/her primary inpatient unit instead of being quickly assigned to a secondary unit with current bed availability?
- *Magnitude:* How much improvement can be achieved if a hospital adopts an effective policy for dynamically assigning ED admitted patients to their primary or secondary inpatient units?

To gain insights and answer these questions, we start by utilizing a Markov decision process (MDP) and modeling the flow process as a multi-class queueing network problem with "flexible"

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



(a) Average boarding time

(b) Non-stationary arrival, bed request, and departure rates

Figure 1 ED boarding times based on collected data from our partner hospital

servers. In this model, the servers are defined as the downstream inpatient unit beds that are "flexible," in that they can serve different classes of patients. The literature on hospital-like multi-class queueing systems with flexible servers that can address the appropriateness of bed assignment decisions is not vast. We contribute to this literature by considering (a) a stochastic penalty cost that reflects the reduction in service quality when a patient is assigned to a secondary inpatient unit, and (b) stochastic risk of adverse events that can occur due to prolonged ED boarding times. By analyzing our MDP setting, we find that the optimal assignment policy is a state-dependent threshold-type policy: keeping patients in the ED for their primary inpatient unit to become available pays off, but only up to a certain threshold that depends on the number and status of outstanding ED bed requests. That is, *patience is a virtue, but only up to a point*.

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

We also use our simulation framework to generate insights into hospital conditions under which such improvements can be most significant. Our results suggest that hospitals with higher congestion levels (e.g., busy teaching hospitals) would benefit more than other hospitals (e.g., less busy community hospitals) from utilizing our proposed policy as a way to strike a better balance between patient safety, quality of care, and operational efficiency. Our results also suggest that, under specific conditions on adverse event rates and number of patients boarded in the ED, keeping an inpatient bed idle for potential future bed requests is beneficial. This practice of intentional bed idling is currently used in some inpatient units such as the ICU. However, our results provide support for implementation across a wider range of inpatient units, and reveal that bed idling should be used more broadly in hospitals.

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

The rest of this paper is organized as follows. Section 2 reviews the related studies on patient flow and dynamic assignment policies. Section 3 presents a model of patient flow, and develops an MDP framework that captures the trade-offs in the model. Section 4 identifies the structure of the optimal policy. In Section 5, we describe our proposed heuristic bed assignment policy, and compare its performance with the optimal policy. In Section 6, we describe our detailed simulation model of patient flow, and use it to perform various sensitivity analyses. Finally, we present our concluding remarks in Section 7. All proofs are provided in Online Appendix A.

2. Literature Review

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

2.1. Related Studies on ED Patient Flow

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

Harrison et al. (2005) use discrete-event simulation to analyze the effect of bed capacity on overflow rates. The authors indicate that seasonality of arrivals is one of the main triggers of overflow in hospitals. Thompson et al. (2009) study a capacity utilization-based patient allocation problem. In their model, patients may be transferred between different units to minimize the total cost under a preemptive service policy assumption, where assignment to each unit is accompanied by a reward/cost. Similar to Thompson et al. (2009), we consider different levels of quality of care that can be provided in different inpatient units. However, unlike that study, we also model the risk of adverse events (ROAE) that can occur because of prolonged waiting in the ED. This allows us to provide a system-wide view that, in addition to operational efficiency, considers both patient safety and quality of care concerns. Another related study is Mandelbaum et al. (2012), which considers the fair routing of patients to inpatient units, where fair routing means targeting the same level of idleness among all servers. Unlike Mandelbaum et al. (2012), we consider patient routing as a mechanism to eliminate prolonged ED boarding times. Furthermore, the study of Mandelbaum et al. (2012) analyzes a model with a single customer class, whereas we consider heterogeneous patient classes in order to gain insights into the questions we raised in the Introduction.

Teow et al. (2012) use data mining techniques to identify factors that trigger overflow decisions. Unlike Teow et al. (2012), our study attempts to identify conditions under which it is optimal to overflow a patient to a secondary inpatient unit. Shi et al. (2015) focus on patient flow from ED to inpatient units, and propose early discharge policies in inpatients units as a mechanism to reduce and flatten ED boarding times. Our study focuses on a similar patient flow from the ED to inpatient units; however, unlike the predetermined trigger times in Shi et al. (2015), we (a) optimize bed assignment decisions based on the number of boarded patients in the ED, and (b) consider both patient safety and quality of care metrics. Furthermore, a policy of changing physician discharge

routines that is described in Shi et al. (2015) might be hard to implement in many hospitals due to cultural issues such as difference in physicians' preferences. Our study offers guidelines on alternative ways of improving the patient flow.

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

2.2. Related Studies on Dynamic Assignment and Routing in Queueing Systems

Our model captures the system characteristics as a multi-class queueing system where the bed requests for ED admitted patients are considered as arrivals, and inpatient unit beds are considered as servers. In multi-class queueing systems, the customers can be differentiated based on service rates, holding costs, arrival rates, or service requirements. Under an average holding cost objective, Cox and Smith (1961) demonstrate that the widely-used $c\mu$ policy is optimal for both preemptive and non-preemptive cases service protocols. The $c\mu$ policy is also shown to be the optimal policy in various more complex queueing networks (see, e.g., Kakalik and Little (1971), Buyukkoc et al. (1985), and Walrand (1988)). A version of the $c\mu$ rule, generalized $c\mu$, is proved to be the optimal policy for different queueing structures under heavy traffic (see, e.g.,Van Mieghem (1995), Mandelbaum and Stolyar (2004)). Saghafian and Veatch (2016) establish the optimality of the $c\mu$ rule for queueing systems with flexible servers and two tier structures, where one tier is served by one server while the second tier can be served by all the servers.

In Lin and Kumar (1984), the authors show that when two types of servers with different service speeds are available—a setting termed "slow server problem"—the optimal assignment policy is a threshold-type policy: customers/jobs are assigned to the slow server whenever the queue length reaches a certain threshold. Our model resembles similar characteristics to the "slow server problem," because (a) patient service times in inpatient units are not identical, and (b) there is some flexibility in assignments (for some patients). However, instead of heterogeneous servers, we consider heterogeneous patient types with different service rates, since it better matches the hospital

patient flow we study. This differentiates our study from the above-mentioned studies in the literature since in such studies the resulted optimal policy typically depends on the difference between service rates of servers (see, e.g., Bell and Williams (2001)). However, our data analysis shows that service durations in primary and secondary units are not statistically different (for patients of the same type).

Dynamic assignment problems in queueing networks are extensively analyzed in the literature (see, e.g., Mandelbaum et al. (2012), Meyn (2001, 2003), and Palmer and Mitrani (2004)). Armony and Bambos (2003) and Dai and Lin (2005) study dynamic assignment problems considering a throughput maximization objective. Andradóttir et al. (2007) and Saghafian et al. (2011) allow for server disruptions and repairs in systems with heterogeneous flexible servers, and De Véricourt and Zhou (2005) study a call center setting where agents are heterogeneous in terms of both service rate and quality of service (see also Zhan and Ward (2013)). Another related stream of literature that considers flexible servers is the "skill-based routing" literature, where the customers are routed to the servers that have the appropriate skill sets (similar to the routing of patients to primary vs. secondary units in our study). However, unlike our work, the focus of those studies are mostly on settings where (a) servers have multiple skills (e.g., call center agents), and (b) staffing decisions are the primary concerns (see, e.g., Garnett and Mandelbaum (2000), Gans et al. (2003), Wallace and Whitt (2005)). There are also various other studies on routing policies in multi-server, multi-class settings (see, e.g., Gurvich and Whitt (2009), Tezcan and Dai (2010), Armony and Ward (2010), Gurvich and Perry (2012)). However, in these studies only costs related to waiting and losing customers are considered, whereas we focus on the trade-off between waiting and overflows. Moreover, we note that the majority of the above-mentioned studies focus on heavy traffic settings. Unlike them, we seek to address the questions we raised in the Introduction under practical hospital congestion levels. To this end, we do not impose any heavy traffic assumption, and instead make use of actual hospital bed census data as the basis of our analytical and simulation analysis.

3. The Model

A general representation of patient flow through the ED and hospital inpatient wards (IWs) is presented in Figure 2. A patient that arrives to the ED goes through the triage stage, and is assigned an Emergency Severity Index (ESI). If there is an examination room available, the patient immediately starts the ED service; otherwise, he/she will have to wait in a designated ED waiting area. Once the ED treatment is done, the patient is either discharged home or is admitted to the hospital. For an admitted patient, if there is a bed available in his/her primary IW (or the secondary IW if applicable), the patient is transferred out of the ED; otherwise, he/she is kept in the ED until a bed

becomes available. For the goals of this study, we focus on the patient flow within the dashed area of Figure 2.[3]

To gain insights into the questions we raised in the Introduction, we start by modeling the patient flow as a multi-class queueing system with IWs as flexible servers, and analyze it by using an MDP. The patients in the system are classified based on their primary IW, i.e., where they can be best served from a medical standpoint. Ward-level placement is typically determined by a bed placement coordinator, sometimes in consultation with the ED or the admitting physician. Once a patient is moved to an IW, the IW bed is considered as unavailable until the patient is done with the inpatient unit service, and hence, the service processes in IWs are typically non-preemptive. To gain some high level insights, we start by considering each of the IWs as a single "super server," which represents the capacity of the IW as a whole. This pooling of beds within each IW allows us to keep track of availability of capacity in IWs in a computationally tractable way. However, to test the insights we gain from this simplifying assumption, we relax it in Section 6, and consider each IW bed as a server. Similarly, we start by considering the arrival process as a stationary Poisson Process, and assume IW service times are exponential. In Section 6, we also relax these simplifying assumptions by using empirical distributions (for both interarrival and service times) that we have estimated based on our data.

Figure 3 illustrates the patient flow under consideration as a queueing system. Our discussions with medical providers revealed that, for the vast majority of patients, only one IW can be considered as a secondary IW.[4] Hence, as illustrated in Figure 3, the system consists of multiple primary-secondary pairs, where each patient type has only one primary IW and only one secondary IW.

To analyze the patient flow depicted in Figure 3, we let N_p and N_s denote the set of patient classes and servers (IWs), respectively. For $i \in N_p$, we denote by λ_i the arrival (i.e., bed request) rate of class i patients. We model the service process in IWs with class-dependent service rates μ_i where $i \in N_p$. We also let $X_i(t)$ denote the number of class i patients boarded in the ED at time t, and define $\underline{X}(t) = (X_i(t) : i \in N_p)$ as the vector of the number of all such patients. Moreover, for $i \in N_p$ and $j \in N_s$, we let $a_{ij}(t) = 1$ if IW j is serving a class i patient at time t, and $a_{ij}(t) = 0$ otherwise. We model the potential occurrence of adverse events that might occur for patients boarded in the ED (awaiting transfer to an inpatient unit) by class-dependent Poisson process, In particular, we let $\bar{\theta}_i$ denote the per unit of time risk of adverse events (i.e., the rate of the underlying Poisson

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

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

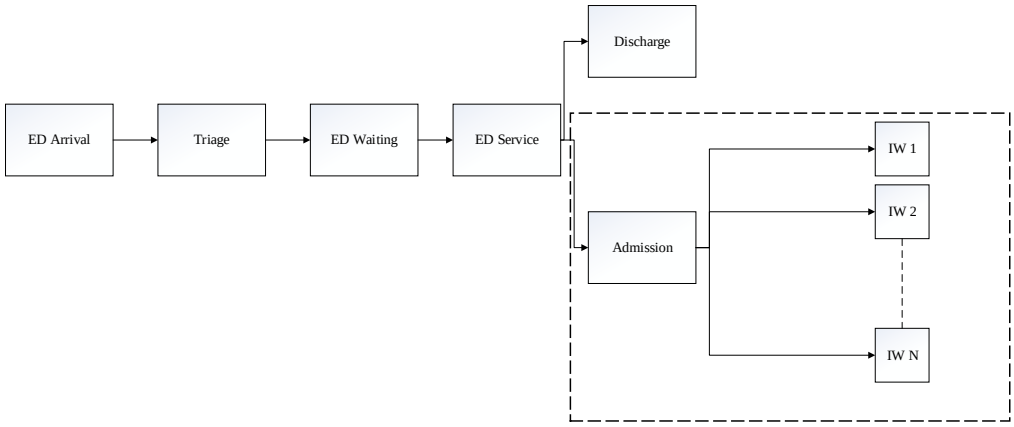


Figure 2 General flow of patients with the dotted area representing the focus of this paper (IW: Inpatient Ward)

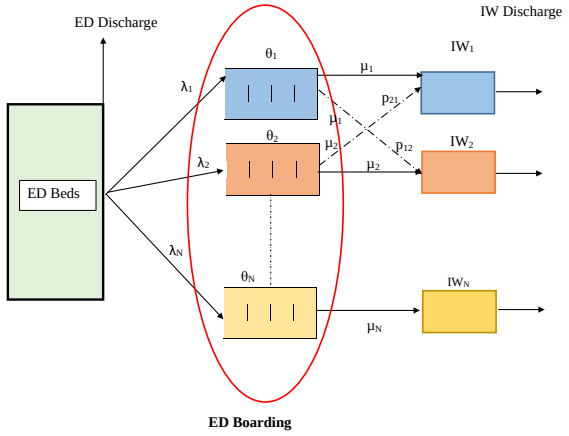


Figure 3 A queueing representation of the patient flow

process) that can occur for a class i patient boarded in the ED, denote by c_i the associated cost per adverse event, define $\theta_i = c_i\bar{\theta}_i$, and let $\underline{\theta} = (\theta_i : i \in N_p)$. In this setting, θ_i plays the role of "expected holding cost" for a patient of class i, and is accrued per unit of time boarding in the ED. However, the actual "holding cost" is random and depends on stochastic deteriorations in the patient's conditions. Similarly, for assignments to secondary inpatient units, we let p_{ij} denote the expected value of a non-negative "penalty cost" (which is random in nature due to its dependency to various patient and provider-dependent conditions) that is accrued due to a lower-than-desired quality of care when a patient of class i is assigned to IW j ($p_{ij} = 0$ if $i = j$).[5]

The objective is to find an optimal assignment policy to control the patient flow in order to minimize the expected total long-run average sum of (a) adverse events (a patient safety concern),

[5]In Section 6, we will discuss how we have used a year of data on patients with chest pain (CP) or congestive heart failure (CHF) to estimate all the parameters required for our model.

and (b) the penalties accrued due to placement in secondary units (a quality of care concern). [6] This optimal objective can be calculated as:

$$Z^* = \inf_{\pi\in\Pi} Z^\pi = \inf_{\pi\in\Pi}\left[\sum_{i\in N_p}\sum_{j\in N_s} p_{ij}O_{ij}^\pi + \sum_{i\in N_p}\theta_i L_i^\pi\right], \tag{1}$$

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

$$L_i^\pi = \limsup_{T\to\infty}\frac{1}{T}\int_0^T E\left[X_i^\pi(s)\right]ds, \tag{2}$$

$$O_{ij}^\pi = \limsup_{T\to\infty}\frac{A_{ij}^\pi(T)}{T}, \tag{3}$$

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

3.1. A Markov Decision Process Formulation

As mentioned earlier, our partner hospital has eight main IWs (see Table 1), and hence, $|N_p| = |N_s| = 8$. However, as noted earlier, because each patient type has only one primary and one secondary IW, the hospital can be viewed as multiple primary-secondary IW pairs. Hence, we expect the insights generated by focusing on a single primary-secondary pair to be useful for the whole hospital system. For this reason, and to gain some clear insights into effective patient flow control policies, we start by considering the simplest case where $N_p = N_s = \{1,2\}$, and later test the insights gained via simulations calibrated with data for a larger system. We let $\underline{a_1} = (a_{11}, a_{21})$ and $\underline{a_2} = (a_{12}, a_{22})$, where $a_{ij} = 1$ if server j is busy with a patient of class i. We assume that all the underlying processes are memoryless, and require that at any point in time $\sum_{i\in N_p} a_{ij} \le 1 \ (\forall j \in N_s)$. With these, we define the system state as $\tilde{X} = (\underline{X}, \underline{a_1}, \underline{a_2})$ with state space $\mathscr{S} = \mathbb{Z}_+^2 \times \{0,1\}^2 \times \{0,1\}^2$. [8] We then use uniformization to transfer the underlying continuous-time Markov chain (CTMC) to a discrete-time Markov chain (DTMC). Let $\psi = \lambda_1 + \lambda_2 + 2\max\{\mu_1,\mu_2\}$ be the uniformization factor. Then, the long-run average cost optimality equation for the DTMC can be

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

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

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

written as:

$$J(\tilde{X})+\hat{Z}^* = \frac{1}{\psi}\Bigg[\underline{\theta}\,\underline{X}^T + \min_{u=\underline{u_{ij}}\in\mathscr{U}(\tilde{X})}\Bigg\{\sum_{i\in N_p}\sum_{j\in N_s}\lambda_i T^{\underline{u_{ij}}}J(\underline{X}+e_i,\underline{a_j}) + \sum_{i\in N_p}\sum_{j\in N_s}\sum_{k\in N_p}a_{kj}\mu_k T^{\underline{u_{ij}}}J(\underline{X},\underline{a_j}-e_k) + \left(\psi-\sum_{i\in N_p}\lambda_i-\sum_{k\in N_p}\sum_{j\in N_s}a_{kj}\mu_k\right)J(\tilde{X})\Bigg\}\Bigg], \tag{4}$$

where $J(\tilde{X})$ is a relative cost function defined as the difference between the total expected cost of starting from state $\tilde{X}$ and a reference state (state $\underline{0}$), $\hat{Z}^*$ is the optimal average cost per uniformized period, the notation "^{T}" represents the transpose operator, and $T^{\underline{u_{ij}}}$ is a functional operator that depends on action vector $\underline{u_{ij}}$. In optimality equation (4), e_i is a vector with the same dimensions as $\underline{X}$ containing a one in the ith position and zeros elsewhere. Thus, the first line inside the minimization in (4) is due to inpatient bed request arrivals from the ED, which occur with rate λ_i for patients of class i. Similarly, the second line in (4) is due to discharges of patients from IWs, and the last line in (4) is due to the self-loop in the underlying DTMC. The control actions u_{ij} in (4) are taken so as to minimize the long-run average cost, where the set of admissible actions is:

$$\mathscr{U}(\tilde{X}) = \left\{u=(u_{ij})_{i\in N_p,j\in N_s}\ \text{s.t.}: u_{ij}\in\{0,1\}, \textstyle\sum_{i\in N_p}u_{ij}\le(1-\sum_{i\in N_p}a_{ij})\ \ \forall j\in N_s, \sum_{j\in N_s}u_{ij}\le X_i\ \ \forall i\in N_p\right\}. \tag{5}$$

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

4. The Optimal Patient-IW Assignment Policy

In Online Appendix A, we show that we can restrict our attention to policies that do not allow idling an IW $j\in N_s$ when there is a patient with IW j as his/her primary IW boarded in ED (See Proposition EC.1 in Online Appendix A).[9] Although this is an expected result in service systems in which preemption is allowed, we note that in non-preemptive services such as the one we model, this insight can be counter intuitive. To establish this non-idling result under our non-preemptive assumption, we first demonstrate a monotonicity property in Online Appendix A (see, Lemma EC.1). Here, we seek to answer the questions we raised in the Introduction, and generate insights into conditions under which patients should be forced to wait in the ED until a bed in their primary inpatient unit becomes available (rather than being transferred to a secondary unit with current bed availability). We start by establishing the following result.

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

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

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

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

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

As is specified in Theorem 1, it is optimal to idle IW 1 when there is no Class 1 patient available and the number of Class 2 patients waiting for an inpatient bed is below a threshold level. This is due the non-zero overflow penalty, which represent the reduction in quality of care when a patient is assigned to a secondary unit. In fact, when p_{ij} is high enough, the optimal policy idles IW j whenever it does not have any primary patient boarded in the ED. For non-extreme overflow penalty cases, when IW 1 does not have a primary patient boarded in the ED, it first idles until the number of Class 2 patients boarded in the ED reaches a certain level, then prioritize Class 2 patients until either a Class 1 patient starts to board in the ED, or the number of Class 2 patients falls below

the threshold. For hospital bed managers, Theorem 1 implies that when a bed becomes available in IW 1, Class 1 patients should be assigned to that IW if there are Class 1 patients boarded in the ED. Otherwise, Class 2 patients should be assigned to IW 1, but only if the number of Class 2 patients boarded in the ED is higher than a certain level. This insight is important, because it sheds light on the fact that an IW 1 bed can be left idle under the optimal policy depending on the congestion level of the ED. By idling such a bed and asking Class 2 patients to continue boarding in the ED, the hospital bed manager can avoid a potential reduction in quality of care, and also prevent a future arriving Class 1 patient from prolonged ED boarding, which in turn may have significant patient safety related consequences.

4.1. Patient Flow to IW 2

To gain further insights into the structure of effective patient-IW assignment policies, we now turn our attention to IW 2, and consider the simplified model illustrated in Figure 4. Recall that IW 2 is the primary IW for Class 2 patients, and the secondary IW for Class 1 patients, where we labeled classes (without loss of generality) such that $\theta_1\mu_1 \geq \theta_2\mu_2$. Thus, IW 2 prefers to serve Class 1 with respect to the $\theta\mu$ index, but Class 2 with respect to the overflow penalty cost parameters. As we will see, understanding the main trade-offs in this simplified model is essential for answering the questions we raised in the Introduction. Put differently, although the model presented in Figure 4 is a stylized version of the complex patient flow in hospitals, it allows us to gain useful insights that we can further test via realistic simulations.

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

$$\mathscr{U}(\underline{Y}) = \left\{ u = (u_{i2})_{i\in\{1,2\}} \ \text{ s.t.} : u_{i2} \in \{0,1\}, u_{i2} \leq Y_i, \sum_{i\in\{1,2\}} u_{i2} \leq 1 \ \ \forall i \in \{1,2\} \right\}. \tag{6}$$

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

$$J(\underline{Y}) = \underline{\theta}\,\underline{Y}^T + \beta \min_{u\in\mathscr{U}(\underline{Y})} \left\{ \sum_{i\in\{1,2\}} \tilde{\lambda}_i J(\underline{Y}+e_i) \right.$$

$$\left. + \sum_{i\in\{1,2\}} \tilde{\mu}_i u_{i2} \left(p_{i2} + J(\underline{Y}-e_i)\right) + \left(1 - \Lambda - \sum_{i\in\{1,2\}} \tilde{\mu}_i u_{i2}\right) J(\underline{Y}) \right\}, \tag{7}$$

where β is the discount factor per uniformized period, the overflow penalty cost parameters p_{12} and p_{22} are scaled so that $p_{22} = 0$, and the vector $\underline{\theta}$ is scaled so that $\underline{\theta}\,\underline{Y}^T$ represents the expected

[10] We realize that allowing service preemption is not fully realistic; however, this assumption is useful for tractability and for gaining sharp insights. We relax this assumption in Section 6, and utilize real-world data along with simulation analyses to verify the insights gained.

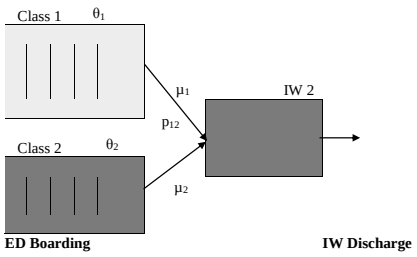


Figure 4 A queueing representation of the simplified system

cost per uniformized period when the system is at state $\underline{Y}$. Moreover, in (7), the uniformization rate is $\bar{\psi} = \lambda_1 + \lambda_2 + \max\{\mu_1, \mu_2\}$, where $\tilde{\mu}_i = \dfrac{\mu_i}{\bar{\psi}}$, $\tilde{\lambda}_i = \dfrac{\lambda_i}{\bar{\psi}}$, and $\Lambda = \tilde{\lambda}_1 + \tilde{\lambda}_2$. Next, we define the functional operators T_a, T_u and T_* (see, e.g, Saghafian and Veatch (2016) for the use of similar operators in a different queueing structure) as:

$$T_\theta J(\underline{Y}) = \underline{\theta}\,\underline{Y}^T, \tag{8}$$

$$T_a J(\underline{Y}) = \sum_{i\in\{1,2\}} \tilde{\lambda}_i J(\underline{Y} + e_i), \tag{9}$$

$$\begin{aligned} T_u J(\underline{Y}) &= \sum_{i\in\{1,2\}} \tilde{\mu}_i u_{i2} \left(p_{i2} + J(\underline{Y} - e_i)\right), \left(1 - \Lambda - \sum_{i\in\{1,2\}} \tilde{\mu}_i u_{i2}\right) J(\underline{Y}), \\ &= (1-\Lambda) J(\underline{Y}) - \sum_{i\in\{1,2\}} \tilde{\mu}_i u_{i2} \left(\Delta_i J(\underline{Y} - e_i) - p_{i2}\right), \end{aligned} \tag{10}$$

$$T_* J(\underline{Y}) = \min_{u\in\mathscr{U}(\underline{Y})} T_u J(\underline{Y}), \tag{11}$$

$$TJ(\underline{Y}) = T_\theta J(\underline{Y}) + \beta \left(T_a J(\underline{Y}) + T_* J(\underline{Y})\right), \tag{12}$$

where $\Delta_i J(\underline{Y}) = J(\underline{Y} + e_i) - J(\underline{Y})$. Using these functional operators, we can simply write the infinite-horizon optimality equation (7) as

$$J(\underline{Y}) = TJ(\underline{Y}). \tag{13}$$

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

Using the above-mentioned setting, we next consider the following two properties

$$(i)\ \ \mu_1\Delta_1 J(\underline{Y}) - \mu_2\Delta_2 J(\underline{Y} + e_1 - e_2) \geq \mu_1\Delta_1 J(\underline{Y} - e_1) - \mu_2\Delta_2 J(\underline{Y} - e_2) \ \text{ for all } \ \underline{Y} \geq (1,1), \tag{14}$$

$$(ii)\ \ \mu_1\Delta_1 J(\underline{Y} - e_1) - \mu_2\Delta_2 J(\underline{Y} - e_2) \geq \mu_1\Delta_1 J(\underline{Y} + e_2 - e_1) - \mu_2\Delta_2 J(\underline{Y}) \ \text{ for all } \ \underline{Y} \geq (1,1). \tag{15}$$

Property *(i)* implies that assigning Class 1 patients to IW 2 becomes more desirable as the number of boarded Class 1 patients increases, and property *(ii)* implies that assigning Class 2 patients to IW 2 becomes more desirable as the number of boarded Class 2 patients increases. Let $\mathscr{F}$ be the set of real-valued functions defined on $\mathscr{S} = \mathbb{Z}_+^2$ such that if $F \in \mathscr{F}$ then F satisfies properties

(14)-(15). The following lemma shows that, if $\theta_1\mu_1 \geq \theta_2\mu_2$, the functional operator T defined in (12) preserves properties (14)-(15).

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

Utilizing Lemma 1, we can establish the following result.

THEOREM 2 **(Optimality of a Threshold-Type Policy)**. *If $\theta_1\mu_1 \geq \theta_2\mu_2$, then the optimal policy obtained from (7) is of a threshold type: IW 2 should prioritize its primary patients until the number of Class 1 patients boarded in the ED reaches a threshold that depends on the number of Class 2 patients still waiting for a bed assignment.*

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

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

effective bed assignment heuristics in the next section. We then use a variety of simulation experiments (calibrated with hospital data that we have collected) to evaluate their effectiveness under realistic conditions, and generate more detailed insights for hospital administrators via sensitivity analyses.

5. Heuristic Policies

When we consider non-preemptive service policies (which better represent the current practice in most hospitals) under the general system structure discussed in Section 3, our numerical computations show that the optimal policy is complex: it has a state-dependent threshold that depends on all the elements in the system state, including IW bed availabilities. Our numerical results also show that the optimal policy has a structure similar to the optimal control of the "N" structure queueing network, where one server works as a *shared* server while the other works as a *dedicated* server. That is, the primary unit of Class 1 patients (IW 1) typically prioritizes its primary patients (i.e., works as a dedicated unit whenever its queue of boarded patients is not empty), and primary unit of Class 2 patients (IW 2) typically first serves its primary patients until the number of Class 1 patients boarded in ED exceeds a threshold, then helps IW 1 by serving Class 1 patients (see Online Appendix A for some numerical experiments supporting this observation). In what follows, we take advantage of this (as well as our earlier findings) to develop easy-to-implement heuristic policies for use in hospitals.

5.1. A Birth-and-Death Process to Approximate the Optimal Threshold

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

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

T, we can calculate $Z(T)$ as:

$$Z(T) = \theta_1 L^1(T) + \theta_2 L^2(T) + p_{12} O^1(T). \tag{16}$$

The objective is to find the value of T that minimizes $Z(T)$. To calculate (16), we use the above-mentioned birth-and-death processes to estimate $L^1(T)$, $L^2(T)$, and $O_1(T)$. To this end, we first need to obtain the steady-state probability P_i^j which is the probability that the length of queue $j \in \{1,2\}$ equals to $i \geq 0$. From the balance equations, we have:

$$P_i^1 = (\frac{\lambda_1}{\mu_1})^i P_0^1 \qquad \forall\ i \leq T, \tag{17}$$

$$P_i^1 = (\frac{\lambda_1}{\mu_1})^T (\frac{\lambda_1}{2\mu_1})^{i-T} P_0^1, \quad \forall\ i > T. \tag{18}$$

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

$$P_0^1(T) = \frac{(1-\rho_1)(1-\rho_2)}{\rho_1^T(\rho_2-\rho_1)+(1-\rho_2)}, \tag{19}$$

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

$$L^1(T) = \sum_{i=0}^{T} i\rho_1^i P_0^1(T) + \sum_{i=T+1}^{\infty} i\rho_1^T \rho_2^{i-T} P_0^1(T). \tag{20}$$

Also, $O^1(T) = \frac{1}{2}\sum_{i=T+1}^{\infty} i\rho_1^T \rho_2^{i-T} P_0^1(T)$ by assuming that Class 1 patients in the queue will be served equally by IW 1 and IW 2 after the number of Class 1 patients boarded in ED reaches the threshold.[11] To calculate the average queue length of Class 2 patients, $L_2(T)$, we first calculate the following:

$$P^2(T) = P(x_1 \leq T) = P_0^1(T)\frac{1-\rho_1^{T+1}}{1-\rho_1}. \tag{21}$$

The average queue length for Class 2 patients is then:

$$L^2(T) = \frac{\lambda_2}{P^2(T)\mu_2 - \lambda_2}. \tag{22}$$

These allow us to calculate $Z(T)$ via (16), and find the optimal threshold value $T^* = \arg\min_{T\geq 0} Z(T)$. However, the threshold level T^* does not have a closed-form solution, and the function $Z(T)$ can be non-convex in general. Nevertheless, we can utilize numerical approaches (e.g., bisection search) to find the value that minimizes (16). We term the heuristic policy that controls the patient flow based on this threshold as the *birth-and-death threshold (BDT)* policy.

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

Our results in Section 4 reveal that there exists a threshold type optimal policy that optimizes performance by following the primary-then-$c\mu$ rule (see, e.g., Theorem 2). This policy tends to serve the primary patient type with the lower $c\mu$ value until the cost differences of serving the secondary patients exceeds the overflow penalty cost (see the discussion in Online Appendix A, proof of

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

Lemma 1). This insight suggests that instead of using a heuristic policy to directly approximate the threshold—the idea behind the BDT policy—there might be value in following a heuristic that balances the costs associated with different queues. Thus, as our second heuristics, we develop a modified version of the Largest Expected Workload Cost (LEWC) policy proposed by Saghafian et al. (2011) for general parallel queueing systems. The LEWC policy dynamically balances the expected workload cost of queues by prioritizing the queue with the largest expected workload cost (ROAE in our setting).[12] In order to also incorporate the additional penalty cost of serving patients in their secondary IW—a main factor for the patient flow focus of this study— we propose a penalty-adjusted version of LEWC, which we term *LEWC-p*. To this end, similar to Saghafian et al. (2011), we first use the following Linear Program (LP). In this LP, the objective is to find the optimal server allocations to maximize the minimum percentage excess capacity among all patient types:

$$\text{Max} \quad \tau \tag{23}$$

Subject to:

$$\sum_{j\in N_s} y_{ij}\mu_i \geq \lambda_i(1+\tau) \qquad \forall i \in N_p, \tag{24}$$

$$\sum_{i\in N_p} y_{ij} \leq 1 \qquad \forall j \in N_s, \tag{25}$$

$$y_{ij} \geq 0 \qquad \forall i \in N_p, \forall j \in N_s. \tag{26}$$

In this LP, y_{ij} is the decision variable that represents the long-run proportion of time that IW j serves patient class i. Constraint (24) ensures that the objective function maximize the minimum excess capacity among all patient classes. Constraint (25) guarantees that the total proportion of time for each IW does not exceed 1, and Constraint (26) enforces the proportions to be non-negative.

Next, when a bed in IW j becomes available, we calculate an index, $I_{ij}(x_i)$, for each queue $i \in N_p$ (class of patients boarded in the ED) to approximate the penalty-adjusted expected workload cost of that queue given that its current length is x_i:

$$I_{ij}(x_i) = \frac{\theta_i x_i}{\sum_{j\in N_s} y^*_{ij}\mu_i} - p_{ij}\frac{x_i y^*_{ij}}{\sum_{j\in N_s} y^*_{ij}}, \tag{27}$$

where y^*_{ij}'s are the solution to LP (23)-(26). The first part of the index approximates the cost associated with risk of adverse events for class i patients: since there are x_i patients in the queue, it will take approximately $\frac{x_i}{\sum_{j\in N_s} y^*_{ij}\mu_i}$ units of time to serve them, and the cost due to adverse events is

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

θ_i per unit of time per patient boarded. The second part of the index approximates the associated penalty cost. In this term, $\frac{y^*_{ij}}{\sum_{j\in N_s} y^*_{ij}}$ represents the proportion of patients of class i served by IW j.[13]

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

Step 1. Solve LP (23)-(26) to derive optimal allocations y^*_{ij}.

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

5.3. Comparison of the Proposed Heuristic Policies

We now compare the performance of the proposed BDT and LEWC-p heuristic policies with the optimal policy. As a benchmark, we also use the generalized $c\mu$ (G$c\mu$) rule. Under the G$c\mu$ policy, the available bed in IW j is assigned to the class that has the highest $\theta_i\mu_i x_i$ value. We use this policy as a benchmark since it (a) takes the queue lengths into account, and (b) is known to work well in a variety of queueing systems.[14]

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

Figure 5 illustrates our computational results over the test suite by constructing the empirical Cumulative Distribution Function (CDF) for the percentage optimality gap of each of the non-optimal policies (BDT, LEWC-p, and G$c\mu$). The results presented in this figure show that LEWC-p and BDT policies can both be considered as "nearly-optimal" policies. However, the mean and standard deviation of LEWC-p optimality gap is smaller than that of the BDT policy, so we can conclude that it is the better policy. The performance of $Gc\mu$ is, however, significantly worse than both the LEWC-p and BDT policies. This is mainly because $Gc\mu$ does not consider penalties associated with secondary unit assignments. However, even when the underlying penalty parameter is zero, we observe that $Gc\mu$ is not the best policy for all cases. When the penalty parameter is

[13] In using (27), we assume that LP (23)-(26) has a unique optimal solution with $y^*_{ij} \neq 0$ whenever $i \neq j$. For systems in which this solution is not unique (e.g., balanced systems where $\frac{\lambda_i}{\mu_i} = \kappa, \quad \forall i \in N_p$) ties need to be broken based on cost parameters.

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

zero, both of the proposed heuristic policies (BDT and LEWC-p) perform close to each other while BDT performs slightly better due to the assumption that IW 1 only serves Class 1 patients (under the $c\mu$ policy both of the IWs serve Class 1 whenever feasible).

Table 2 compares the optimality gap of LEWC-p, BDT, and $Gc\mu$ policies for various congestion levels in the system. All of the policies show a smaller mean optimality gap in moderate to high congestion levels than in the low congestion level. This observation suggests that implementing them in crowded systems (e.g., in busy teaching hospitals) is better than doing so in less crowded systems (e.g., in less busy urban hospitals). Finally, Table 3 compares the policies based on various penalty parameter settings and shows that all policies perform best when the underlying penalty parameter is high. Moreover, LEWC-p is more robust than the BDT policy to changes in the penalty parameter. This is intuitive, since the BDT policy only uses one threshold level, while the LEWC-p policy dynamically adjusts the assignments based on the number of patients of different classes that are boarded in the ED.

6. Simulation Analysis Using Hospital Data

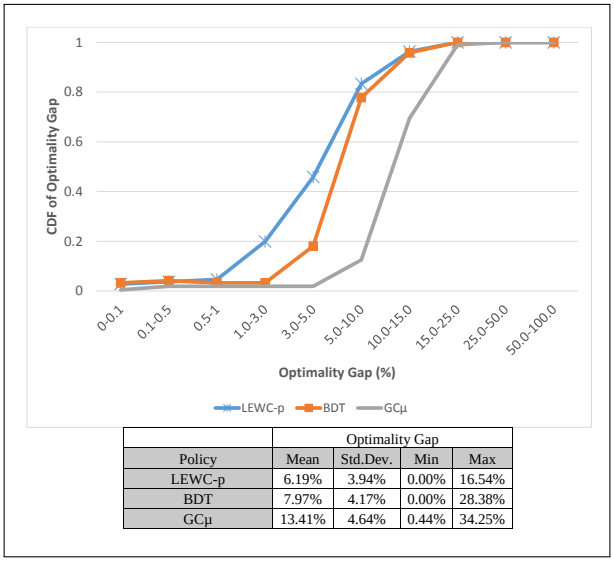
To gain more insights into effective policies for assigning ED patients to their primary or secondary inpatient units, we use a discrete-event simulation model of ED patient flow, and calibrate it with a year of hospital data that we have from our partner hospital. This enables us to relax some of the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

6.1. Patient Flow and IWs in Our Partner Hospital

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

IWs. Patients from the three admission sources described above require a bed from one of the eight inpatient units based on their diagnosis. The name of IWs, their descriptions, and number of beds in each of them in our partner hospital can be found in Table 1 (see Section 1).

Patient Types. To gain clear insights into effective assignment policies, we focus on patients who were admitted via the ED of our partner hospital with an admission diagnosis of either *chest pain*



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

Figure 5 Performance of LEWC-p, BDT, and $Gc\mu$ relative to the optimal policy over the entire test suite (216 problem instances)

Table 2 Optimality gap of policies for various congestion levels

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

Table 3 Optimality gap of policies for various penalty cost parameters

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

(CP) or *congestive heart failure* (CHF). These patients are often assigned to a secondary IW; the primary IW for both CP and CHF patients is 4 West (4W), and their secondary IW is 5 West (5W) (see Table 1 for more information regarding these IWs). There are two types of CP and CHF patients: Type 1 patients are those considered to be more sensitive to a secondary bed assignment

(i.e., are subject to higher reduction in quality of care if assigned to a secondary inpatient unit). Type 2 patients are those who are less sensitive to a secondary bed assignment. We develop a classification scheme using simple laboratory findings and based on our discussions with medical experts at our partner hospital. We define Type 1 CP patients as those who have an elevated serum *troponin* (Tn) level, and Type 2 CP patients as those who have a normal troponin level. We define Type 1 CHF patients as those who have a *B-type natriuretic peptide* (BNP) level of 4,000 pg/ml or greater, and Type 2 CHF patients as those with BNP levels below 4,000 pg/ml. Our empirical analyses show that, among patients of same type, there is no statistically significant difference in the mean IW service time between primary and secondary units (see Table EC.7 in Online Appendix C).

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

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

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

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

our physician collaborators, and are intended to represent values that are realistic while covering possible differences among hospitals.

Performance Measures. In addition to the overall objective we introduced in Section 3, we use the overflow proportion (the ratio of patients assigned to a secondary IW to the total number of patients of same type served) and the average ED boarding time (the average time between a request and bed occupancy) as other performance measures. We also use the 2-hour boarding rate (the fraction of patients that are boarded for two hours or more)[16] as another performance measure. We do so because reducing excessive boarding times (and not just average boarding times) is also important for most EDs.

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

Base Case Scenario. We consider the base case scenario to be a reflection of the current system in our partner hospital based on a year of data that we have collected. We use this scenario as a benchmark to analyze the potential changes that may occur due to implementing our proposed policies. Thus, we use the level of performance measures in the base case scenario (e.g., 2-hour boarding rate, average boarding time in the ED, etc.) for CP and CHF patients as a point of reference, and compare the results of our proposed policy with those metrics. To this end, we focus on patient flow from ED to the two IWs that can serve CP and CHF patients: 4 West and 5 West. In addition to CP and CHF patients, we simulate the flow of other patients that require a bed from 4 West or 5 West, but note that these patients are not eligible for overflows, and can only be assigned to their primary inpatient units. We include these patients in our simulation model to represent the capacity utilization in 4 West and 5 West more accurately, thereby increasing the fidelity of our simulations. Figure 6 illustrates the patient flow under consideration.[17]The dashed lines in Figure 6 show assignments of patients to secondary IWs (overflows that incur a penalty cost) while the solid lines show assignments to primary IWs. In the current practice, there is no specific rule for assigning patients to their primary vs secondary units. Thus, for our base case scenario, we use the FIFO rule for the primary bed assignments, and model the overflows to secondary IWs by using the proportions that are obtained from our data analyses.

[16] As we discussed in Section 1, the current 2-hour boarding rate at our partner hospital based on our data set is around 30%.

[17] In Online Appendix E, we extend our simulation analysis to the whole patient flow depicted in Figure 3 with all the 8 IWs listed in Table 1. However, since this requires estimating various parameters for each and every patient type served in the hospital, our simulations lose fidelity. Thus, here we stay with CP and CHF patients (i.e., patients for which we have more accurate data).

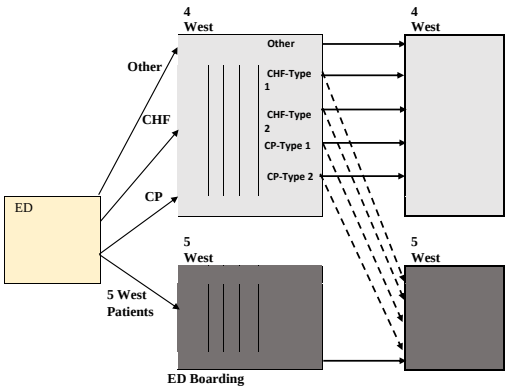


Figure 6 Patient flow in the simulation model

6.2. Validating the Simulation Model

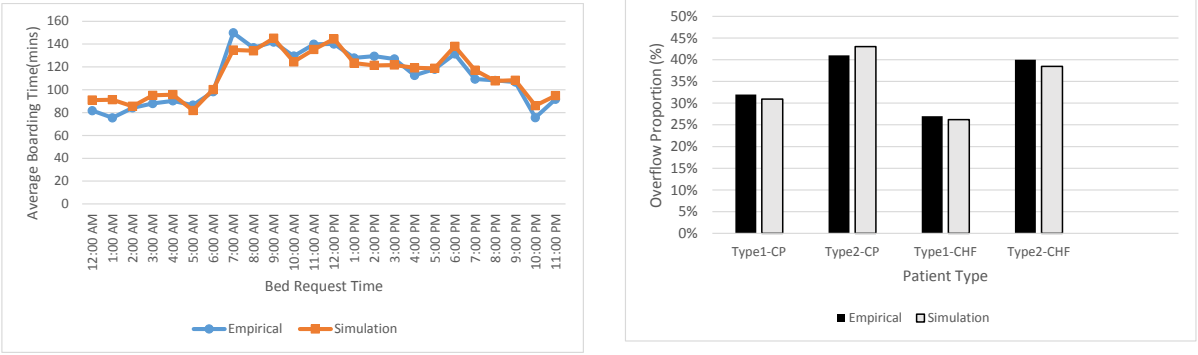
To validate our simulation model, we compare our empirical results obtained directly from our data set with those obtained from our simulation model. Figures 7(a) and 7(b) compare the resulting time-dependent boarding time of patients as well as the resulted overflow rates of the simulation model with that of the empirical data. Using the t-test for the equality of means, we observe no statistical difference between outputs of our simulation model and those from empirical data (p-value = 0.412). Similarly, using Kolmogorov-Smirnov tests for comparing the distributions of outputs (e.g., boarding time distributions) with the empirical distributions from our data, we do not observe any significant mismatch. These results give us confidence that our simulation model is relatively of high fidelity, and accurately matches the current practice.

6.3. Performance of the Proposed LEWC-p Policy

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

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

We next test the sensitivity of the gained benefits to the penalty costs and costs associated with adverse events. As we increase the latter, the improvement in the 2-hour boarding rate and the average ED boarding time increases (see Figures 8(a) and 8(b)). Furthermore, we observe that as we increase the penalty cost, IWs start to work as dedicated units tending to only serve their primary patients. Hence, after increasing the penalty cost, we observe improvements in overflow propor-



(a) Hourly average ED boarding time

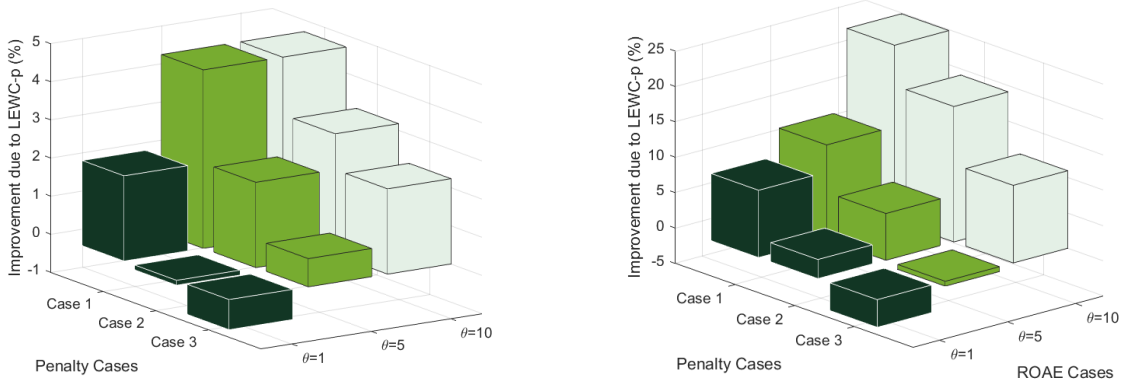
(b) Overflow proportion

Figure 7 Validating the simulation model

tions, but the average ED boarding time and the costs associated with adverse events increase. This result is similar to what we observed from the optimal policy of the analytical model: as we increase (decrease) the penalty cost, the LEWC-p policy mimics the optimal policy by decreasing (increasing) the assignments to secondary IWs. Similarly, as we increase (decrease) costs associated with adverse events that may occur during ED boarding, the LEWC-p policy mimics the optimal policy by increasing (decreasing) the assignments to secondary IWs.

Figures 9(a) and 9(b) illustrate the change in total number of boarded patients in the ED and overflow proportion as ROAE and penalty cost parameters change under the LEWC-p policy. As the ROAE cost increases, the proposed policy starts to assign patients to their secondary IW more aggressively. This leads to lower average ED boarding times, and suggests that utilizing a secondary IW is a more attractive option for patients who have a higher ROAE (e.g., those in need of timely care following their ED service). Another implication of Figures 9(a) and 9(b) is that assigning patients to their secondary IWs has a minimal effect on the average ED boarding time when the penalty cost parameter is high. This suggests that hospital administrators should be more patient in assigning beds for patients who are more sensitive to a secondary IW assignment (e.g. Type 1 patients as opposed to Type 2 patients): *the virtue of patience is dependent on patient type*.

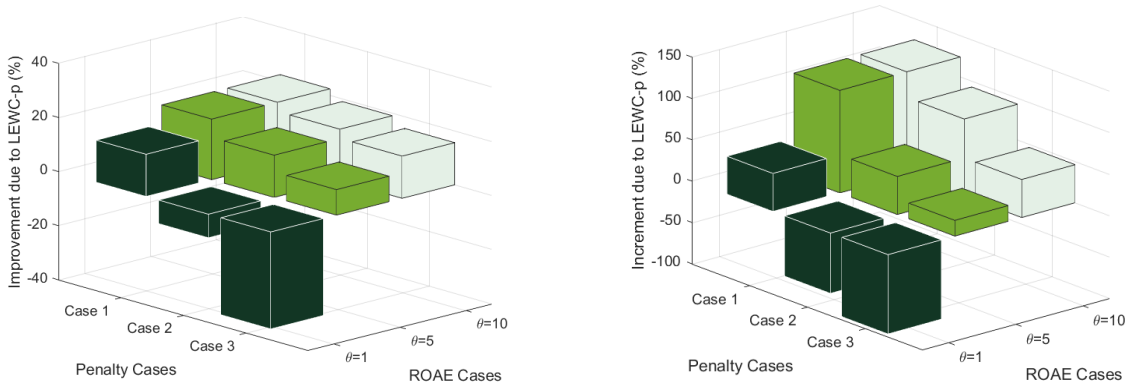
6.3.1. Effect of Idling Beds Our proposed policy allows idling IW beds (in anticipation of future needs) even when there are patients boarded in the ED who need them. However, hospital beds are valuable assets, and keeping them idle while patients are waiting for them might not be perceived as attractive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



(a) Improvement in 2-hour boarding rate

(b) Improvement in average boarding time

Figure 8 Improvement due to LEWC-p compared to current practice for various penalty and ROAE parameters



(a) Average number of patients boarded in ED

(b) Overflow proportion

Figure 9 The effect of ROAE and penalty cost parameters on the average number of patients boarded and overflow proportion due to LEWC-p

OBSERVATION 2 **(Nonidling Policy)**. *Non-idling flow policies increase the number of patients overflowed, but does not significantly change the average number of patients boarded in the ED, the average boarding time, and the 2-hour boarding rate.*

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

administrators that beds should not be idled intentionally. We note that this perception might be correct when the ROAE among different patient groups (in our case, 4 West and 5 West patients) is significantly different.[18] However, our results suggest that hospitals should typically refrain from prohibiting idling: *idling IW beds can be beneficial*.

6.3.2. Effect of Inpatient Bed Capacity In our previous simulation experiments, we used the current bed capacity of IWs in our partner hospital (see, Table 1). To gain more insights for other hospitals which might have higher or lower capacities, we now provide sensitivity analysis by altering the number of beds in IWs (both 4 West and 5 West). Figure 10 illustrates the results, and enables us to make the following obsevation:

OBSERVATION 3 **(Effect of Inpatient Bed Capacity)**. *The achievable improvements due to implementing LEWC-p (on the average boarding time, the 2-hour boarding rate, and the total cost) is greater in hospitals with lower inpatient bed capacity (all else equal).*

This observation suggests that hospitals that lack enough inpatient bed capacity (e.g., busy teaching hospitals) will benefit more from implementing the LEWC-p policy. Thus, instead of investing in increasing their capacity—a challenging and extraordinarily expensive undertaking that often requires a certification of need—they can benefit from better bed assignment policies such as LEWC-p, which requires only a minimal investment.

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

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

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

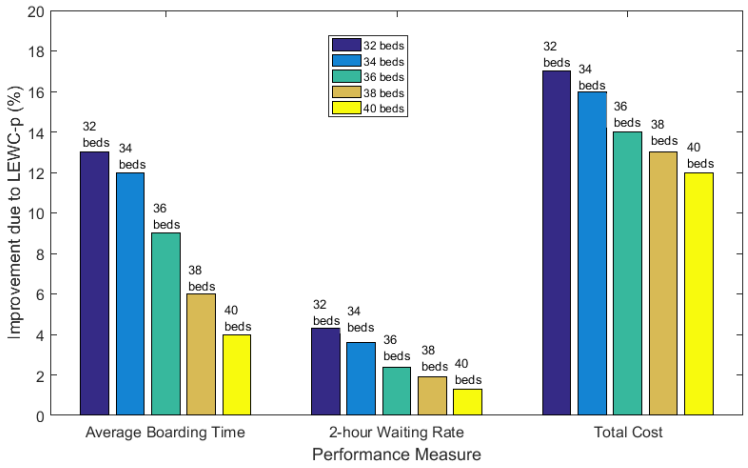


Figure 10 Effect of inpatient bed capacity on the improvements due to LEWC-p

Our results suggest that, when the number of beds usable for overflows decreases, the number of overflows decreases (as expected). Since our proposed policy captures the trade-off between the ROAE and the quality of care, reduction in overflows leads to an increase in the number of patients boarded in the ED. However, the changes in performance measures are not significant when we either drop to 25% bed capacity from 50% bed capacity, or drop to 50% bed capacity from 75% bed capacity. For hospitals with similar characteristics to our partner hospital (in terms of bed request arrivals, inpatient LOS, etc.) this suggests that, to make a statistically significant impact on the performance measures, hospital administrators should consider dramatic changes in the number of beds to be used for overflows.

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

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

This observation suggests that imposing an overflow trigger time that is higher than two hours does not change the performance of our proposed LEWC-p policy. In fact, using a trigger time that is more than two hours typically adds complexity in assignment decisions without any significant change in performance measures. As we noted earlier, our proposed LEWC-p policy improves the average boarding time by approximately 10 minutes per patient compared to the current practice. This results in approximately 100 minutes of average boarding time and 29% of 2-hour boarding rate in the improved system. Setting a trigger time that is lower than two hours can affect more than 70% of the patient population (since 2-hour boarding rate is 29%), which may result in improvements in the average boarding time and the 2-hour boarding rate. However, we find that adding trigger times to LEWC-p does not lead to improvements in the aggregate cost measure, regardless of the level of the trigger time. This further confirms that the proposed LEWC-p policy already strikes a strong balance between concerns related to prolonged ED boarding times and those related to overflows.

7. Conclusion

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

Our results for a simplified model with two patient classes and two IWs suggest that the optimal policy is a threshold-type policy, where the threshold depends on the number of patients boarded in the ED. Under this policy, the primary unit of Class 1 patients (i.e., patients that have a higher $\theta\mu$ value) typically works as a dedicated unit that serves its primary patients whenever such a patient is boarded in the ED. Moreover, the primary unit of Class 2 patients serves them before helping IW 1 on Class 1 patients, and switches to serving Class 1 patients once the number of Class 1 patients boarded in the ED reaches a threshold. These suggest that patience in transferring ED admitted patients to IWs is a virtue, but only up to a point. Contrary to the prevalent perception among hospital administrators, we also find that idling IW beds can be beneficial. In particular, while idling is used in some hospitals and some specific inpatient units, our results indicate evidence for wider implementation of idling policies. We also show that, when the penalties that represent the reduction in quality of care in secondary units are negligible, the optimal policy is a strict priority rule in which both IWs prioritize serving Class 1 patients in order to myopically decrease the risk of adverse of events for patients boarded in the ED.

Our analyses show that the optimal policy is complex in general, and may not be suitable for implementation in practice. Therefore, we use the insights gained from analyzing our simplified models to develop two heuristic policies that are easy to implement. We first use a birth-and-death process to approximate the threshold level that minimizes an aggregate measure of both patient

safety and quality of care. Then, we propose a modified version of the LEWC heuristic termed LEWC-p that enables us to dynamically strike a balance between concerns of patient safety and quality of care. The results show that LEWC-p significantly outperforms other policies, and is also more robust than them in that it has a lower standard deviation of optimality gap. Thus, an important contribution of this study is to introduce LEWC-p as a simple but effective policy that can be implemented in hospitals.

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

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

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

References

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

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

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

Armony M, Israelit S, Mandelbaum A, Marmor YN, Tseytlin Y, Yom-Tov GB (2015). On patient flow in hospitals: A data-based queueing-science perspective. *Stochastic Systems* Forthcoming.

Bell SL, Williams RJ (2001). Dynamic scheduling of a system with two parallel servers in heavy traffic with resource pooling: asymptotic optimality of a threshold policy *The Annals of Applied Probability* 11(3):608–649.

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

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

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

Buyukkoc C, Varaiya P, Walrand J (1985). The cμ rule revisited. *Advances in Applied Probability* 17(1):237–238.

Carr BG, Hollander JE, Baxt WG, Datner EM, Pines JM (2010). Trends in boarding of admitted patients in US emergency departments 2003–2005. *The Journal of Emergency Medicine* 39(4):506–511.

Chan CW, Farias VF, Escobar GJ (2016). The impact of delays on service times in the intensive care unit. *Management Science* 53(2):197–218.

CNN (2008). *Tape shows woman dying on waiting room floor* http://www.cnn.com/2008/US/07/01/waiting.room.death/index.html?eref=rs.

Cox DR, Smith WL (1961). *Queues* (Methuen & Co., London).

Dai JG, Lin W (2005). Maximum pressure policies in stochastic processing networks. *Operations Research* 53(2):197–218.

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

Falvo T, Grove L, Stachura R, Vega D, Stike R, Schlenker M, Zirkin W (2007). The opportunity loss of boarding admitted patients in the emergency department. *Academic Emergency Medicine* 14(4):332–337.

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

Government Accountability Office (GAO) (2003). Hospital emergency departments: crowding vary among hospitals and communities. http://www.gao.gov/new.items/d03460.pdf.

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

Garnett O, Mandelbaum A (2000.) An introduction to skills-based routing and its operational complexities. *Teaching Note, Technion, Israel*.

Griffin J (2012). Improving health care delivery through multi-objective resource allocation. Ph.D. dissertation, School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, GA.

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

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

Guttmann A, Schull MJ, Vermeulen MJ, Stukel TA (2011). Association between waiting times and short term mortality and hospital admission after departure from emergency department: population based cohort study from Ontario, Canada. *British Medical Journal* 342:d2983.

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

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

Kakalik JS, Little JDC (1971). Optimal service policy for the $M/G/1$ queue with multiple classes of arrival. Report, RAND Corporation, Santa Monica, CA.

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

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

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

Mandelbaum A, Stolyar AL (2012). Scheduling flexible servers with convex delay costs: Heavy-traffic optimality of the generalized cμ-rule. *Operations Research* 52(6):836–855.

Meyn SP (2001). Sequencing and routing in multiclass queueing networks part I: Feedback regulation. *SIAM Journal on Control and Optimization* 40(3):741–776.

Meyn SP (2003). Sequencing and routing in multiclass queueing networks part II: Workload relaxations. *SIAM Journal on Control and Optimization* 42(1):178–217.

Palmer J, Mitrani I (2004). *Optimal Server Allocation in Reconfigurable Clusters with Multiple Job Types* (Springer Berlin Heidelberg).

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

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

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

Saghafian S, Hopp WJ, Van Oyen MP, Desmond JS, Kronick SL (2012). Patient streaming as a mechanism to improve responsiveness in Emergency Departments. *Operations Research* 60(5):1080–1097.

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

Saghafian S, Veatch MH (2016). A cμ rule for two-tiered parallel preferences. *IEEE Transactions on Automatic Control* 61(4):1046–1050.

Sennott L (1999) *Stochastic Dynamic Programming and Control of Queueing Systems* (John Wiley & Sons, New York).

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

Teow KL, El-Darzi E, Foo C, Jin X, Sim, J (2012). Intelligent analysis of acute bed overflow in a tertiary hospital in Singapore. *Journal of Medical Systems* 36(3):1873–1882.

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

Thompson S, Nunez M, Garfinkel R, Dean MD (2009). OR practice-efficient short-term allocation and reallocation of patients to floors of a hospital during demand surges. *Operations Research* 57(2):261–273.

Van Mieghem JA (1995). Dynamic scheduling with convex delay costs: The generalized cμ rule. *The Annals of Applied Probability* 5(3):809–833.

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

Walrand J (1988). *An Introduction to Queueing Networks* (Prentice Hall,, Englewood Cliffs, NJ).

Welch PD (1983). The statistical analysis of simulation results. *The computer performance modeling handbook* 22:268–328.

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