Critical care in hospitals: When to introduce a Step Down Unit?

Mor Armony
Stern School of Business, New York University marmony@stern.nyu.edu

Carri W. Chan
Decision, Risk, and Operations, Columbia Business School cwchan@columbia.edu

Bo Zhu
Courant Institute of Mathematical Sciences, New York University zhubo@cims.nyu.edu

In hospitals, Step Down Units (SDUs) provide an intermediate level of care between the Intensive Care Units (ICUs) and the general medical-surgical wards. Because SDUs are less richly staffed than ICUs, they are less costly to operate; however, they also are unable to provide the level of care required by the sickest patients. There is an ongoing debate in the medical community as to whether and how SDUs should be used. On one hand, an SDU alleviates ICU congestion by providing a safe environment for post-ICU patients before they are stable enough to be transferred to the general wards. On the other hand, an SDU can take capacity away from the already over-congested ICU. In this work, we propose a queueing model of patient flow through the ICU and SDU in order to determine when an SDU is needed and what size it should be. Using first and second order analysis, we examine the tradeoff between reserving capacity in the ICU for the most critical patients versus gaining additional capacity achieved by allocating nurses to the SDU due to the lower staffing requirement. Despite the complex patient flow dynamics, we leverage a dimensionality reduction result in our analysis to establish the optimal allocation of nurses to units. We find that under some circumstances the optimal size of the SDU is zero, while in other cases, having a sizable SDU may be beneficial. The insights from our work provide rigorous justification for the variation in SDU use seen in practice.

Key words: Healthcare, critical care, patient flow, queueing, fluid analysis, diffusion analysis, state-space collapse

1. Introduction

Step Down Units (SDUs) provide an intermediate level of care between the Intensive Care Units (ICUs) and the general medical-surgical wards. These units, which are also commonly referred to as intermediate care units and transitional care units, are found in many, but not all, hospitals in developed nations. Typically, these units are staffed at a higher nurse to patient ratio than general medical-surgical wards but not as high as ICUs. ICUs care for the sickest patients and consume a disproportionate share of total health care costs (nearly $82 billion annually (Halpern and Pastores 2010), which amounts to 20-35% of total hospital costs with ICU beds occupying only 5-10 percent of inpatient beds (Joint Commission Resources 2004)). Consequently, a voluminous literature in both the medical and operations communities exists that addresses the need to understand and improve how these units function (see, for example, Chalfin et al. (2007), Chan et al. (2012), Kc and Terwiesch (2012), Kim et al. (2015), Shmueli et al. (2003)). In contrast, very few studies

address these issues with respect to SDUs, despite the fact that, in hospitals that have them, the SDU plays an important role in patient flow through the ICU.

The purpose of an SDU is to treat patients who are more severe than the typical ward patient, but who do not require as intense monitoring as the most critical ICU patients. The basic premise of having an SDU is that it can both care for sicker patients and, at the same time, take pressure off the ICU, thereby resulting in both better patient outcomes as well as increased efficiency (Byrick et al. 1986, Zimmerman et al. 1995). Despite this promise, there is high variation in the presence and size of SDUs as the medical community debates the use of these units. Our goal in this work is to develop a better understanding of the operational role SDUs play in the treatment of critically ill patients.

Semi-critical patients who can be treated in the SDU can generally be treated in the ICU without any impact on their quality of care. Conversely, due to the lower staffing requirements in the SDU, Critical patients who are treated in the SDU will not be able to receive the high level monitoring and care provided in the ICU, resulting in substantial degradation of their quality of care. Hence, not only do ICUs provide care for the sickest patients, they can also be considered 'flexible servers' in the sense that they can also treat moderately severe patients. However, largely due to the high nurse-to-patient ratio requirement, they are more costly to operate than SDUs. In California, an ICU is legally obligated to have at least one nurse for every two ICU patients; in practice, many hospitals operate with one nurse per patient. In contrast, SDUs can be staffed anywhere from one nurse per two to four patients. In particular, the SDU can accommodate more patients for the same number of nurses. This creates an interesting tradeoff between overall capacity gains (SDU) for all critical patient severities versus maintaining more capacity for the most severely ill patients (ICU).

This work was initially motivated by a conversation with the chief intensivist at a large urban hospital. The hospital was considering creating an SDU by reducing capacity in the ICU. The main debate centered on how many SDU beds should be created without modifying the number of nursing staff on budget. The hospital did not want to increase the number of nursing staff on budget due to cost considerations–any physical changes would primarily have a one time occurrence (at the time of change), but staffing costs would perpetuate long into the future. On the other hand, cutting nursing staff would hurt hospital morale and result in substantial backlash by hospital staff which would make it difficult to implement the new plan. The goal was to rotate the current ICU nurses between the ICU and new SDU, so that the main differentiation between the two units would be the nurse-to-patient ratio. The decision to use critical-care nurses in the SDU was clinically strategic–management wanted to ensure that the nurses were capable of dealing with any complications which could arise in the unit. Other hospitals have also used critical-care nurses to staff the SDU (e.g. Eachempati et al. (2004), Harding (2009)). While some hospitals (e.g. Aloe et al. (2009)) use medical-surgical nurses in their SDU, our primary focus will be on the hospitals which use critical care nurses in both the ICU and SDU.

Patient flows into SDUs can come from various sources. For instance, patients can be directly admitted to an SDU from the Emergency Department if they are deemed too sick for the ward, but not so sick that they require ICU care. Alternatively, some SDUs are used for post-operative patients with fairly standard recovery patterns, but who need additional monitoring in the event of complications due to surgery. While the original intent of the SDU was to provide 'Step-Down' care for patients post-ICU, patients are sometimes placed in the SDU prior to ICU care if the ICU is too congested to immediately admit the patient. These complex flow patterns make studying SDUs quite challenging. A number of hospitals (e.g. Cady et al. (1995) and Eachempati et al. (2004)) only admit post-ICU patients into their SDU, while others allow different admission patterns as described above. In order to maintain tractability and gain some insight into the role of SDUs in the care of Critical patients, we focus on the case where the SDU is a true 'Step Down Unit' and patients are admitted only after being discharged from the ICU.

We introduce a queueing model of Critical patients who arrive to the ICU. If there is an available bed, a patient will be treated immediately. If there is a long queue of critical patients waiting for an ICU bed, the patient will immediately balk and be sent for care at another hospital. Otherwise, he will be treated in another hospital bed while waiting to be admitted to the ICU. If the wait is too long, the patient will eventually recover and no longer need ICU care or, in the most extreme case, die due to the long wait–we refer to such events as patient 'abandonment'. A Critical patient who is admitted to the ICU will be treated until reaching either a stable enough state to leave the ICU/SDU system or a Semi-critical state where he can be treated in the SDU or stay in the ICU. To capture the fact that demand pressures from sicker patients can lead to patient discharges from the ICU (Kc and Terwiesch 2012), we allow for Semi-critical patients to be bumped out of the ICU if a Critical patient requires a bed.

Our objective is to determine the size of the SDU and ICU and the balking threshold in order to minimize the costs associated with patient balking, abandonment, holding in queue, and bumping. Cost minimization and reward maximization formulations are common in the healthcare literature (see for example, Green et al. (2006a), Chan et al. (2012), Mills et al. (2013), Mason et al. (2014), Best et al. (2015), Mills et al. (2015) among others).

Our main contributions can be summarized as follows:

- We start with first order analysis of our queueing system via a fluid approximation and provide justification for the highly varied use of SDUs observed in practice. In particular, we find there exist two operational regimes which depend on the relative costs between lack of access for Critical and Semi-critical patients. In one–the ICU Driven (ID) regime–virtually all nurses are allocated to the ICU (so the SDU is very small or is of size zero), and the system only incurs costs related to the bumping of Semi-Critical patients. While in the other–the Capacity Driven (CD) regime–a significant number of nurses are allocated to both units, and only costs related to Critical patients (balking, abandonment and holding) are incurred. Surprisingly, this can occur even when these per-patient costs for critical patients are greater than the per-patient cost of

bumping semi-critical patients. Moreover, our results are very robust to variation in system parameters as long as the system is away from the switching point between the two regimes.

- Using second order analysis (via a diffusion approximation), we develop better insight into how a more refined characterization of system dynamics plays into the optimal policy. In contrast to the first order analysis, costs for lack of access to care (via balking, abandonment, queueing, and bumping) for *both* Critical and Semi-critical become significant when considering second order terms. Additionally, fine-tuned optimization of the balking threshold becomes important.

Our second order analysis suggests that a consistently full ICU does not necessarily imply that the ICU is the system bottleneck. In some cases, it is the shortage of SDU beds that results in having many ICU beds being occupied by semi-critical patients.

- Via numeric and simulation analysis, we find that the solutions obtained from our fluid and diffusion approximations result in good outcomes compared to an exhaustive search. This holds even under moderate traffic. Moreover, we find that in the Capacity Driven regime, it can be highly suboptimal to not have an SDU. In the ICU Driven regime, it can be optimal to have a non-trivial balking threshold, depending on the relative magnitude of the per-patient balking versus abandonment plus holding cost. We also find that fine-tuning this threshold only has a second order effect on the cost.

1.1. Literature Review

Our work is most related to three bodies of research: 1) medical literature on ICU and SDU care, 2) work in healthcare operations management on capacity and patient flow management, and 3) the queueing literature.

While there exists an extensive body of literature in the medical community on ICUs–there are multiple journals, including *Critical Care* and *Intensive Care Medicine*, devoted to this topic–much less attention has been directed towards SDUs. The majority of work related to SDUs has focused on the impact of SDUs on ICU care. Though there may not be a general consensus as to whether SDUs can be cost-effective for treating semi-critical patients (Keenan et al. 1998), there are a number of studies focused on either specific ailments or at individual institutions which suggest the presence of an SDU can benefit patients. For instance, having an SDU can reduce ICU LOS (Byrick et al. 1986); this is intuitive because patients do not have to reach as high a level of stability to be discharged from the ICU to the SDU rather than the general medical/surgical floor. In a study of patients with Acute Myocardial Infarction, the presence of an SDU was shown to reduce cost by $1.5 million a year for the treatment of patients with moderate risk (Tosteson et al. 1996). It is also argued there that high risk patients should not be treated in the SDU.

There has been some work in operations management looking at staffing in healthcare (e.g. Green et al. (2006b), de Véricourt and Jennings (2008), Yankovic and Green (2011), Yom-Tov and Mandelbaum (2013)). Most of the prior work focuses on a single unit and have not considered the impact of the SDU. In recent work, Best et al. (2015) takes a utilization maximization approach to partitioning hospitals into

different units. The focus is on how many beds to allocate to each *type* of medical service in the general ward. In contrast, we consider multiple *levels* of care: the ICU and SDU. Chan et al. (2014a) also looks at patient flows through the ICU and SDU, but takes an empirical and simulation based approach to consider how SDU capacity impacts patient outcomes. The authors find that when the ICU capacity is fixed, adding more SDU capacity improves patient outcomes, but the gains are marginally decreasing. In contrast, this work uses a queueing approach to gain insights into management of ICU and SDU capacity and patient flows in a scenario where increasing the capacity of the SDU necessarily results in reduced ICU capacity. Indeed, we find scenarios where, due to this capacity tradeoff, it is optimal to have no SDU. Recent work by Mathews and Long (2015) uses a simulation model to examine the role of an SDU in critical care. In contrast to our work, the authors do not consider the operational impact of proposed changes. As such, they find, for example, that allocating all beds to the ICU results in the best outcomes; however, they do not consider the need to hire additional nurses to enable such a configuration.

In capturing the patient flow dynamics through an ICU and an SDU, we consider a modification to the commonly used N-model queueing system (see Figure 16 in Gans et al. (2003)). The N-model arises in our case due to the fact that the ICU consists of flexible beds (servers), while the SDU does not. In our setting, once a Critical patient completes treatment (service) in the ICU, he may transition into a Semi-critical patient who can be treated in either the ICU or SDU. This patient flow dynamic introduces a feedback into our model, which is not captured by existing N-models. In various settings, a threshold priority policy for routing patients to the flexible servers (Bell and Williams 2001, Tezcan and Dai 2010, Ghamami and Ward 2012), and a generalized C-μ priority policy (Mandelbaum and Stolyar 2004, Dai and Tezcan 2008, Gurvich and Whitt 2009b) have been shown to minimize costs for the N-model in heavy traffic asymptotic regimes. With the exception of Wallace and Whitt (2005) and Gurvich and Whitt (2010), in all of these works, prioritization and routing of customers is the primary concern. In contrast, in the hospital setting, routing is largely dictated by medical necessity, so we focus on the question of staffing and sizing of units while assuming that a prioritization and routing rule is given.

There is a rich literature on flexibility in queueing systems (e.g. Green (1985), Hopp et al. (2004), Iravani et al. (2005), Ata and Van Mieghem (2009), Bassamboo et al. (2012), Tsitsiklis and Xu (2012)). An important aspect discussed in this literature is how to design the network topology (pairing, chaining, full flexibility, etc.). Another focus is quantifying how to split the resources between flexible and dedicated servers. For example, there has been a series of recent work which considers this question with respect to tandem systems (Andradottir et al. 2013, Zhang and Ayhan 2013, Kirkizlar et al. 2013). Our work is related to this second category as we determine how to allocate the nurses between the ICU (flexible) and the SDU (dedicated). While we also look at a tandem system, the flow patterns exhibit different dynamics, such as bumping, which arise in a hospital setting.

In developing an understanding of the hospital system, we utilize a number of analytic methods. To start, we examine the system using fluid analysis (e.g. Whitt (2006), Bassamboo and Randhawa (2010)), that uses law-of-large-number principles to evaluate cost terms that are of the order of the arrival rate. Next, we refine our analysis by using diffusion approximations as in Jagerman (1974), Garnett et al. (2002), Mandelbaum and Zeltyn (2009), Kocaga and Ward (2010), that leverage central-limit-theorem type results to evaluate fluctuations about the fluid limit that are of order square-root of the arrival rate. Through the diffusion analysis, we establish a state-space collapse result similar to Gurvich and Whitt (2009a), albeit for different dynamics in a different queueing system. Using these methodologies, we are able to evaluate the average abandonment, holding, balking and bumping costs and optimize the balking threshold and the size of the units to minimize these costs. In our asymptotic analysis we take formal fluid and diffusion limits of the nurse allocation problem and then analyze the corresponding fluid and diffusion optimization problems directly. Using simulations we demonstrate the efficacy of the asymptotic solutions for the original system. This approach is similar to the one taken by Harrison and Zeevi (2004), Rubino and Ata (2009), Kostami and Ward (2009), Akan et al. (2013) and Ata et al. (2013).

2. Model

During a patient's hospital stay, his health state evolves over time. For tractability, and in order to highlight the main tradeoffs, we consider two possible health states for each patient: Critical or Semi-critical (such an approach to patient classification was also considered in Mathews and Long (2015)). If a patient is in the Critical state, he *must* be treated in the ICU. Once the patient is admitted to the ICU, the time he is physiologically considered to be in the Critical state is exponentially distributed with rate μ_C. Once a patient is no longer considered to be in the Critical state, he will become a Semi-critical patient with probability p; with probability $1-p$ he leaves the system, which can practically correspond to a number of different situations, such as the patient being transferred and treated in the ward, being discharged home, or dying. Semi-critical patients can be treated in the SDU or ICU. Regardless of the type of bed, the time a patient is considered to be Semi-critical is exponentially distributed with rate μ_{SC}. Note that these rates specify 'service times', defined as the expected time a patient is in a specific health state when being treated in one of the units; these times do not necessarily correspond to the time a patient is treated in any particular unit.

We consider a system with a fixed number of N nurses. These nurses are flexible in the sense that they can work in either the ICU or SDU. While not all hospitals use critical-care nurses to staff the SDU, many–such as that in Eachempati et al. (2004)–do. For safety reasons, a strict nurse-to-patient ratio must be maintained in each unit. Let r_I ($< r_S$) be the given number of patients each nurse can manage in the ICU (SDU). Our goal is to determine how to allocate nurses between the two units, which is analogous to determining the number of ICU and SDU beds, B_I and B_S. We consider *budget neutral* allocations of nurses, so that we must allocate up to N nurses on salary. No additional nurses can be hired. This means that

$$\frac{B_I}{r_I} + \frac{B_S}{r_S} \leq N \tag{1}$$

so that we allocate up to N nurses to the ICU and SDU while satisfying the nurse-to-patient ratios. We refer to any pair (B_I, B_S) of non-negative integers that satisfy (1) as a feasible bed (nurse) allocation. As critical-care is often a bottleneck in the hospital (Ryckman et al. 2009, Kc and Terwiesch 2012, Beck 2011), we will assume there is ample space in the general medical-surgical ward. This will allow us to focus on the flow of critical and semi-critical patients.

See Figure 1 as an example of an allocation of nurses amongst the ICU and SDU. The nurse-to-patient ratio–i.e. the maximum number of patients a nurse can treat at once–in the ICU is $r_I = 1$ and in the SDU it is $r_S = 3$. There are $N = 8$ nurses who are allocated to $B_I = 6$ ICU beds and $B_S = 6$ SDU beds.

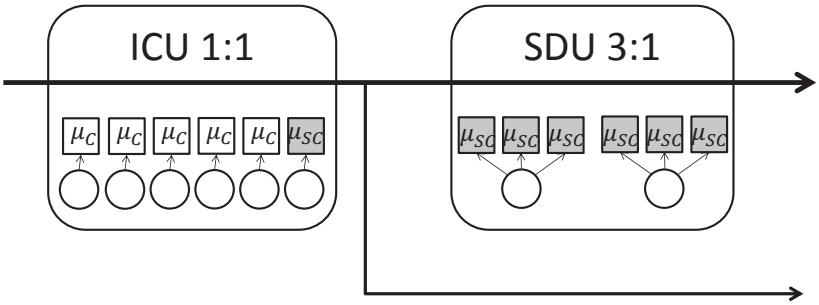


Figure 1 **Nurses are depicted as circles, patients are depicted at squares. Critical patients are served in the ICU. A Critical patient may become a Semi-critical patient upon finishing service in the ICU. Semi-critical patients are depicted in gray and are served in the SDU or ICU. One Semi-critical patient is currently being served in the ICU.**

New Critical patients arrive to the ICU according to a Poisson process with rate λ. If there is space in the ICU, the patient will begin treatment immediately. If there is no space in the ICU, he will wait in a virtual queue. For instance, the patient could wait for ICU admission in the Emergency Department (ED). This queue has length of up to $K \in [0, \infty]$, which is a design parameter the system administrator must select. That is, if a new Critical patient arrives and there are already K Critical patients waiting for ICU admission, the new patient will balk and be sent to a different hospital for care. A cost of w_C^B is incurred for each Critical patient who balks from the queue.

Each Critical patient in the queue incurs a holding cost with rate w_C^H to capture the undesirability of making Critical patients wait. This is undesirable in terms of patient care as well as operationally, as these patients must be treated elsewhere–often in the ED, consuming many resources. If the Critical patient waits too long, he will abandon the queue after an exponential time with rate θ and an abandonment cost of w_C^A is incurred. Note that abandonment corresponds to a patient waiting for ICU care and then eventually rescinding the request after receiving care elsewhere, recovering or dying. This is in contrast to balking

which occurs when a patient's request for ICU care is immediately cancelled upon arrival. For tractability, we use costs for patient balking, abandonment, and holding to capture the undesirability of lack of access to ICU care. Other adverse events of patient wait, such as an increase in LOS (Chan et al. 2013), could also be considered.

If there is a Semi-critical patient in the ICU and all ICU beds are occupied, he can be bumped out by an incoming Critical patient. If there is space for him in the SDU, this bumping comes at no cost. However, if there is no space in the SDU, a current semi-critical patient will be bumped to the general ward resulting in cost w_{SC}. Our queueing model is depicted in Figure 2.

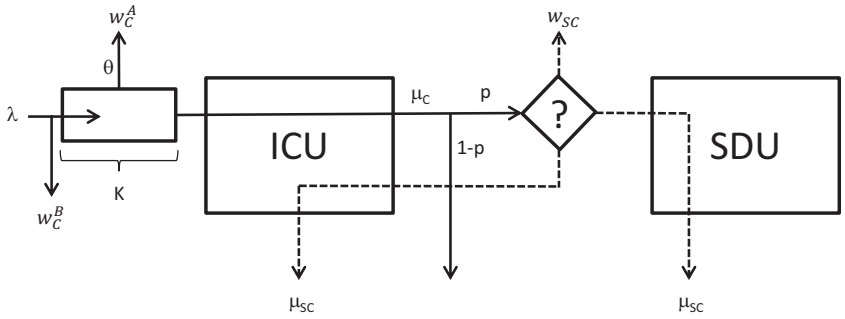


Figure 2 **ICU-SDU queueing model: The '?' represents the assignment decision of a Semi-critical patient. Solid lines depict Critical patient flows while dotted lines depict Semi-critical patient flows.**

Our objective is to minimize the long time average balking, holding, abandonment, and bumping costs. Let $Z_C(t)$ and $Z_{SC}(t)$ denote the number of Critical and Semi-critical patients in the ICU or SDU at time t. $Q(t)$ denotes the number of Critical patients *waiting* in a (virtual) queue. We define a balking function $\xi(Q(t)) : \mathbb{Z}_+ \to \{0, 1\}$ as a function which specifies whether a new arrival would enter the queue given queue length $Q(t)$. In particular, if $Q(t) \geq K$, the patient balks and $\xi = 0$; if $Q(t) < K$, the patient enters the queue and $\xi = 1$. $\psi(Q(t), Z_C(t), Z_{SC}(t)) : \mathbb{Z}_+^3 \to \{0, 1\}$ is a function which specifies whether a Semi-Critical patient will be bumped given system state $(Q(t), Z_C(t), Z_{SC}(t))$. Note that a patient cannot be bumped if he departs the system without becoming a Semi-critical patient (either by balking, abandoning or leaving after completing ICU service). Additionally, a patient cannot abandon if he balks upon arrival. Our objective is thus to determine the allocation of nurses to specify the number of ICU and SDU beds as well as the balking threshold, K, in order to minimize the following cost function:

$$\limsup_{T\to\infty} \frac{1}{T} \int_0^T \left[w_C^B \lambda \xi(Q(t)) + w_C^Q Q(t) + w_{SC}(p\mu_c[B_I \wedge Z_C(t)] + \lambda)\psi(Q(t), Z_C(t), Z_{SC}(t)) \right] dt, \quad (2)$$

where $w_C^Q \triangleq w_C^H + w_C^A \theta$, and $\wedge$ denotes the minimum function. The first component of (2) corresponds to the balking costs; the second component represents the queue length costs, which is the sum of the holding plus the abandonment costs; and the third captures the bumping costs.

In this work, we examine a stylized model of the ICU and SDU. Byrick et al. (1986) found that having an SDU can reduce ICU LOS–this reduction is captured by our service requirements of Critical and Semi-critical patients. With an SDU, the mean LOS of a patient in the ICU will be $1/\mu_C$ plus some additional time depending on if there is space in the ICU to treat him while in the Semi-critical state. However, without an SDU, more Semi-critical patients will be treated in the ICU, thus increasing overall ICU LOS. While there are some practical elements our model does not capture, such as external arrivals to the SDU, readmissions, or treatment of Critical patients in the SDU, it does capture the essence of the tradeoff between increasing capacity for all patient severities versus maximizing capacity for the most vulnerable patients. For tractability, we focus on the patient flows described in this section and find that, in doing so, we can gain many insights into the role of the SDU. In Section 6, we consider some extensions to this initial model.

In considering the possible types of patient dynamics in our system, we found a general consensus amongst physicians we consulted with that Critical patients are typically given priority over Semi-critical patients in the ICU. In what follows, we will assume that strict priority is given to Critical patients, so that a Semi-critical patient will be bumped out of the ICU if a new Critical patient needs the bed. Formally, we make the following assumption throughout the paper:

Assumption 1 *Critical patients obtain strict preemptive priority over Semi-critical patients in the ICU.*

Note that Assumption 1 implies that a Critical patient never balks or queues if there are Semi-critical patients in the ICU.

Remark 1 *In theory, having a single large unit where the level of care of each bed can be dynamically flexed up or down is likely to result in lower costs than fixing the nurse allocation. While a few hospitals have tried to implement units with these flexible capabilities, achieving such benefits in practice has been extremely challenging due to a number of logistical hurdles (e.g. scheduling staff) (see Kwan (2011) and related references). Unit reconfigurations typically occur once or twice a year, if they happen at all. As such, we focus on the strategic decision of nurse allocation to determine the fixed ICU and SDU capacity.*

2.1. Motivating our asymptotic approach

In theory, one could calculate the steady-state distribution of balking, waiting, abandonment and bumping given a balking threshold and a fixed allocation of nurses to the ICU and SDU. Then, an exhaustive search would reveal the balking threshold and the allocation that obtain the lowest cost. Unfortunately, the numerical approach provides little intuition for the general model as to the impact of various system parameters on the optimal solution. In fact, calculating steady-state performance via exact analysis is also extremely difficult because while Critical patients follow an $M/M/B_I/K+M$ queueing model, the number of Semi-critical patients strongly depends on the number of Critical patients in a non-trivial way. The result is a 2-dimensional Markov chain with no known closed-form expression for the steady-state distribution. Hence,

our goal is to develop an understanding of the main drivers of system performance by considering different operational regimes of our ICU and SDU hospital system. The asymptotic regime we consider is one with many nurses. In particular, we consider a sequence of systems indexed by the number of nurses N, with N and λ growing to ∞, while the rest of the parameters do not change. While the average hospital has 15-40 ICU beds (8-40 ICU nurses), we will see via numeric examples in Section 5 that the asymptotic analysis can be quite accurate even with a moderate number of nurses. Our first-order analysis relies on fluid scaling which considers terms of the order of N. Our second-order analysis relies on diffusion scaling, in which we consider fluctuations of the order of $\sqrt{N}$.

3. First Order Analysis

We begin our analysis via a fluid modeling approach. Because ICUs and SDUs are so expensive to operate, hospital administrators do not want to have many empty beds in these units at all times. As a consequence, these units are often operated at or above capacity (Green 2002, Pronovost et al. 2004). With that in mind, we consider a system that is heavily loaded, even if all of the available nurses are optimally allocated between the ICU and the SDU. This assumption is in line with focusing on minimizing costs incurred for limited access to care under the worst-case scenario of some patients being unable to obtain access to a bed. Clearly, during less congested periods the corresponding costs will be lower. If, hypothetically, there were no capacity constraints, it would reasonable to treat all Critical patients in the ICU and all Semi-Critical patients in SDU to minimize the number of busy nurses. Thus, the offered load of Critical patients in the ICU (i.e. the mean number of nurses needed in the ICU) is $\frac{\lambda}{r_I\mu_C}$, while for Semi-Critical patients in the SDU it is $\frac{\lambda p}{r_S\mu_{SC}}$. Our overloaded assumption is such that there are not enough nurses to satisfy this demand. More formally, we postulate the following assumption:

Assumption 2 *The system operates in overload. That is,*

$$\lambda\left(\frac{1}{r_I\mu_C}+\frac{p}{r_S\mu_{SC}}\right)>N. \tag{3}$$

Our asymptotic approach is to consider a sequence of systems indexed by the number of nurses N, in which both N and λ grow without bound, while the rest of the system parameters remain fixed. For notational compactness, we omit the indexing of λ by N. The following proposition justifies our definition of the overloaded regime.

Proposition 1 1. *If $\lambda\left(\frac{1}{r_I\mu_C}+\frac{p}{r_S\mu_{SC}}\right)\leq N$, then there exists a feasible bed allocation and balking threshold such that the total cost rate in Eqn.* (2) *is $o(N)$, where $f(x):=o(x)$ if $f(x)/x\to 0$ as $x\to\infty$.*

2. *Otherwise, if $\lambda\left(\frac{1}{r_I\mu_C}+\frac{p}{r_S\mu_{SC}}\right)>N$, then for any feasible bed allocation and balking threshold the total cost rate in Eqn.* (2) *is at least $\mathcal{O}(N)$, where $f(x):=\mathcal{O}(x)$ if $f(x)/x\leq c>0$ as $x\to\infty$.*

Under the overloaded assumption, we wish to examine the optimal allocation of nurses and balking threshold given the balking, queue length, and bumping cost parameters. To do this, we turn to fluid analysis. The fluid analysis is based on scaling the arrival rate and the number of beds and nurses by $1/N$ and ignoring quantities that are $o(N)$. This way, we can focus on the main drivers of the balking threshold and nurse allocation. We begin by defining our fluid scaling. Let $\bar{\lambda} := \lambda/N, b_i := B_i/N$ for $i = I, S$ and note that by (1),

$$\frac{b_I}{r_I} + \frac{b_S}{r_S} \leq 1. \tag{4}$$

3.1. Balking Threshold

In this section we consider an arbitrary nurse allocation and show that, in the fluid scaling, the optimal balking threshold is either ∞ or 0, independent of this allocation. In determining the optimal balking threshold, K^*, we must consider two cases depending on a relationship between the abandonment rate, the balking cost and the queue length cost.

- **Queue-Dominated Case ($w_C^Q/\theta \leq w_C^B$):** Because the queue length cost is less than that of balking, it is easy to see that patients should never balk. By allowing each Critical patient into the system, at worst, he will wait and abandon, incurring expected cost w_C^Q/θ, rather than the larger w_C^B if the patient is blocked upon arrival. Indeed, following Proposition 1 of Kocaga et al. (2014) we have that, in this case $K^* = \infty$.

- **Balking-Dominated Case ($w_C^Q/\theta > w_C^B$):** Due to the overloaded assumption, for any fixed K, the queue length will be equal to K as long as K is small enough, i.e. $K \leq \bar{q}_{\max}$, where $\bar{q}_{\max} \triangleq \lambda - \mu_C b_I \geq 0$ denotes the maximum queue length on the fluid scale if balking were not allowed. Then the corresponding queue length cost incurred is $w_C^Q K$. The balking cost is $(\bar{\lambda} - b_I\mu_C - \theta K)w_C^B$. Because we are in the overloaded regime, the ICU is *always* filled with Critical patients. As such, the balking threshold only impacts the queue length and balking costs, but not the bumping costs. We determine threshold K^*, which minimizes the cost function $\min_{0\leq K<\infty}\left\{(w_C^Q - \theta w_C^B)K + w_C^B(\bar{\lambda} - b_I\mu_C)\right\}$. Since $w_C^Q/\theta > w_C^B$, we have that $K^* = 0$. That is, having no queue is optimal.

The following proposition summarizes the above discussion.

Proposition 2 *In the fluid model, under the overloaded regime, the optimal balking threshold is given as:*

$$K^* = \infty, \text{ if } w_C = w_C^Q/\theta \leq w_C^B;$$
$$K^* = 0, \quad \text{if } w_C = w_C^B < w_C^Q/\theta.$$

The proof is embedded in the above discussion and is hence omitted.

3.2. Nurse Allocation

We now consider the optimal nurse allocation. We start by defining a critical cost as:

$$w_C = \min\{w_C^Q/\theta, w_C^B\}$$

Note that w_C captures the costs of lack of ICU access for Critical patients. If $w_C = w_C^Q/\theta$ (Queue-Dominated Case), there is no balking. Under our overloaded assumptions we have that $b_I^* \leq \frac{\bar{\lambda}}{\mu_C}$. This is because further increasing the number of nurses allocated to the ICU increases the bumping costs without affecting the queue length cost. Thus, the fluid scaled abandonment rate is equal to the scaled arrival rate minus the scaled service capacity, or $\left(\bar{\lambda} - b_I\mu_C\right)$. Under this allocation, the ICU is always full with Critical patients as there is not enough (or just enough) capacity to serve all Critical patients. Hence, there is no room for Semi-critical patients in the ICU. Thus, the fluid-scaled queue length is equal to the scaled aggregate abandonment rate divided by the individual abandonment rate: $(\bar{\lambda} - b_I\mu_C)/\theta$. This results in an expected scaled queue length cost equal to $\frac{w_C^Q}{\theta}(\bar{\lambda} - b_I\mu_C) = w_C(\bar{\lambda} - b_I\mu_C)$. Using a similar argument, if $w_C = w_C^B$ (Balking-Dominated Case), then there is no queue and, under our overloaded assumptions, the fluid scaled balking rate is equal to $\left(\bar{\lambda} - b_I\mu_C\right)$. Thus, in both cases, the total balking and queue length costs incurred will be: $w_C\left(\bar{\lambda} - b_I\mu_C\right)$.

The fluid-scaled bumping rate from the SDU is equal to the positive part of the scaled SDU arrival rate minus its service rate: $(b_I\mu_C p - b_S\mu_{SC})^+$. Combining these two expressions together gives us the average cost. Recognizing that constraint (4) is satisfied as an equality under the optimal allocation, we can specify our fluid objective in terms of b_I. Our goal is thus to determine, $0 \leq b_I \leq \left(r_I \wedge \frac{\bar{\lambda}}{\mu_C}\right)$ and $0 \leq b_S \leq r_S$, the allocation of nurses to ICU and SDU beds, respectively, so as to minimize the cost function:

$$\min_{0\leq b_I\leq\left(r_I\wedge\frac{\bar{\lambda}}{\mu_C}\right)} \left\{ w_C\left(\bar{\lambda} - b_I\mu_C\right) + w_{SC}\left(b_I\mu_C p - r_S\left(1 - \frac{b_I}{r_I}\right)\mu_{SC}\right)^+ \right\} \tag{5}$$

We can solve the preceding optimization problem to determine how to allocate nurses between the ICU and SDU. We find that the optimal policy is highly dependent on the relationship between w_C and w_{SC}. More formally, we have:

Proposition 3 *In the fluid model, under the overloaded regime, the optimal allocation of nurses can be split into two cases. The cost minimizing allocation of nurses to ICU beds is given by:*

$$b_I^* = \begin{cases} r_I \wedge \frac{\bar{\lambda}}{\mu_C}, & \text{if } \frac{w_C}{w_{SC}} > \frac{r_I p\mu_C + r_S\mu_{SC}}{r_I\mu_C}, \textbf{ ID regime} \\ \frac{r_I r_S\mu_{SC}}{r_I\mu_C p + r_S\mu_{SC}}, & \text{if } \frac{w_C}{w_{SC}} \leq \frac{r_I\mu_C p + r_S\mu_{SC}}{r_I\mu_C}, \textbf{ CD regime} \end{cases} \quad \text{and} \quad b_S^* = r_S\left(1 - \frac{b_I^*}{r_I}\right)$$

Our proposed nurse allocation to ICU and SDU beds, respectively, based on fluid analysis is thus:

$$B_I^* = b_I^* N, \qquad B_S^* = b_S^* N.$$

Note that for notational simplicity, from here on we ignore the integrality constraints. Naturally, our numerical solutions in Section 5 will incorporate integrality constraints. The proof of Proposition 3 is trivial and, hence, omitted. Note that one must verify that the value of b_I^* under the second scenario does not exceed $\bar{\lambda}/\mu_C$, which is true due to the overloaded condition. We have two regimes of interest. When the cost for

lack of ICU access (w_C) is very large, we see the optimal policy is to allocate as many nurses to the ICU as needed in order to satisfy all Critical patients demand (if possible). If there are not enough nurses to meet all of this demand (i.e. $r_I\mu_C < \bar{\lambda}$), then all nurses should be allocated to the ICU. We call this regime the ICU-Driven (ID) regime. On the other hand, when the cost of lack of access to care for Semi-Critical patients (w_{SC}) is close to that for Critical patients, then the optimal policy is to allocate some nurses to the SDU and reduce access to care for Critical patients. We call this regime the Capacity-Driven (CD) regime: the larger the capacity gained by transferring a nurse from the ICU to the SDU (increasing $\frac{r_S\mu_{SC}}{r_I\mu_C}$), the more likely the CD regime is to be optimal. Additionally, if many Critical patients become Semi-critical (large p) the SDU becomes more beneficial.

We observe that the regimes are set such that, whenever possible, one would incur either Critical patients related costs or Semi-critical patients related costs, but not both. Indeed, in the ID regime only bumping costs are incurred, as long as there is enough capacity to accommodate all Critical patients. In contrast, in the CD regime, the system will only incur Critical patients related costs. Moreover, in the latter regime, the system incurs either balking costs or queue costs, but not both. We additionally observe that the bed-allocation scheme proposed by our fluid analysis is very *robust* with respect to the system parameters, as long as the system operates away from the threshold $\frac{w_C}{w_{SC}} = \frac{r_Ip\mu_C + r_S\mu_{SC}}{r_I\mu_C}$.

In further interpreting the results of Proposition 3, we have that in the CD regime, the SDU size is selected such that the SDU is *critically* loaded, $\lambda_{SDU} \approx B_I^*\mu_C p \approx B_S^*\mu_{SC}$, while the ICU is strictly overloaded (by Proposition 1). This is surprising because it occurs even when lack of access to the ICU, via balking or queue length costs, is more costly than bumping an SDU patient. Yet, this allocation results in having balking rate (or queue length costs) which is of order N and bumping rate which is of order $o(N)$. In the CD regime, the capacity gains of allocating nurses to the SDU are more substantial than the gain of keeping the nurses in the ICU to serve the high priority (Critical) patients. In the ID regime, the needs of the Critical patients dominate. In fact, we see that in both the ID and CD regime, if it is possible, the optimal solution is such that enough nurses should be allocated to one of the two units to make it critically loaded, necessarily making the other unit overloaded. The dominating unit depends on the relationship between the system parameters, $w_C = \min\{w_C^Q/\theta, w_C^B\}$ and w_{SC}.

In practice, we see that some hospitals have SDUs while others do not. We expect hospitals to have differing system parameters based on varied patient mixes and regulations; moreover, they may view the relative costs between balking, abandonment, holding and bumping differently. Our analysis suggests then, that the variation of SDU use in practice may be warranted.

4. Second Order Analysis

In this section, we consider refining our analysis from Section 3 by examining the impact of reallocating a small number of nurses to either the ICU or SDU. Our starting point is the analysis of the fluid approximation in Section 3. Under the ID regime it is optimal to have as big of an ICU as necessary/possible, while

in the CD regime, it is optimal to have an SDU which is comparable in size to the ICU. In this section, we consider how the reallocation of a small number of nurses may help. We find that in some cases, this reallocation can be quite helpful. The fluid analysis finds the optimal allocation of nurses to the ICU and SDU up to an order of $o(N)$. In particular, the fluid analysis excludes these lower ordered terms and so it might still be beneficial to reallocate a small number of nurses, say of order $\mathcal{O}(\sqrt{N})$ to the SDU or ICU. We will use *diffusion* analysis to examine these two regimes.

4.1. Diffusion Analysis in the ID regime

Recall that in the ID regime, the fluid solution allocates all nurses possible/required to the ICU so that the queue plus balking cost is $o(N)$ (if possible), or negligible in fluid scale. We now explore the benefits of reallocating a *small* number of nurses, of order $\mathcal{O}(\sqrt{N})$ between the ICU and the SDU. In this section we assume that

$$\frac{w_C}{w_{SC}} > \frac{r_I\mu_C p + r_S\mu_{SC}}{r_I\mu_C},$$

and therefore, on a fluid level, it is optimal to operate the system in the ID regime. Additionally, suppose that the number of nurses is large enough to satisfy

$$Nr_I \geq \frac{\lambda}{\mu_C} + o(N). \tag{6}$$

That is, the number of beds allocated to the ICU is $B_I^* = \lambda/\mu_C + o(N)$, and the ICU is critically loaded with respect to the Critical patients.

We now postulate the following refinement of the above nurse allocation scheme:

$$B_I = \frac{\lambda}{\mu_C} + \beta\sqrt{\frac{\lambda}{\mu_C}} + o(\sqrt{N}), \quad B_S = \frac{r_S}{r_I}\left(Nr_I - \frac{\lambda}{\mu_C} - \beta\sqrt{\frac{\lambda}{\mu_C}}\right) + o(\sqrt{N}), \tag{7}$$

where β is only restricted by the non-negativity constraints on B_I and B_S. In particular, the ICU is critically loaded when focusing on Critical patients and works under the QED regime (Halfin and Whitt 1981, Garnett et al. 2002) with respect to the same patients. At the same time, due to our overloaded condition and by Proposition 1, the SDU is overloaded.

It is not clear that in this operating regime, the ICU and SDU will always be full with Critical and Semi-critical patients, respectively, as is the case under the fluid scaling. Because the ICU may not be full of Critical patients, the dynamics of our queueing system and, specifically, the flow of the Semi-critical patients is more complex. Before we can determine the optimal allocation of nurses, we must first understand more precisely when and to what extent Semi-critical patients will be treated in the ICU.

4.1.1. State-Space Collapse In order to develop an understanding of the patient flow dynamics, one can examine the two-dimensional process with state $(Q + Z_C, Z_{SC})$ (recall that Q denotes the queue length and Z_C (Z_{SC}) denotes the number of Critical (Semi-critical) patients occupying a bed). This process is clearly a Markov process under the strict priority of Critical patients over Semi-critical patients in the ICU; however, the dynamics of this process are intricate. While the dynamics of the Critical patients follow that of a fairly standard multiserver queue with finite/infinite waiting room and abandonment, the dynamics of the Semi-critical patients cannot be analyzed separately from the Critical patients; the dynamics of the Critical patients determine precisely the arrival rate into the Semi-critical state and also how many beds are available in the ICU to treat these patients.

Given our goal is to gain some insights as to how to allocate the nurses between the two units in this case, it is important to be able to characterize the patient flows through the ICU and SDU. Despite the challenges which arise with the two-dimensional Markovian model, we are able to show that this two-dimensional process may be accurately approximated by a one-dimensional process. Let

$$\hat{Z}_C^N := \frac{1}{\sqrt{\lambda}}\left(Z_C^N - B_I^N\right), \quad \hat{Z}_{SC}^N := \frac{1}{\sqrt{\lambda}}\left(Z_{SC}^N - B_S^N\right),$$

describe the diffusion scaled number of patients occupying a bed within each of the two states, respectively. Also, let $\Rightarrow$ represent weak convergence. Then we have:

Theorem 1 ***(State-Space Collapse)*** *In the ID regime and under the nurse allocation of (7) we have a state-space collapse. More formally, assuming that at time 0, $\hat{Z}_C^N(0) + \hat{Z}_{SC}^N(0) \Rightarrow 0$, as $N \to \infty$, then*

$$\hat{Z}_C^N + \hat{Z}_{SC}^N \Rightarrow 0, \ \text{ as } N \to \infty,$$

where the convergence is in D the space of all RCLL (Right Continuous with Left Limits) functions with values in $\mathbb{R}$, equipped with the Skorohod J_1 metric (see Whitt (2002)).

According to Theorem 1, in the diffusion scale, all beds are always full. In particular, it is sufficient to know the value of the one dimensional process $X_C^N := Q^N + Z_C^N$ in order to figure out the value of the two dimensional process (X_C^N, Z_{SC}^N) (up to order $o(\sqrt{N})$). For example, if there is no queue ($X_C^N \leq B_I$ so that $Q^N = 0$), then we know that any ICU bed which is not occupied by a Critical patient will be used to treat a Semi-critical patient. Hence the term 'State-space collapse'. Specifically, the dynamics of our system can be summarized as follows:

1. The ICU is operated in the QED-regime with respect to Critical patients, so the number of Critical patients can be approximated by the diffusion analysis of an Erlang-A model with finite or infinite buffer (Garnett et al. 2002, Kocaga and Ward 2010) with B_I servers.

2. The SDU is always full. If there are fewer than B_I Critical patients in the system, then Semi-critical patients fill the remaining ICU beds.

The second point implies that even if the ICU is not overly crowded with critical patients it will always be full and thus appear as if it is operating in the overloaded regime. This raises an important practical insight: when examining hospital data a unit that is always full may appear to be a system bottleneck where, in fact, the reason why it is full could be due to spillover from other units. In the ID regime this observation applies to an always full ICU, with some Semi-critical patients who are treated there due to lack of capacity is the SDU. While a natural reaction to observing ICUs which are constantly full is to add more ICU capacity, the real culprit of such congestion may be inadequate SDU capacity.

The intuition behind Theorem 1 is as follows: The SDU is overloaded. In particular, the rate at which it is losing patients due to lack of space is of order N. At the same time the ICU is in the QED regime with respect to Critical patients. In particular, the number of ICU beds that are not occupied by Critical patients is at most of order $\mathcal{O}(\sqrt{N})$. As soon as some of these beds are empty they almost instantaneously become occupied by Semi-critical patients. Hence all beds are always full.

We now leverage our results from above to examine the nurse allocation and balking threshold problem. Our aim is to derive expressions for the cost function using a diffusion approximation. Given the state-space collapse result that applies to the process $(Q + Z_C, Z_{SC})$, it is reasonable to expect that a similar state-space collapse applies in steady-state as well. Establishing this requires a formal justification of a limit interchange argument as in Theorem 9.10 of Ethier and Kurtz (1985). To avoid a lengthy and rather technical mathematical argumentation here we simply postulate that the same state-space collapse holds in steady-state as well.

Let $\hat{Q}^N := \frac{Q^N}{\sqrt{\lambda}}$ and $\hat{I}^N = \frac{I^N}{\sqrt{\lambda}}$ be the scaled queue length and "idleness" processes, where I^N is the number of ICU beds not occupied by Critical patients. Note that due to Theorem 1, I^N is also approximately equal to the number of Semi-critical patients who are being treated in the ICU. With a slight abuse of notation we also let $\hat{Q}^N$ and $\hat{I}^N$ represent these quantities in steady-state. Also, let $\hat{L}^N$ be the steady-state balking rate.

4.1.2. Diffusion cost function: The Queue-Dominated case Recall that in the queue-dominated case ($w_C^Q/\theta \leq w_C^B$) it is never optimal to let a patient balk (Proposition 1 of Kocaga et al. (2014)), and therefore the optimal balking threshold is $K^* = \infty$. In particular, in this case, the ICU operates as an $M/M/N+M$ system with respect to the critical patients. To evaluate the steady-state cost, we begin by stating a result that follows directly from results in Garnett et al. (2002) and Browne and Whitt (1995). Note, one could also consider using an alternative approximation, such as that in Baron and Milner (2009).

Theorem 2 ***(Erlang-A in Steady-State)*** *In the ID regime, and under the nurse allocation in (7), we have that* $(\hat{Q}^N, \hat{I}^N) \Rightarrow (\hat{Q}, \hat{I})$*, as* $N \to \infty$*, with*

$$E[\hat{Q}] = \left(1 + \frac{h(\beta\sqrt{\mu_C/\theta})}{\sqrt{\mu_C/\theta} \cdot h(-\beta)}\right)^{-1} \cdot \left(-\frac{\sqrt{\mu_C}\beta}{\theta} + \sqrt{\frac{1}{\theta}} \cdot h\left(\beta\sqrt{\frac{\mu_C}{\theta}}\right)\right)$$

and

$$E[\hat{I}] = \frac{1}{\sqrt{\mu_C}} \left(1 - \left(1 + \frac{h(\beta\sqrt{\mu_C/\theta})}{\sqrt{\mu_C/\theta} \cdot h(-\beta)} \right)^{-1} \right) \cdot (\beta + h(-\beta)),$$

where $h(x) = \frac{\phi(x)}{1-\Phi(x)}$ *is the hazard rate function of the Standard Normal distribution.*

Note that Theorem 2 states the weak convergence of the steady-state random variable $(\hat{Q}^N, \hat{I}^N)$ but does not argue that convergence in expectation applies as well. This requires an additional technical argument which we omit. We simply postulate the convergence in expectation applies as well.

We now derive diffusion approximations for the bumping rates. Let Bm denote the steady-state bumping rate. The starting point is that the bumping rate is equal to the Semi-critical arrival rate minus its total service rate. The arrival rate may be expressed as: $E[Z_C]\mu_C p$. Similarly, and assuming that the SSC result of Theorem 1 holds in steady-state, the departure rate may be expressed as: $B_S\mu_{SC} + E[I]\mu_{SC} + o(\sqrt{N})$. Putting all of the above together we see that, under the ID regime and the nurse allocation (7), the cost function (centered by $w_{SC}\left(\lambda p + \frac{r_S}{r_I}\left(Nr_I - \frac{\lambda}{\mu_C}\right)\mu_{SC}\right)$ and scaled by $1/\sqrt{\lambda}$) may be approximated by:

$$C(\beta) := w_C^Q E[\hat{Q}] + w_{SC}\left[\beta\sqrt{\mu_C}p + \frac{r_S\beta\mu_{SC}}{r_I\sqrt{\mu_C}} - (\mu_{SC} + \mu_C p)E[\hat{I}]\right], \tag{8}$$

where the expressions for $E[\hat{Q}]$ and $E[\hat{I}]$ are explicitly given in Theorem 2.

Let $\beta^* := \arg\min_\beta C(\beta)$, where we choose the supremum on β if there are multiple values of β that minimize the cost $C(\beta)$. Then our proposed solution in the ID regime is:

$$B_I^* := \frac{\lambda}{\mu_C} + \beta^*\sqrt{\frac{\lambda}{\mu_C}}, \quad B_S^* = \frac{r_S}{r_I}\left(Nr_I - \frac{\lambda}{\mu_C} - \beta^*\sqrt{\frac{\lambda}{\mu_C}}\right).$$

Note that we have not imposed upper and lower bounds on β^*. In particular, it is plausible that β^* is so small (including $\beta^* = -\infty$), that B_I^* is in fact smaller than what is proposed by the CD regime, even though, by assumption, the system operates in the ID regime. To remedy this, we set a lower bound on B_I^* and an upper bound on B_S^* that are dictated by the fluid solution. In doing so, the allocation of nurses is given by:

$$B_I^* := \max\left\{\frac{\lambda}{\mu_C} + \beta^*\sqrt{\frac{\lambda}{\mu_C}},\ \frac{r_I r_S \mu_{SC}}{r_I\mu_C p + \mu_{SC} r_S}N\right\},$$

and

$$B_S^* = \min\left\{\frac{r_S}{r_I}\left(Nr_I - \frac{\lambda}{\mu_C} - \beta^*\sqrt{\frac{\lambda}{\mu_C}}\right),\ \frac{r_I r_S \mu_C p}{r_I\mu_C p + r_S\mu_{SC}}N\right\}.$$

In Section 5, we will see through numeric experiments, that indeed, in the ID regime the optimal value of β might be such that a small number of beds should be allocated to the SDU.

In the ID regime, the ICU is operated in QED with respect to the Critical patients. Hence, some Semi-critical patients will be treated in the ICU, so we can see that the reallocation of beds in this regime translates to balancing the tradeoff between capacity for the most critical patients (ICU beds) versus overall capacity (SDU beds). Note that in the ID regime, this tradeoff only arises in this second order analysis.

4.1.3. Diffusion cost function: The Balking-Dominated case In this case, the per-patient balking cost w_C^B is less expensive than the worst-case queue cost w_C^Q/θ. Therefore, in the fluid scale, it is optimal to have all critical patients for whom an ICU bed is not immediately available balk. In the diffusion scale, things are more subtle. Here, it might be worthwhile to let critical patients wait in queue and not balk in the hope that a bed would become available to them after a relatively short amount of time, so that the queue cost per patient is small. This tradeoff was explored in Kocaga and Ward (2010). Consistent with that paper, we consider a balking threshold K^N which is of order $\mathcal{O}(\sqrt{N})$. For simplicity assume that $K^N = k\sqrt{N}$. The next result follows directly from results in Kocaga and Ward (2010).

Theorem 3 ***(ID Diffusion performance)*** *In the ID regime, and under the nurse allocation in (7) and for balking threshold $K^N = k\sqrt{N}$, we have that $(\hat{Q}^N, \hat{I}^N, \hat{L}^N) \Rightarrow (\hat{Q}, \hat{I}, \hat{L})$, as $N \to \infty$,*

$$E[\hat{Q}] = \frac{1}{\theta\sqrt{\mu_C}} \cdot \frac{1 - \exp\left(\frac{-\theta}{\sigma^2}\left(k^2 + 2\frac{m}{\theta}k\right)\right) + \frac{2}{\sigma}\sqrt{\frac{\pi}{\theta}} m e^{\frac{m^2}{\theta\sigma^2}} \left(\Phi\left(\frac{m}{\sigma}\sqrt{\frac{2}{\theta}}\right) - \Phi\left(\frac{\sqrt{2\theta}}{\sigma}\left(k + \frac{m}{\theta}\right)\right)\right)}{\frac{2}{\sigma}\sqrt{\pi}\left(\frac{1}{\sqrt{\mu_C}} e^{\frac{m^2}{\mu_C\sigma^2}} \Phi\left(\frac{m}{\sigma}\sqrt{\frac{2}{\mu_C}}\right) + \frac{1}{\sqrt{\theta}} e^{\frac{m^2}{\theta\sigma^2}} \left(\Phi\left(\frac{\sqrt{2\theta}}{\sigma}\left(k + \frac{m}{\theta}\right)\right) - \Phi\left(\frac{m}{\sigma}\sqrt{\frac{2}{\theta}}\right)\right)\right)}$$

and

$$E[\hat{I}] = \frac{1}{\sqrt{\mu_C}} \cdot \frac{\frac{1}{\mu_C}\left(1 + \frac{2}{\sigma}\sqrt{\frac{\pi}{\mu_C}} m e^{\frac{m^2}{\mu_C\sigma^2}} \Phi\left(\frac{m}{\sigma}\sqrt{\frac{2}{\mu_C}}\right)\right)}{\frac{2}{\sigma}\sqrt{\pi}\left(\frac{1}{\sqrt{\mu_C}} e^{\frac{m^2}{\mu_C\sigma^2}} \Phi\left(\frac{m}{\sigma}\sqrt{\frac{2}{\mu_C}}\right) + \frac{1}{\sqrt{\theta}} e^{\frac{m^2}{\theta\sigma^2}} \left(\Phi\left(\frac{\sqrt{2\theta}}{\sigma}\left(k + \frac{m}{\theta}\right)\right) - \Phi\left(\frac{m}{\sigma}\sqrt{\frac{2}{\theta}}\right)\right)\right)}$$

where $m := \beta\mu_C$ and $\sigma^2 = 2\mu_C$. Additionally, we have that the scaled balking rate is:

$$\hat{L} = \frac{1}{\sqrt{\mu_C}} \cdot \frac{e^{\frac{-\theta}{\sigma^2}\left(k^2 + 2\frac{m}{\theta}k\right)}}{\frac{2}{\sigma}\sqrt{\pi}\left(\frac{1}{\sqrt{\mu_C}} e^{\frac{m^2}{\mu_C\sigma^2}} \Phi\left(\frac{m}{\sigma}\sqrt{\frac{2}{\mu_C}}\right) + \frac{1}{\sqrt{\theta}} e^{\frac{m^2}{\theta\sigma^2}} \left(\Phi\left(\frac{\sqrt{2\theta}}{\sigma}\left(k + \frac{m}{\theta}\right)\right) - \Phi\left(\frac{m}{\sigma}\sqrt{\frac{2}{\theta}}\right)\right)\right)}.$$

Similar to the previous case, under the ID regime, the balking-dominated case, and the nurse allocation (7), the cost function (centered by $w_{SC}\left(\lambda p + \frac{r_S}{r_I}\left(Nr_I - \frac{\lambda}{\mu_C}\right)\mu_{SC}\right)$ and scaled by $1/\sqrt{\lambda}$) may be approximated by:

$$C(m,k) := w_C^B \hat{L} + w_C^Q E[\hat{Q}] + w_{SC}\left[\frac{m}{\sqrt{\mu_C}} p + m\frac{r_S\mu_{SC}}{r_I\mu_C^{3/2}} - (\mu_{SC} + \mu_C p)E[\hat{I}]\right], \tag{9}$$

where the expressions for $\hat{L}$, $E[\hat{Q}]$ and $E[\hat{I}]$ are explicitly given in Theorem 3.

Let $(m^*, k^*) := \arg\min_{m,k} C(m,k)$, where we choose the supremum on m (and k) if there are multiple values of m (k) that minimize the cost $C(m,k)$. Then our proposed solution in the ID regime is:

$$B_I^* := \frac{\lambda}{\mu_C} + \frac{m^*}{\mu_C}\sqrt{\frac{\lambda}{\mu_C}}, \quad B_S^* = \frac{r_S}{r_I}\left(Nr_I - \frac{\lambda}{\mu_C} - \frac{m^*}{\mu_C}\sqrt{\frac{\lambda}{\mu_C}}\right), \quad K^* = k^*\sqrt{N}.$$

Similar to the queue-dominated case, we set a lower bound on B_I^* and an upper bound on B_S^* that are dictated by the fluid solution. In doing so, the allocation of nurses is given by:

$$B_I^* := \max\left\{\frac{\lambda}{\mu_C} + \frac{m^*}{\mu_C}\sqrt{\frac{\lambda}{\mu_C}},\ \frac{r_I r_S \mu_{SC}}{r_I\mu_C p + \mu_{SC} r_S} N\right\},$$

and

$$B_S^* = \min\left\{\frac{r_S}{r_I}\left(N r_I - \frac{\lambda}{\mu_C} - \frac{m^*}{\mu_C}\sqrt{\frac{\lambda}{\mu_C}}\right),\ \frac{r_I r_S \mu_C p}{r_I\mu_C p + r_S\mu_{SC}} N\right\}.$$

In Section 5, we see cases where the values of both the optimal k and m are non-trivial.

4.2. Diffusion Analysis in the CD regime

Recall that the fluid analysis identified two operating regimes for the system: the ID and CD regimes. Now we take a closer look at the CD regime. In particular, we focus on the case where

$$\frac{w_C}{w_{SC}} \le \frac{r_I\mu_C p + r_S\mu_{SC}}{r_I\mu_C}.$$

In this case, according to Proposition 3, we have that

$$B_I^* = b_I^* N + o(N), \quad b_I^* = \frac{r_I r_S \mu_{SC}}{r_I\mu_C p + r_S\mu_{SC}}, \text{ and } B_S^* = b_S^* N + o(N), \quad b_S^* = \frac{r_I r_S \mu_C p}{r_I\mu_C p + r_S\mu_{SC}}.$$

In particular, we have that the ICU is overloaded and the SDU is critically loaded. Our aim here is to see whether an order of $\sqrt{N}$ refinement for the $o(N)$ terms above can lead to a lower cost. We further assume that $\lambda = \mathcal{O}(N)$ so that the ICU operates in the efficiency-driven (ED) regime (Gans et al. 2003). Otherwise, the ICU will be "super" overloaded, and refinements of this order will not make a noticeable difference. Set

$$B_I = b_I^* N + o(N) = \gamma R_I + \delta\sqrt{R_I} + o(\sqrt{R_I}), \quad R_I := \frac{\lambda}{\mu_C}, \tag{10}$$

where $\gamma = \frac{N r_I r_S \mu_{SC}\mu_C}{\lambda(r_I\mu_C p + r_S\mu_{SC})}$ is less than 1 due to our overloaded regime condition. Also, let

$$B_S = b_S^* N + o(N) = R_S + \beta\sqrt{R_S} + o(\sqrt{R_S}), \quad R_S := \frac{B_I\mu_C p}{\mu_{SC}}, \tag{11}$$

where β and δ are only restricted by the non-negativity constraints on B_I and B_S. R_I is the offered load of the ICU, by definition. We argue that R_S is the offered load of the SDU. To see this, note that, since $\gamma < 1$, the ICU is indeed operated in the overloaded regime. In particular, all ICU beds are full with Critical patients all the time, almost surely. Hence, the arrival rate into the SDU is equal to $B_I\mu_C p$, and the offered load is indeed equal to $\frac{B_I\mu_C p}{\mu_{SC}}$. Note that, as expected, the SDU is critically loaded, and operates in the QED regime. Finally, using the relation $\frac{B_I}{r_I} + \frac{B_S}{r_S} = N + o(\sqrt{N})$ we obtain that

$$\delta := \delta(\beta) = -\sqrt{\frac{N}{\lambda}}\frac{\beta r_I\mu_C\mu_{SC}\sqrt{\frac{r_I r_S p}{r_I\mu_C p + r_S\mu_{SC}}}}{r_I\mu_C p + r_S\mu_{SC}}.$$

We aim to find a value for β that minimizes the expected balking plus queue plus bumping cost.

We first argue that in the CD regime, the optimal balking threshold is $K^* = \infty$ or $K^* = o(\sqrt{N})$ depending on whether the queue or the balking dominated case holds, respectively. This implies that the system incurs either queue or balking cost, but not both (up to an order of $o(\sqrt{N})$. We have already established that in the queue-dominated case the optimal threshold is equal to ∞. The balking-dominated case is more involved. We wish to show that in this case $K^* = o(\sqrt{N})$. This requires a few steps. First notice that by Theorem 4.3 of Mandelbaum and Zeltyn (2009), we have that in this regime, when no balking occurs

$$\lim_{\lambda\to\infty} \frac{EQ^N}{\lambda} = \frac{1-\gamma}{\theta}.$$

We now argue that whenever the balking threshold K^N is smaller than EQ^N, then the queue length is always equal to K^N up to an order of $o(\sqrt{N})$.

Proposition 4 ***(balking threshold in the CD regime)*** *In the CD regime and under the nurse allocation of (10) and (11) if a threshold policy is used with threshold K^N that satisfies*

$$\limsup_{N\to\infty} \frac{K^N}{(1-\gamma)\lambda/\theta} = 1-\eta, \quad 0<\eta\leq 1, \tag{12}$$

then, the buffer is always full. More formally, assuming that at time 0, $\frac{Z_C^N(0)+Q^N(0)-(B_I^N+K^N)}{\sqrt{N}} \Rightarrow 0$, as $N\to\infty$, then

$$\frac{Z_C^N + Q^N - (B_I^N + K^N)}{\sqrt{N}} \Rightarrow 0, \ \text{ as } N\to\infty,$$

where the convergence is in D the space of all RCLL (Right Continuous with Left Limits) functions with values in $\mathbb{R}$, equipped with the Skorohod J_1 metric.

Corollary 1 *Under the conditions of Proposition 4, we have that the number of ICU beds that are occupied by critical patients is equal to $B_I - o(\sqrt{N})$.*

Corollary 2 *Under the conditions of Proposition 4, the optimal threshold in the balking-dominated case satisfies $K^{*N} = o(\sqrt{N})$.*

An interesting conclusion from the results above is that, in the CD regime, the system will either incur queue costs or balking costs but not both. In the balking-dominated case the balking rate is equal to $\lambda - \mu_C B_I + o(\sqrt{N})$, and the corresponding balking cost is $w_C^B \cdot (\lambda - \mu_C B_I) + o(\sqrt{N})$. In the queue-dominated case we have that the average queue length satisfies $EQ = \frac{\lambda-\mu_C B_I}{\theta} + o(\sqrt{N})$, and the corresponding queue cost is $w_C^Q \cdot \frac{\lambda-\mu_C B_I}{\theta} + o(\sqrt{N})$. Thus, recalling that $w_C = \min\{w_C^Q/\theta, w_C^B\}$, we have that the total queue plus balking cost in the CD regime is

$$w_C \cdot (\lambda - \mu_C B_I) + o(\sqrt{N}) = w_C \cdot \lambda \left(1 - \gamma - \delta/\sqrt{\frac{\lambda}{\mu_C}}\right) + o(\sqrt{N}).$$

For the bumping rate, from Jagerman (1974), we have that

$$Pr\{Bm\} = \frac{1}{\sqrt{B_S}} h(-\beta) + o(1/\sqrt{\lambda}).$$

Adding the two cost components together, centering by $w_C\lambda(1-\gamma)$, scaling by $1/\sqrt{N}$, and letting $N \to \infty$, we obtain that the relevant cost function may be approximated by

$$C(\beta) = \mu_{SC}\sqrt{\frac{r_I r_S \mu_C p}{r_I\mu_C p + r_S\mu_{SC}}}\left(w_C \frac{\beta r_I \mu_C}{r_I\mu_C p + r_S\mu_{SC}} + w_{SC}h(-\beta)\right). \tag{13}$$

Let $\beta^* := \arg\min_\beta C(\beta)$, and let $\delta^* := \delta(\beta^*)$. Analogously to the ID regime, it is plausible that β^* is so small that the proposed B_I^* is larger than what is proposed by the ID regime. We set an upper bound on B_I^* and a lower bound on B_S^* that are dictated by the fluid solution. Then our proposed solution in the CD regime is:

$$B_I^* = \min\left\{\gamma R_I + \delta^*\sqrt{R_I}, r_I N, \frac{\lambda}{\mu_C}\right\}, \quad R_I := \frac{\lambda}{\mu_C}, \tag{14}$$

and

$$B_S^* = \max\left\{R_S^* + \beta^*\sqrt{R_S^*}, \frac{r_S}{r_I}\left(Nr_I - \frac{\lambda}{\mu_C}\right)\right\}, \quad R_S^* := \frac{B_I^*\mu_C p}{\mu_{SC}}. \tag{15}$$

5. Numeric Results

We have utilized fluid and diffusion analysis to determine how to allocate nurses to ICU and SDU beds. We find that two operational regimes exist: the ID regime in which the SDU has very few beds, if any, and the CD regime in which the SDU is comparable in size to the ICU. We now use numerical approaches to examine the quality of our approximations.

5.1. Empirical Data

To start, we must first calibrate the parameters of our model. To do this, we leverage the existing medical literature. Given the limited literature on SDUs, we identified two articles which specify the necessary parameters for our queueing model. The first article looks at the impact of adding an SDU for the cardiothoracic ICU at the University of Missouri Hospitals (Cady et al. 1995). The second article also considers the impact of introducing an SDU, but this time for the surgical ICU at New York-Presbyterian Hospital (Eachempati et al. 2004). The parameters from these articles are summarized in Table 1. We let $\mu_{SC} = 1/ICULOS$ and $\mu_C = 1/SDULOS$. Note that this ignores Semi-Critical patients who may be treated in the ICU as well as censored observations due to abandonment and bumping. Based on conversations with medical professionals who suggested that patients could wait on average up to 1 day for an ICU bed, we set $\theta = 1$. We will vary the cost parameters to see how they impact the balking and staffing decisions.

Source	ICU LOS	SDU LOS	p	r_I	r_S
Cady et al. (1995)	2.5 days	1.2 days	0.65	1†	2-3
Eachempati et al. (2004)	4.8 days	2.3 days	0.8	2	4

Table 1 Summary of ICU and SDU patient flow parameters. †The ICU nurse-to-patient ratio is not given in this article, so we assume it to be one-to-one.

5.2. Simulation Results

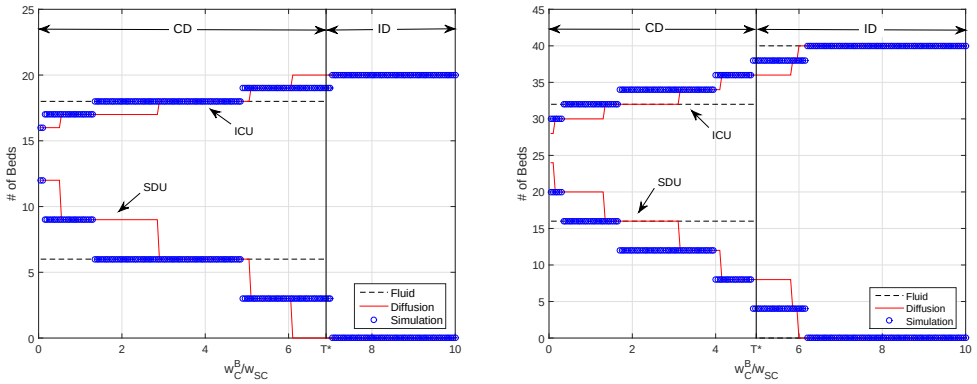
We now leverage the parameters from Table 1 to simulate patient flows through the ICU and SDU. Using an exhaustive search over simulations which examine the average costs incurred under every combination of nurse allocations and balking thresholds from 0 to 100 as well as ∞, we find the optimal number of ICU and SDU beds and the optimal balking threshold. We then compare this to the allocation of nurses given by our fluid and diffusion analysis.

We start by assuming the arrival rate is such that the ICU is critically loaded in case all the nurses are allocated to the ICU. Specifically, $\lambda = N\mu_C r_I$. For the Queue-dominated (Balking-dominated) case, we set $w_{SC} = 1$ and $w_C^B = 15$ ($w_C^Q/\theta = 15$), while we vary the critical cost: w_C^Q/θ (w_C^B).

In considering the staffing level in the ICU, we expect the number of ICU beds to be non-decreasing in the ratio between the critical cost and bumping cost: w_C/w_{SC}. It turns out that because we have two different solution regimes (ID and CD) at the diffusion level, it is possible the monotonicity is violated near the transition between these two regimes, i.e. when $w_C/w_{SC} = T^* := \frac{r_I\mu_C p + r_S\mu_{SC}}{r_I\mu_C}$. Indeed, we encounter this issue in our numeric analysis in some scenarios. For such scenarios, in order to translate our diffusion solution to maintain the desired monotonicity, at T^*, we assigned the number of ICU beds to be the average between the ID and CD diffusion solutions. That is, let $B_I^*(\text{ID}, T^*)$ be the ID solution (minimizes Eqn. (8) or (9)) and let $B_I^*(\text{CD}, T^*)$ be the CD solution (minimizes Eqn. (13)) when $w_C/w_{SC} = T^*$. Then, our diffusion solution is $B_I^* = \frac{1}{2}[B_I^*(\text{ID}, T^*) + B_I^*(\text{CD}, T^*)]$, which also serves as a lower (upper) bound for the number of ICU beds in the ID (CD) regime.

Figure 3 compares the nurse allocation from our analysis to the exhaustive search when there are 20 nurses to split amongst the ICU and SDU in the Balking-dominated case. As we can see in these figures, the solution determined by minimizing the cost in (9) and (13) is very close to the solution determined by using exhaustive search over simulations. The fluid model is fairly accurate for many different weights, but can be quite coarse at times. Additionally, the accuracy of our approximation depends on the size of the system, with better results for larger systems. Because the nurse-to-patient ratios in Eachempati et al. (2004) require fewer nurses per patient than in Cady et al. (1995), the size of the units is twice as large for the Eachempati et al. (2004) parameters. As such, the quality of the solution from the diffusion analysis is more accurate in Figure 3b than in Figure 3a. We also find that when we increase the number of nurses to allocate (for instance, to $N = 100$), the approximations become even more accurate. Because the queue cost, $w_C^Q/\theta = 15$, is so large compared to the balking costs, $w_C^B/ \in [0, 10]$, the optimal balking threshold is

0 for the fluid and exhaustive search. It is also 0 for the diffusion solution in the CD regime, but ranges from 1 to 3 in the ID regime. We saw in Section 4.1.3 that it can be optimal to have a balking threshold of order $\sqrt{N}$; our numerics confirm this and we explore it further in Section EC-2.1 of the Electronic Companion (EC).



(a) Hospital Parameters from Cady et al. (1995)

(b) Hospital Parameters from Eachempati et al. (2004)

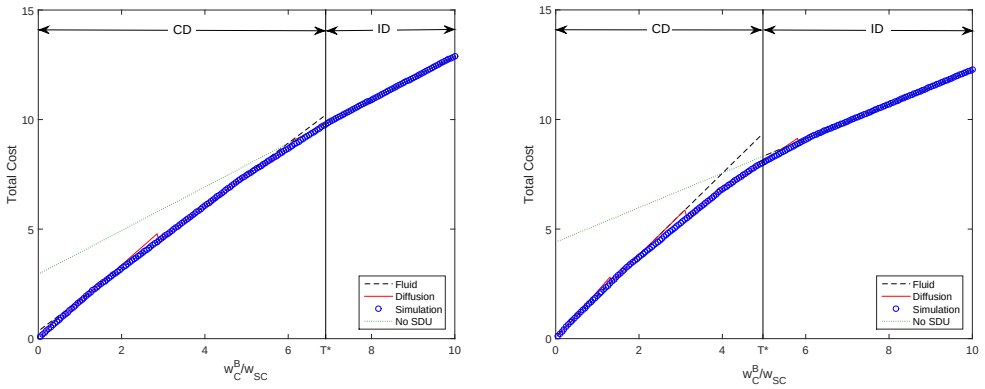
Figure 3 The Balking-Dominated Case: Optimal allocation of nurses to beds via fluid and diffusion analysis and exhaustive search. $N = 20$ **nurses.** $w_{SC} = 1$**,** $w_C^Q/\theta = 15$**.**

Figure EC.1 in the EC is the analog of Figure 3 in the Queue-dominated case. Again, the diffusion solution given by (8) and (13) is very close to the solution determined by using exhaustive search over simulations.

Though we see discrepancies in the number of beds in the ICU and SDU under the diffusion approximations, we find that the actual average cost incurred is quite close to optimal. Figure 4 compares the simulated costs under the diffusion and fluid solutions to the minimum cost achieved via exhaustive search. Figure EC.2 (in the EC) compares the same when split by expected queue length, balking and bumping rates. We also provide a benchmark of not having any SDU. The cost differences under the diffusion solutions are always less than 13% and are typically within less than 1% of optimal. On the other hand, the fluid solution can incur more than 4 times as much costs compared to optimal. Depending on the operating parameter regime, it may be sufficient to implement the fluid solution. In other instances, the diffusion solution can provide an important refinement to reduce costs.

We can also see that in the ID regime, it is certainly reasonable to put all nurses in the ICU. When the system is in the CD regime, it is very important to consider introducing an SDU; not having an SDU can result in costs which are an order of magnitude higher than that achieved via the optimal allocation.

5.2.1. Moderately Heavy Traffic Our fluid and diffusion analysis assumed an overloaded regime where the units were nearly always full. A number of hospitals strive for a target utilization of 85% and within New York, the average ICU occupancy level was 75% (Green 2003). We note that these utilization



(a) Hospital Parameters from Cady et al. (1995) (b) Hospital Parameters from Eachempati et al. (2004)

Figure 4 The Balking-Dominated Case: Average cost incurred under optimal allocation of nurses to beds via fluid and diffusion analysis and exhaustive search. $N = 20$ **nurses.** $w_{SC} = 1$**,** $w_C^Q/\theta = 15$**.**

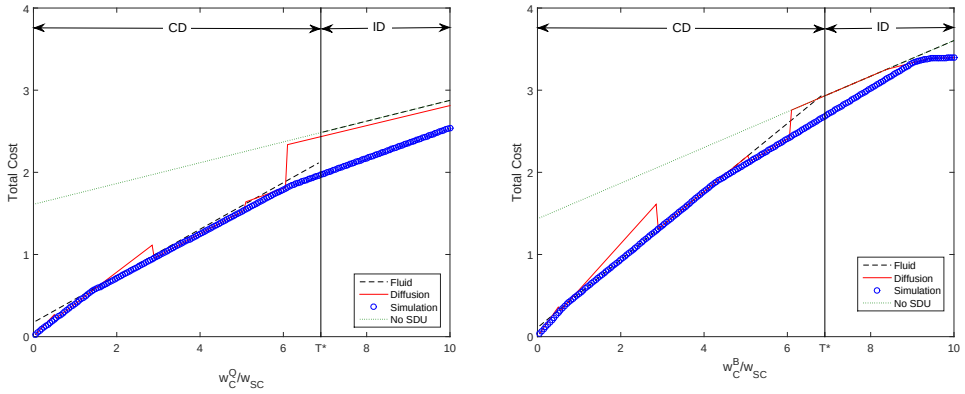
metrics are censored measures of the true demand due to adaptive techniques–such as balking, abandonment and bumping–which can divert arrivals and reduce length-of-stay. Still, there may be periods when the ICU is not in overload, so we also consider the quality of our analysis in a 'moderately' heavy traffic regime. The traffic load in this case is such that if all nurses are allocated to the ICU, the nominal load of patients is 85%, i.e. $\frac{\lambda(r_I\mu_C p + r_S\mu_{SC})}{N r_I r_S \mu_C \mu_{SC}} = .85$. While the optimal allocation of nurses changes slightly in this case, we see in Figure 5 that the diffusion and fluid solutions still perform reasonably well in terms of costs in the CD regime. We notice that in the ID regime, the asymptotic approach can result in poor performance. This is because, in this moderate traffic, the ICU is very far from operating in the overloaded regime and the quality of the approximations noticeably degrades. Still, we see that our solution always outperforms the simple benchmark of having no SDU.

6. Model Extensions

Thus far, the focus of this work has been on the model presented in Section 2. We now consider a number of extensions to our initial model which capture additional dynamics which can arise in various hospital settings. In particular, we explicitly consider readmissions, variants to the budget neutral nursing constraint, and time-varying arrivals.

6.1. Returns to Critical State

We start by considering a stylized model which incorporates patient readmissions. To streamline the discussion, we focus on the Queue-dominated case, so that $K^* = \infty$ and there is no balking. When a patient leaves the ICU-SDU system, there is some probability he will return to the Critical state (note that a patient can leave the ICU-SDU system, but still remain in the hospital in the general medical-surgical ward). The probability a patient will return to Critical state depends on how the patient left the system. We let p_C^A



(a) Queue-dominated ($w_C^B = 15$) (b) Balking-dominated ($w_C^Q/\theta = 15$)

Figure 5 **Average cost incurred under optimal allocation of nurses to beds via diffusion analysis and exhaustive search.** $N = 20$ **nurses. Moderate traffic:** $\frac{\lambda(r_I\mu_C p + r_S\mu_{SC})}{Nr_I r_S \mu_C \mu_{SC}} = .85$**. Hospital Parameters from Cady et al. (1995).** $w_{SC} = 1$**.**

and p_C^N denote the return probability for patients who abandoned or departed naturally as Critical patients, respectively. Similarly, p_{SC}^B and p_{SC}^N denote the return to Critical state probability if the patient is bumped as a Semi-critical patient or if the patient naturally completed service as a Semi-critical patient in the ICU or SDU. As return patients are typically worse off, we will assume that all return patients are served and they will not abandon and cannot balk or be bumped. Thus, the expected length of stay of a return patient, not including waiting time, is $E[LOS_R|Return] = \frac{1}{\mu_C} + \frac{p}{\mu_{SC}}$. Finally, the expected readmission load is then $p_R E[LOS_R|Return]$, where p_R denotes the return risk of the patient and depends on how the patient departs the system.

In the Electronic Companion, we formally introduce this model with patient returns. Additionally, we establish the stability condition of such a system. Similar to Chan et al. (2012), we find that minimizing the expected readmission load corresponds to maximizing throughput.

Proposition 5 *If the abandonment and bumping costs capture the* **increase in readmission load** *associated with these events, then the allocation of nurses which minimizes the average abandonment and bumping costs will also minimize the number of nurses necessary to stabilize the readmission queue.*

Now, we use simulation to compare the the quality of our nurse allocation derived from our original model when considering a model which incorporates readmissions. We assume the following readmission probabilities: $p_C^A = .10$, $p_{SC}^B = .05$, and $p_C^N = p_{SC}^N = .02$. We assume the time to return to Critical state is exponentially distributed with mean $1/\delta = 5$ days. We consider the nurse allocation for our original model in Section 2 which minimizes the return rate on the diffusion scale derived in Section 4. We then evaluate the performance of this solution via a simulation model that does have returns to Critical state to the solution achieved via an exhaustive search for the model with returns. We consider the case with $N = 20$ nurses.

	Original Model (without Returns)	Exhaustive Search via simulation	All nurses in ICU (No SDU)
Eachempati et al. (2004)	18.6%	18.8%	20.5%
Cady et al. (1995)	31.0%	31.4%	36.5%

Table 2 A system with returns to Critical state: Comparison of return rates for solution which ignores returns (Original Model) to solution established via exhaustive search.

Table 2 summarizes our simulation results for a system with returns. We can see that the number of returns achieved via our diffusion solution for a model *without* returns, but with cost appropriately defined as the increase in return risk due to abandonment and/or bumping, is very close to the minimum percentage of returns. As a benchmark, we see that when there is no SDU, the percentage of returns increases.

6.2. Relaxing the nursing constraint

Thus far, we have considered the ICU and SDU sizing decision under the assumption that the number of nurses must be held constant. This budget neutral constraint appears in a number of settings. However, it is conceivable that the joint ICU and SDU sizing decision may not have such a strict constraint on the number of nurses. For instance, a hospital may consider hiring M additional nurses and must determine whether to allocate them all to the ICU or SDU or split the nurses across both units. Alternatively, a hospital may not want to completely resize the units and may just want to consider 2 potential options.

Our analysis provides some insight into these other problem formulations. In particular, given an allocation a specific number of ICU and SDU beds, B_I and B_S, one can easily calculate the number of nurses N. Given the arrival rate λ at the hospital, one can use the analysis from Sections 3 and 4 to evaluate the operational parameter regime and assess the performance–in terms of balking rate, queue length, and bumping rate–of such a configuration. That is, our results are also useful for *performance analysis*.

6.3. Time-varying arrivals

In practice, hospitals tend to have arrival rates that are highly time variable (Green et al. 2006b, Armony et al. 2010), while the unit sizes remain fixed for a while. Accounting for this time variation when determining staffing levels in the Emergency Department (ED) can lead to much better provision of care (Green et al. 2006b, Yom-Tov and Mandelbaum 2013). As many ICU patients originate from the ED, the time-varying arrival rates to the hospital translate to time-varying arrival rates to the ICU. However, unlike the ED, the service times in the ICU are very long ($\sim$2-4 days as seen in Table 1) whereas the variation in arrival rates is on the order of hours. This difference in time scale suggests that it is not essential to capture time variation when establishing staffing levels in the ICU. For more discussion of this see Yom-Tov and Mandelbaum (2013) as well as Section 5.2 and Figure 13 in Chan et al. (2014b).

7. Conclusions and Discussion

Within the medical community, there is a lot of uncertainty on how to manage and size SDUs. In this work, we consider the optimal allocation of nurses for the inpatient units used to treat the hospitals most critical

patients: the ICU and SDU. In doing so, we provide insight into when and how the SDU can be useful in managing patient flow.

We propose a queueing model which allows us to examine how to optimally tradeoff flexibility and capacity given the costs associated with lack of access to ICU and/or SDU care. Via our fluid analysis, we identify two parameter regimes–the ICU-Driven and Capacity Driven regimes–which dictate the optimality of allocating a very small (including zero) or a substantial number of nurses to the SDU. Depending on the regime, only costs associated with Critical or Semi-Critical patients will be incurred, but not both. On the other hand, costs associated with both Critical and Semi-Critical patients will be incurred at the diffusion level. We leverage a state-space collapse result to evaluate and optimize the staffing allocation and balking threshold in the diffusion scale. We also find that in the ID regime, it can be optimal to have a non-zero balking threshold on the diffusion level, so that balking, queue, and bumping costs are all incurred. Numerically, we find that our analysis in these asymptotic regimes can be quite accurate, even as we relax some of our initial model assumptions.

In practice, there is high variation across hospitals as to whether it has an SDU and if so, how large the unit is in comparison to the ICU. On the surface, this variation could be attributed to the fact there is limited consensus in the medical community as to the management of SDUs. However, our analysis provides justification for this variation. The optimal size of an SDU is highly dependent on patient mix (including differences in service times and the likelihood of becoming a Semi-critical patient following ICU care), staffing requirements in the ICU versus SDU, as well as the relative cost of lack of access to care for a Critical versus Semi-critical patient. Because these factors are likely to vary substantially across different hospitals and geographic areas, it is reasonable–and highly desirable–that hospitals utilize and size SDUs in a heterogenous manner. One size does not fit all.

This work suggests several potential directions for future research. For instance, if a new hospital were being built, it would be useful to consider the staffing decision without the budget neutral constraint. In such a setting, a third tradeoff arises: staffing costs versus flexibility and capacity. Another direction would be to consider other patient flows through the SDU. In this work, we only consider SDU patients who originate in the ICU; however, some hospitals will admit patients into the SDU who have never visited the ICU. One could also consider different priority rules, so that in some cases a Critical patient will have to wait (and potentially abandon), even if there is a Semi-critical patient in the ICU which could be bumped. Finally, in this paper we have focused on sizing the ICU and SDU, while ignoring the size of the general wards. This is because the ICU is often considered the hospital bottleneck. An interesting direction for future research is to explicitly model the size and dynamics of the general ward along with the other two units.

Despite some of these limitations of our model, our work provides an important first step into addressing the substantial debate in the medical community as to if and how SDUs should be used. The prevailing sentiment amongst SDU supporters is that they are a cost effective way to provide care to Semi-critical

patients. This is true in some cases (CD regime). However, in the ID regime, we see that the need of the high priority patients outweighs the additional capacity generated by moving nurses to the SDU. Still, even in this regime, a *small* SDU can be beneficial in serving as a buffer between the ICU and the hospital wards. The insights from our work will be useful for hospital managers to assess the pros and cons of SDUs and whether one is warranted at their hospital. Indeed, we are currently working with a large academic hospital which treats an underserved population that recently opened a new SDU. This unit will only be used as a true Step-Down Unit, so that patients will only be admitted following ICU discharge. Upon learning of our findings, the Critical Care team reached out to us for help assessing the management of their new SDU. We are currently working with them to collect data in order to calibrate system parameters for their patient population. While we do not expect the hospital to directly implement the precise sizing and balking threshold decision our model recommends, we do expect to be able to assess i) whether a sizable SDU is warranted and ii) whether most Critical patients should wait or balk immediately upon arriving to a full ICU.

Acknowledgements

We are grateful to Feryal Erhun, Linda Green and Avi Mandelbaum for providing us with invaluable feedback throughout the various stages of this project. We also thank Alexej Proskynitopoulos for his research assistance with our numerical study.

References

Akan, M., B. Ata, T. Olsen. 2013. Congestion-based leadtime quotation for heterogeneous customers with convex-concave delay costs: Optimality of a cost-balancing policy based on convex hull functions. *Operations Research* **60**(6) 1505–1519.

Aloe, K., L. Raffaniello, M. Ryan, L. Williams. 2009. Creation of an Intermediate Respiratory Care Unit to Decrease Intensive Care Utilization. *Journal of Nursing Administration* **39**(11) 494–498.

Andradottir, S., H. Ayhan, H. Eser Kirkizlar. 2013. Flexible servers in tandem lines with setups. *Working paper, Georgia Institute of Technology* .

Armony, M., S. Israelit, A. Mandelbaum, Y.N. Marmor, Y. Tseytlin, G.B. Yom-tov. 2010. Patient flow in hospitals: A data-based queueing-science perspective. *Working Paper, Stern School of Business* .

Ata, B., B.L. Killaly, T.L Olsen, R.P. Parker. 2013. On hospice operations under medicare reimbursement policies. *Management Science* **59(**5) 1027–1044.

Ata, B., J. A. Van Mieghem. 2009. The Value of Partial Resource Pooling: Should a Service Network Be Integrated or Product-Focused? *Management Science* **55**(1) 115–131.

Baron, O., J. Milner. 2009. Staffing to maximize profit for call centers with alternate service-level agreements. *Operations Research* **57**(3) 685–700.

Bassamboo, A., R. S. Randhawa. 2010. On the accuracy of fluid models for capacity sizing in queueing systems with impatient customers. *Operations Research* **58** 1398–1413.

Bassamboo, A., R.S. Randhawa, J.A. Van Miegham. 2012. A Little Flexibility is All You Need: On the Asymptotic Value of Flexible Capacity in Parallel Queuing Systems. *Operations Research* **60**(6) 1423–1435.

Beck, M. 2011. Critical (Re)thinking: How ICUs are getting a much-needed makeover. *Wall Street Journal, March 28* .

Bell, S. L., R. J. Williams. 2001. Dynamic Scheduling of a System with Two Parallel Servers in Heavy Traffic with Resource Pooling: Asymptotic Optimality of a Threshold Policy. *Annals of Applied Probability* **11**(3) 608–649.

Best, T., B. Sandikci, D. Eisenstein, D. Meltzer. 2015. Managing hospital inpatient bed capacity through partitioning care into focused wings. *MSOM, to appear* .

Browne, S., W. Whitt. 1995. Piecewise-linear diffusion processes. Dshalalow, ed., *Advances in queueing: Theory, methods, and open problems*. CRC Press, Boca Raton, FL, 463–480.

Byrick, R.J., J.D. Power, J.O. Ycas, K.A. Brown. 1986. Impact of an intermediate care area on ICU utilization after cardiac surgery. *Critical care medicine* **14**(10) 869.

Cady, N., M. Mattes, S. Burton. 1995. Reducing Intensive Care Unit Length of Stay: A Stepdown Unit for First-Day Heart Surgery Patients. *Journal of Nursing Administration* **25**(12) 29–35.

Chalfin, D. B., S. Trzeciak, A. Likourezos, B. M. Baumann, R. P. Dellinger. 2007. Impact of delayed transfer of critically ill patients from the emergency department to the intensive care unit. *Critical Care Medicine* **35** 1477–1483.

Chan, C. W., V. F. Farias, N. Bambos, G. Escobar. 2012. Optimizing ICU Discharge Decisions with Patient Readmissions. *Operations Research* **60** 1323–1341.

Chan, C. W., V. F. Farias, G. Escobar. 2013. The Impact of Delays on Service Times in the Intensive Care Unit. *Working paper, Columbia Business School* .

Chan, C. W., L. V. Green, L. Lu, G. Escobar. 2014a. The role of a step-down unit in improving patient outcomes. *working paper, Columbia Business School* .

Chan, C.W., G. Yom-Tov, G. Escobar. 2014b. When to use Speedup: An Examination of Service Systems with Returns. *Operations Research* **62**(2) 462 – 482.

Chen, L. M., C. M. Martin, S. P. Keenan, W. J. Sibbald. 1998. Patients readmitted to the intensive care unit during the same hospitalization: clinical features and outcomes. *Critical Care Medicine* **26** 1834–1841.

Dai, J. G., T. Tezcan. 2008. Optimal Control of Parallel Server Systems with Many Servers in Heavy Traffic. *Queueing Systems* **59** 95–134.

de Véricourt, F., O.B. Jennings. 2008. Dimensioning large-scale membership services. *Operations Research* **56**(1) 173–187.

Durbin, C.G., R.F. Kopel. 1993. A Case-Control Study of Patients Readmitted to the Intensive Care Unit. *Critical Care Medicine* **21** 1547–1553.

Eachempati, S. R., L. J. Hydo, P. S. Barie. 2004. The effect of an intermediate care unit on the demographics and outcomes of a surgical intensive care unit population. *Archives of Surgery* **139**(3) 315–319.

Ethier, S.N., T.G. Kurtz. 1985. *Markov processes, characterization and convergence*. John Wiley & Sons.

Gans, N., G. Koole, A. Mandelbaum. 2003. Telephone Call Centers: Tutorial, Review, and Research Prospects. *Manufacturing & Service Operations Management* **5**(2) 79–141.

Garnett, O., A. Mandelbaum, M. Reiman. 2002. Designing a call center with impatient customers. *Manufacturing & Service Operations Management* **4**(3) 208–227.

Ghamami, S., A. R. Ward. 2012. Dynamic Scheduling of a Two-Server Parallel Server System with Complete Resource Pooling and Reneging in Heavy Traffic: Asymptotic Optimality of a Two-Threshold Policy. *Mathematics of Operations Research (to appear)* .

Green, L. 1985. A Queueing System with General-Use and Limited-Use Servers. *Operations Research* **33** 168–182.

Green, L. V. 2003. How many hospital beds? *Inquiry* **39**(4) 400–412.

Green, L. V., S. Savin, B. Wang. 2006a. Managing patient service in a diagnostic medical facility. *Operations Research* **54**(1) 11–25.

Green, L. V., J. Soares, J. F. Giglio, R. A. Green. 2006b. Using queuing theory to increase the effectiveness of emergency department provider staffing. *Academic Emergency Medicine* **13** 61–68.

Green, L.V. 2002. How many hospital beds? *Inquiry-The Journal Of Health Care Organization Provision And Financing* **39** 400–412.

Gurvich, I., W. Whitt. 2009a. Queue-and-Idleness-Ratio Controls in Many-Server Service Systems. *Mathematics of Operations Research* **34** 363–396.

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

Gurvich, I, W Whitt. 2010. Service-level differentiation in many-server service systems via queue-ratio routing. *Operations Research* **58**(2) 316–328.

Halfin, S., W. Whitt. 1981. Heavy-Traffic Limits for Queues with Many Exponential Servers. *Operations Research* **29** 567–588.

Halpern, N.A., S.M. Pastores. 2010. Critical care medicine in the United States 2000-2005: an analysis of bed numbers, occupancy rates, payer mix, and costs. *Crit Care Med* **38** 65–71.

Harding, A. D. 2009. What Can An Intermediate Care Unit Do For You? *Journal of Nursing Administration* **39**(1) 4–7.

Harrison, J.M., A. Zeevi. 2004. Dynamic scheduling of a multiclass queue in the halfin and whitt heavy traffic regime. *Operations Research* **52** 243–257.

Hopp, W.J., E. Tekin, M.P. Van Oyen. 2004. Benefits of skill chaining in serial production lines with cross-trained workers. *Managemet Science* **50**(1) 83–98.

Iravani, S.M.R., M.P. Van Oyen, K.T. Sims. 2005. Structural flexibility: A new perspective on the design of manufacturing and service operations. *Management Science* **51**(2) 151–166.

Jagerman, D. L. 1974. Some properties of the erlang loss function. *Bell Systems Tech. J.* **53** 525–551.

Joint Commission Resources. 2004. *Improving Care in the ICU*. Joint Commission on Accredidation of Healthcare Organizations.

Kc, D., C. Terwiesch. 2012. An econometric analysis of patient flows in the cardiac intensive care unit. *Manufacturing & Service Operations Management* **14**(1) 50–65.

Keenan, S. P., W. J. Sibbald, K. J. Inman, D. Massel. 1998. A Systematic Review of the Cost-Effectiveness of Non-cardiac Transitional Care Units. *Chest* **113** 172–177.

Kim, S-H, C. W. Chan, M. Olivares, G. Escobar. 2015. ICU Admission Control: An Empirical Study of Capacity Allocation and its Implication on Patient Outcomes. *Management Science* **61** 19–38.

Kirkizlar, H. Eser, S. Andradottir, H. Ayhan. 2013. Flexible servers in understaffed tandem lines. *POMS, to appear* .

Kocaga, Y. L., M. Armony, A. R. Ward. 2014. Staffing call centers with uncertain arrival rates and co-sourcing. *Production and Operations Management (to appear)* .

Kocaga, Y. L., A. R. Ward. 2010. Admission control for a multi-server queue with abandonment. *Queueing Systems* **65**(3) 275–323.

Kostami, V., A.R. Ward. 2009. Managing service systems with an offline waiting option and customer abandonment. *Manufacturing & Service Operations Management* **11**(4) 644–656.

Kwan, M.A. 2011. Acuity-adaptable nursing care: Exploring its place in designing the future patient room. *Health environments research & design* **5**(1) 77 – 93.

Loynes, R.M. 1963. The stability of a queue with non-independent interarrival and service times. *Proceedings of the Cambridge Philisophical Society* **58** 497–530.

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

Mandelbaum, A., S. Zeltyn. 2009. Staffing many-server queues with impatient customers: constraint satisfaction in call centers. *Operations Research* **57** 1189–1205.

Mason, J.E., B.T. Denton, N.D. Shah, S.A. Smith. 2014. Using electronic health records to monitor and improve adherence to medication. *working paper, University of Virgina* .

Mathews, K. S., E. F. Long. 2015. A conceptual framework for improving critical care patient flow and bed utilization. *Annals of the American Thoracic Society, to appear* .

Mills, A., N. T. Argon, S. Ziya. 2013. Resource-based patient prioritization in mass-casualty incidents. *MSOM* **15** 361–377.

Mills, A., N. T. Argon, S. Ziya. 2015. Dynamic distribution of casualties to medical facilities in the aftermath of a disaster. *working paper, Kelley School of Business, Indiana University* .

Pronovost, P.J., D.M. Needham, H. Waters, C.M. Birkmeyer, J.R. Calinawan, J.D. Birkmeyer, T. Dorman. 2004. Intensive care unit physician staffing: Financial modeling of the leapfrog standard*. *Critical care medicine* **32**(6) 1247–1253.

Reiman, M.I. 1984. Some diffusion approximations with state space collapse. F. Baccelli, G. Fayolle, eds., *Modelling and Performance Evaluation Methodology*. Springer-Verlag, 209–240.

Rubino, M., B. Ata. 2009. Dynamic control of a make-to-order parallel-server system with cancellations. *Operations Research* **57**(1) 94–108.

Ryckman, F.C., P.A. Yelton, A.M Anneken, P.E. Kiessling, P.J. Schoettker, U.R. Kotagal. 2009. Redesigning intensive care unit flow using variability management to improve access and safety. *Joint Commission journal on quality and patient safety / Joint Commission Resources* **35** 535–43.

Shmueli, A., C.L. Sprung, E.H. Kaplan. 2003. Optimizing admissions to an intensive care unit. *Health Care Management Science* **6**(3) 131–136.

Snow, N., K.T. Bergin, T.P Horrigan. 1985. Readmission of Patients to the Surgical Intensive Care Unit: Patient Profiles and Possibilities for Prevention. *Critical Care Medicine* **13** 961–985.

Tezcan, T., J.G. Dai. 2010. Dynamic Control of N-Systems with Many Servers: Asymptotic Optimality of a Static Priority Policy in Heavy Traffic. *Operations Research* **58** 94–110.

Tosteson, A., L. Goldman, I. S. Udvarhelyi, T. H. Lee. 1996. Cost-effectiveness of a coronary care unit versus an intermediate care unit for emergency department patients with chest pain. *Circulation* **94**(2) 143–150.

Tsitsiklis, J.N., K. Xu. 2012. On the power of (even a little) resource pooling. *Stochastic Systems* **2** 1–66.

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

Whitt, W. 2002. *Stochastic-Process limits: An Introduction to Stochastic Process Limits and their applications to Queues*. Spring-Verlag, New York.

Whitt, W. 2006. Fluid Models for Multiserver Queues with Abandonments. *Operations Research* **54** 37–54.

Yankovic, N., L. Green. 2011. Identifying Good Nursing Levels: A Queuing Approach. *Operations Research* **59** 942–955.

Yom-Tov, G., A. Mandelbaum. 2013. Erlang-r: A time-varying queue with reentrant customers, in support of healthcare staffing. *Manufacturing & Service Operations Management (to appear)* .

Zhang, Bo, H. Ayhan. 2013. Optimal admission control for tandem queues with loss. *IEEE Transactions on Automatic Control, to appear* .

Zimmerman, J.E., D.P. Wagner, W.A. Knaus, J.F. Williams, D. Kolakowski, E.A. Draper. 1995. The use of risk predictions to identify candidates for intermediate care units. *Chest* **108**(2) 490.

Electronic Companion
EC-1. Miscellaneous Proofs

PROOF OF PROPOSITION 1:

1. Suppose that $\limsup_{N\to\infty} \lambda\left(\frac{1}{r_I\mu_C}+\frac{p}{r_S\mu_{SC}}\right)\le N$. Note, this implies that the offered load in the ICU can be met: $\limsup_{N\to\infty}\frac{\lambda}{r_I\mu_C}\le N$. Consider the case where there is no balking, i.e. $K=\infty$. Then, the number of Critical patients in the ICU behaves like an $M/M/B_I+M$ queue. With traffic intensity $\frac{\lambda}{B_I\mu_C}\le 1$, we have that, by (Garnett et al. 2002, Theorem 4) with $\beta>-\infty$, the rate of abandonment is equal to $[\lambda-B_I\mu_C]^+ + o(N) = o(N)$.

As for the Semi-Critical patients, the arrival rate into this state is equal to $p\mu_C EZ_C$, where EZ_C stands for the expected steady-state number of ICU beds that are occupied by critical patients. The service rate is equal to $(B_S+B_I-EZ_C)\mu_{SC}$. By Little's law, $EZ_C=(\lambda-o(N))/\mu_C$, where the $o(N)$ term is contributed by the Critical patient abandonment rate. The bumping rate is hence equal to

$$[p\mu_C EZ_C-(B_S+B_I-EZ_C)\mu_{SC}]^+ = \mu_{SC}\left[\mu_T(\lambda+o(N))-(B_S+B_I)\right]^+ = o(N).$$

2. Suppose now that $\liminf_{N\to\infty}\lambda\left(\frac{1}{r_I\mu_C}+\frac{p}{r_S\mu_{SC}}\right)>N$. We let $1/\mu_T=\left(\frac{1}{\mu_C}+\frac{p}{\mu_{SC}}\right)$ be the mean amount of time a new patient should be treated while in the Critical and Semi-Critical states if the system has ample capacity. For any bed allocation (B_I,B_S), we let $\rho_C=\frac{\lambda}{B_I\mu_C}$ and $\rho_T=\frac{\lambda}{(B_I+B_S)\mu_T}$. In this case, we have that for any sequence of bed allocation (B_I,B_S), either $\liminf_{N\to\infty}\rho_C>1$, or $\liminf_{N\to\infty}\rho_T>1$, or both. If $\limsup_{N\to\infty}\rho_C>1$, then we have that the aggregated abandonment and balking rate is at least $\lambda-b_I\mu_C$, which is $O(N)$ (it could be less if Semi-Critical patients are occupying ICU beds, so that less than b_I beds are available to treat Critical patients). On the other hand, if $\limsup_{N\to\infty}\rho_C\le 1$, then by 1. the abandonment is $o(N)$. Therefore, the bumping rate is again equal to

$$[p\mu_C EZ_C-(B_S+B_I-EZ_C)\mu_{SC}]^+ = \mu_{SC}\left[\mu_T(\lambda+o(N))-(B_S+B_I)\right]^+ = O(N).$$

If neither of these cases applies, the argument works analogously when considering converging subsequences such that either $\lim_{N\to\infty}\rho_C>1$ or $\lim_{N\to\infty}\rho_C\le 1$.

□

PROOF OF THEOREM 1: Suppose that (3) holds in the limit. That is, assume that

$$\liminf_{N\to\infty}\frac{\lambda(r_I\mu_C p+r_S\mu_{SC})}{Nr_Ir_S\mu_C\mu_{SC}}>1. \tag{EC-1}$$

Additionally, assume that the system operates in the ID regime and that (6) and (7) hold. Let $\hat{U}^N:=\hat{Z}_C^N+\hat{Z}_{SC}^N$. And suppose that $\hat{U}^N(0)=0$. It is our goal to show that for any $\epsilon>0$,

$$P\left\{\inf_{0\le t\le 1}\hat{U}^N(t)<-\epsilon\right\}\to 0, \text{ as } N\to\infty.$$

The proof follows along the lines of Reiman (1984). Fix $\epsilon > 0$ and let

$$\tau_N = \inf\{t \geq 0;\ \hat{U}^N(t) < -\epsilon\} \text{ and } \tau'_N = \sup\{t \leq \tau_N;\ \hat{U}^N(t) \geq -\epsilon/2\}.$$

During $[\tau'_N, \tau_N]$ there are empty beds in either the ICU or SDU (or both), so no bumping will occur. In particular, during this interval

$$Z_C^N(t) + Z_{SC}^N(t) = Z_C^N(\tau'_N) + Z_{SC}^N(\tau'_N) + A^N(\tau'_N, t) + \Phi^N(\tau'_N, t) - D_C^N(\tau'_N, t) - D_{SC}^N(\tau'_N, t),$$

where, for $s < t$, $A^N(s,t)$ is the number of critical patients that arrived directly into the ICU (and did not wait in queue) during $(s,t]$, $\Phi^N(s,t)$ is the number of critical patients arrivals into the ICU from the queue in $(s,t]$. Also, $D_C^N(s,t]$ is the number of critical patients who have completed their stay in the ICU and did not switch to a semi-critical state during $(s,t]$. Finally, $D_{SC}^N(s,t)$ is the number of service completions of semi-critical patients in $(s,t]$. More specifically, let S_i, $i = 1,2,3$ be independent unit Poisson processes, then

$$\begin{aligned}
A^N(s,t) + \Phi^N(s,t) &= S_1\left(\int_s^t \lambda 1_{\{Z_C^N(r) < B_I\}} + \mu_C Z_C^N(r) 1_{\{Z_C^N(r) = B_I,\ Q > 0\}} \cdot dr\right) = (t-s)\cdot(\lambda + o(\lambda)),\\
D_C^N(s,t) &= S_2\left((1-p)\mu_C \int_s^t Z_C^N(r)\cdot dr\right) = (t-s)\cdot((1-p)\lambda + o(\lambda)),\\
D_{SC}^N &= S_3\left(\mu_{SC}\int_s^t Z_{SC}^N(r)\cdot dr\right) \leq S_3\left(\mu_{SC}\int_s^t \left(B_S^N + B_I^N - Z_C^N(r)\right)\cdot dr\right)\\
&= (t-s)\cdot\left(\frac{\mu_{SC} r_S}{r_I}\left(N r_I - \frac{\lambda}{\mu_C}\right) + o(\lambda)\right).
\end{aligned} \tag{EC-2}$$

Recall that the ICU is operating in the QED regime with respect to Critical patients; therefore, $\mu_C Z_C^N = \lambda + o(\lambda)$ and $B_I - Z_C^N = o(\lambda)$. Finally, we have:

$$\begin{aligned}
P\left\{\inf_{0\leq t\leq 1} \hat{U}^N(t) < -\epsilon\right\} &\leq P\left\{\inf_{0\leq s\leq t\leq 1} \frac{A^N(s,t) + \Phi^N(s,t) - D_C^N(s,t) - D_{SC}^N(s,t)}{\sqrt{\lambda}} < -\epsilon/2\right\}\\
&= P\left\{\inf_{0\leq s\leq t\leq 1} \frac{\frac{t-s}{\mu_C r_I}\cdot(\lambda\cdot(p r_I \mu_C + \mu_{SC} r_S) - \mu_{SC}\mu_C r_S r_I N) + o(\sqrt{\lambda})}{\sqrt{\lambda}} < -\epsilon/2\right\}\\
&\to 0, \text{ by (EC-1).}
\end{aligned}$$

□

PROOF OF PROPOSITION 4: Suppose that (EC-1) holds. Additionally, assume that the system operates in the CD regime and that (10) and (11) hold. Let $\hat{U}^N := \frac{Z_C^N + Q^N - (B_I^N + K^N)}{\sqrt{\lambda}}$, and suppose that $\hat{U}^N(0) = 0$. It is our goal to show that for any $\epsilon > 0$,

$$P\left\{\inf_{0\leq t\leq 1} \hat{U}^N(t) < -\epsilon\right\} \to 0, \text{ as } N \to \infty.$$

The proof follows along the lines of Reiman (1984). Fix $\epsilon > 0$ and let

$$\tau_N = \inf\{t \geq 0;\ \hat{U}^N(t) < -\epsilon\} \text{ and } \tau'_N = \sup\{t \leq \tau_N;\ \hat{U}^N(t) \geq -\epsilon/2\}.$$

During the interval $[\tau'_N, \tau_N]$, we have that $Z_C^N + Q^N < B_I^N + K^N$, so no balking would occur. In particular, during this interval

$$Z_C^N(t) + Q^N(t) = Z_C^N(\tau'_N) + Q^N(\tau'_N) + A^N(\tau'_N, t) - D_C^N(\tau'_N, t) - \Phi^N(\tau'_N, t),$$

where, for $s < t$, $A^N(s,t)$ is the number of critical patients that arrived to the system during $(s,t]$, $D_C^N(s,t]$ is the number of critical patients who have completed their stay in the ICU and either switched to a semi-critical state during $(s,t]$ or not. Finally, $\Phi^N(s,t)$ is the number of abandonment from the queue in $(s,t]$. More specifically, let S_i, $i = 1,2,3$ be independent unit Poisson processes, and let $\tau'_N \le s < t \le \tau_N$. Then

$$\begin{aligned}
A^N(s,t) &= S_1\left(\int_s^t \lambda 1_{\{Q^N(r) < K^N\}} \cdot dr\right) = (t-s)\cdot(\lambda + o(\lambda)),\\
D_C^N(s,t) &= S_2\left(\mu_C \int_s^t Z_C^N(r)\cdot dr\right) \le S_2\left(\mu_C B_I(t-s)\right) = (t-s)\cdot(\gamma\lambda + o(\lambda)),\\
\Phi^N(s,t) &= S_3\left(\theta \int_s^t Q^N(r)\cdot dr\right) \le S_3\left(\theta K^N(t-s)\right)\\
&\le S_3\left((1-\gamma)(1-\eta/2)(t-s) + o(\lambda)\right) = (t-s)\cdot\left((1-\gamma)(1-\eta/2) + o(\lambda)\right).
\end{aligned} \tag{EC-3}$$

Finally, we have:

$$\begin{aligned}
P\left\{\inf_{0\le t\le 1} \hat{U}^N(t) < -\epsilon\right\} &\le P\left\{\inf_{0\le s\le t\le 1} \tfrac{A^N(s,t) - D_C^N(s,t) - \Phi^N(s,t)}{\sqrt{\lambda}} < -\epsilon/2\right\}\\
&= P\left\{\inf_{0\le s\le t\le 1} \tfrac{(t-s)\cdot\lambda\cdot(1-\gamma)\eta/2 + o(\lambda)}{\sqrt{\lambda}} < -\epsilon/2\right\}\\
&= P\left\{\inf_{0\le s\le t\le 1} (t-s)\cdot\sqrt{\lambda}\cdot(1-\gamma)\,\eta/2 + o(\sqrt{\lambda}) < -\epsilon/2\right\}\\
&\to 0, \quad \text{by (EC-1)}.
\end{aligned}$$

□

PROOF OF COROLLARY 2: By the fluid analysis we have that $K^{*N} = o(N)$. Therefore, K^{*N} satisfies (12) with $\eta = 1$. By Corollary 1, the number of ICU beds available for semi-critical patients is $o(\sqrt{N})$ and therefore, the bumping cost is independent of the threshold level K^N (up to $o(\sqrt{N})$). It is therefore sufficient to focus on the queue and balking costs. As a function of the threshold level K^N we have that, by Proposition 4, the total queue plus balking cost rate is equal to

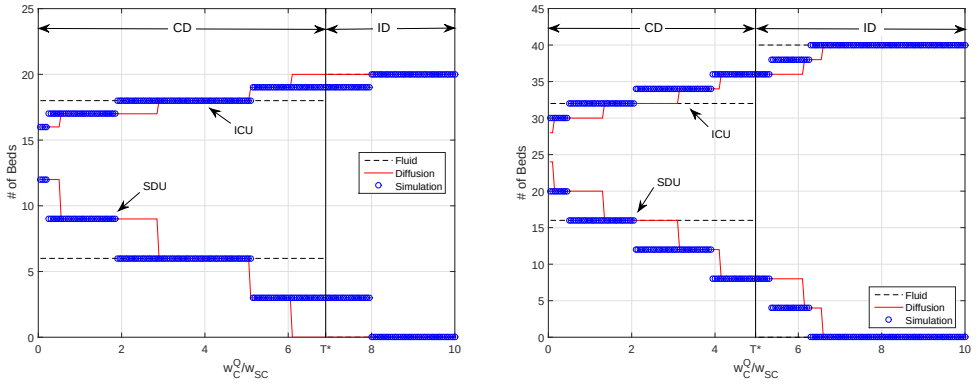
$$w_C^Q K^N + w_C^B \cdot \left(\lambda - \mu_C B_I - \theta K^N\right) + o(\sqrt{N}) = \theta K^N \cdot \left(w_C^Q/\theta - w_C^B\right) + w_C^B \cdot (\lambda - \mu_C B_I) + o(\sqrt{N}).$$

Under the balking-dominated case, the cost above is minimized by $K^N = o(\sqrt{N})$. □

EC-2. Additional Numerics

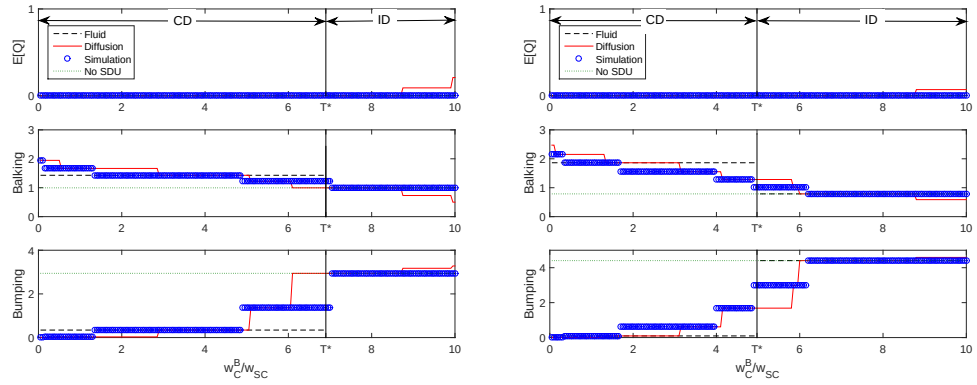
Figures EC.1 and EC.2 are supplemental to the numerical analysis of Section 5.2. Figure EC.1 is a parallel of Figure 3 for the Queue-Dominated case. The qualitative results are similar. Because the balking cost,

$w_C^B = 15$, is larger than the queue costs, $w_C^Q/\theta \in [0, 10]$, the balking threshold is $K^* = \infty$ for all solutions in this case. Figure EC.2 shows a breakdown of the costs in Figure 4, into expected queue lengths, as well as balking and bumping rates. Since in this scenario the Balking-dominated case applies, the expected queue lengths are, as expected equal to 0, except for high values of the balking cost w_C^B. As expected, the balking (bumping) rate is decreasing (increasing) in the cost ratio w_C^B/w_{SC}.



(a) Hospital Parameters from Cady et al. (1995) (b) Hospital Parameters from Eachempati et al. (2004)

Figure EC.1 The Queue-Dominated Case: Optimal allocation of nurses to beds via fluid and diffusion analysis and exhaustive search. $N = 20$ **nurses.** $w_{SC} = 1$**,** $w_C^Q/\theta = 15$**.**



(a) Hospital Parameters from Cady et al. (1995) (b) Hospital Parameters from Eachempati et al. (2004)

Figure EC.2 Performance measures: Average queue length, and balking and bumping rates (in # patients per day) under optimal allocation of nurses to beds via fluid and diffusion analysis and exhaustive search. $N = 20$ **nurses.** $w^{SC} = 1$**,** $w_C^Q/\theta = 15$**.**

EC-2.1. Balking Threshold: Balking-dominated case, ID regime

In Section 4.1.3, we found that in the Balking-dominated case ($w_C^B < w_C^Q/\theta$), ID regime ($w_C/w_{SC} > T^*$), the diffusion solution may have a non-zero optimal balking threshold. Our previous simulations suggest this is most likely to occur when w_C^Q/θ and w_C^B are close in magnitude.

To examine the impact of carefully optimizing the balking threshold, we consider the optimal (scaled) balking threshold and bed allocation, k^* and m^*, for the diffusion cost in Eqn. (9) when the queue costs are 1% larger than the balking costs: $w_C^Q/\theta = 1.01 \times w_C^B$. We then compare the cost incurred when using the optimal bed allocation, $\hat{m}^*_{\{k=0\}}$, when the balking threshold is fixed at 0, which is the threshold the fluid solution suggests. We consider both the diffusion cost and the unscaled and un-centered cost (i.e. $\sqrt{\lambda}C(m,k) + w_{SC}\left(\lambda p + \frac{r_S}{r_I}\left(Nr_I - \frac{\lambda}{\mu_C}\right)\mu_{SC}\right)$). Table 3 summarizes these findings.

	Scaled & Centered		Unscaled & Un-centered	
w_C^B	k^*	$\frac{C(\hat{m}^*_{\{k=0\}},0)}{C(m^*,k^*)}$	$K^* = k^*\sqrt{N}$	$\frac{C(\bar{B}^*_{I,\{k=0\}},0)}{C(B^*_I,K^*)}$
5.0	0.5	1.938	3	1.042
5.2	1.5	1.756	7	1.097
5.4	2.2	1.639	10	1.111
5.6	3.0	1.566	14	1.116
5.8	3.7	1.516	17	1.118
6.0	3.7	1.479	17	1.119

Table 3 Balking-dominated case, ID regime: Optimal balking threshold and suboptimal cost ratio for having no queue ($K = 0$). The scaled and centered results come from the diffusion solution in Eqn. 9. The unscaled and un-centered results transforms the diffusion costs to the case where $N = 20$ nurses. Hospital Parameters given by Eachempati et al. (2004). $w_{SC} = 1$, $w_C^Q/\theta = 1.01 \times w_C^B$.

We see that as w_C^B increases, so does the balking threshold, K^*. This makes sense as the absolute difference between the balking and queue costs are increasing (the relative difference is fixed at 1%), making it more desirable to have a (small) queue. We also see that simply fixing the balking threshold at 0 and optimizing the bed allocation can result in very poor performance. When $w_C^B = 5$, the resulting cost is over 93% higher than the optimal cost in the diffusion scale. While we find that in the Balking-dominated case, ID regime, it is important to calculate the balking threshold accurately as not doing so can have significant cost implications on the diffusion scale, we emphasize these are second order effects. In particular, when considering the unscaled and un-centered costs, the cost difference between optimizing the balking threshold versus fixing it at 0 is 4.2%-11.9%.

EC-3. A System with Returns to Critical State

We now consider a stylized model which explicitly accounts for patient returns to the Critical state. Note that throughout this discussion, we assume that $w_C^Q/\theta < w_C^B$, so that no balking occurs. For simplicity, we also assume that $w_C^H = 0$, so that the queue cost includes only the abandonment cost. We consider the following setup:

1. N nurses are reserved to treat first-time arrivals. These nurses can be allocated amongst B_I ICU and B_S SDU beds as desired. Any reference to system state will be understood to correspond to the number of *first-time* Critical and Semi-critical patients.

2. Without loss of generality, patients who depart naturally from the ICU or SDU will not return as Critical patients.

3. A first-time Critical patient who abandons from the ICU queue returns for to the Critical state with probability p_C^A, and has 'readmission' ICU LOS which is exponentially distributed with mean L_C^A. We let $w_C^A = p_C^A L_C^A$.

4. A first-time Semi-critical patient who is bumped from the ICU returns the Critical state with probability p_{SC}^B, and has readmission ICU LOS which is exponentially distributed with mean L_{SC}^B. We let $w_{SC}^B = p_{SC}^B L_{SC}^B$.

5. The return queue is served First-Come-First-Serve by C beds. Return patients are treated in the ICU until they are stable enough to be transferred to the Ward, i.e. they do not go through the SDU. Return patients will not abandon or balk from the return queue, nor can they be bumped from the ICU.

In practice, readmitted patients tend to be much sicker, with higher mortality rates and longer LOS (Snow et al. 1985, Durbin and Kopel 1993, Chen et al. 1998). Thus, it is desirable to provide high quality care for these return patients, which we capture by requiring they are treated in the ICU and cannot abandon, balk, or be bumped.

The total number of ICU beds in this setting is $C + B_I$. Given the N nurses to treat first-time arrivals, our goal is to determine the allocation of nurses to the ICU and SDU (B_I and B_S) such that we minimize the number of nurses, C/r_I, required to staff the readmission queue so that the queue remains stable. That is, if we let $\{W_n\}$ denote the waiting time the n^{th} readmitted patient experiences, we require that for any subsequence of $\{W_n\}$ there exists a sub-subsequence which converges to a random variable which is finite almost surely.

We start by examining the stability condition of the return queue. Let $\{\sigma_n, T_n\}$ denote the service requirement and interarrival time for readmitted patient n under some allocation of nurses between the ICU and SDU. Then the stability condition stems from a classical result of Loynes (1963), which requires that $E[\sigma_0]/E[T_0] < C$ for the readmission queue to be stable. We let π denote the steady-state distribution of the first-time patients, where the state is denotes by $S = (Q, Z_C, Z_{SC})$. For notational compactness, we suppress the dependence of this distribution on the nurse allocation. Relating the stability condition to our original problem setting of Section 2 we have:

Lemma 1 *The return queue is stable if and only if:*

$$\limsup_{T\to\infty} \frac{1}{T} \int_0^T \left[w_C^A \theta Q(t) + w_{SC}^B p \mu_C [B_I \wedge Z_C(t)] \psi(Q(t), Z_C(t), Z_{SC}(t)) \right] dt < C \qquad \text{(EC-4)}$$

PROOF: To start, we denote by a_n and b_n as indicator variables which equal 1 if patient n, who arrives at time t_n, abandons or is bumped, respectively. It is easy to see that the condition in (EC-4) is equivalent to:

$$\limsup_{T\to\infty} \frac{1}{T}\left[\sum_{n=0}^{\infty}\left[w_C^A a_n + w_{SC}^B b_n\right] 1_{\{t_n\le T\}}\right] < C$$

We start by focusing on first-time patients, and let $v = \lambda + Q_{\max}\theta + B_I \max\{\mu_C, \mu_{SC}\} + B_S\mu_{SC}$ be the maximum possible transition rate. We can determine the probability that the next event is a Critical patient abandonment or a Semi-critical patient bumping:

$$P(\text{Abandonment or Bumping}) = \sum_S \pi_S \frac{\left[\theta Q + \lambda 1_{\{Z_C = B_I + B_S - Z_{SC}, Z_{SC} > B_S\}}\right]}{v} \tag{EC-5}$$

If an abandonment (bumping) occurs, the patient's readmission ICU LOS is L_C^R (L_{SC}^R) with probability p_C^A (p_{SC}^B) and 0 otherwise; that is, we formally assume that *all* the abandoning and bumped patients are readmitted, but some of them have an ICU LOS of 0. The interarrival time of events is exponentially distributed with rate v. Additionally, the number of events until an abandonment or bumping is Geometrically distributed with mean $1/P(\text{Abandonment or Bumping})$. Thus, the interarrival time of readmitted patients is:

$$E[T_0] = \frac{v}{\sum_S \pi_S \left[\theta Q + \lambda 1_{\{Z_C = B_I + B_S - Z_{SC}, Z_{SC} > B_S\}}\right]} \tag{EC-6}$$

Finally, the expected service requirement of readmitted patients is:

$$E[\sigma_0] = \frac{\sum_S \pi_S \left[\frac{\theta Q}{v} w_C + \frac{\lambda 1_{\{Z_C = B_I + B_S - Z_{SC}, Z_{SC} > B_S\}}}{v} w_{SC}\right]}{\sum_S \pi_S \left[\theta Q + \lambda 1_{\{Z_C = B_I + B_S - Z_{SC}, Z_{SC} > B_S\}}\right]} \tag{EC-7}$$

Combining equations (EC-6) and (EC-7) gives the desired stability condition.

$$\begin{aligned}\frac{E[\sigma_0]}{E[T_0]} &= \sum_S \pi_S \left[\theta Q w_C + \lambda 1_{\{Z_C = B_I + B_S - Z_{SC}, Z_{SC} > B_S\}} w_{SC}\right] \\ &= \limsup_{T\to\infty} \frac{1}{T}\left[\sum_{n=0}^{\infty}\left[w_C a_n + w_{SC} b_n\right] 1_{\{t_n\le T\}}\right] < C \end{aligned} \tag{EC-8}$$

□

We can see that given an allocation of nurses, the readmission queue is stable when there are enough beds C to serve the readmission load. By specifying the costs of abandonment and bumping to be the readmission load associated with these events, the stability condition is to have enough beds C such that it is greater than the optimal average abandonment and bumping costs. Thus, to minimize the number of beds (and, subsequently, nurses) necessary to stabilize the readmission queue, the N nurses dedicated to first-time patients should be allocated such that the average abandonment plus bumping cost is minimized, as captured in equation (2).

Note that Proposition 5 follows directly from Lemma 1.