Critical Care Capacity Management: Understanding the role of 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, what size it should be, and what are the main drivers influencing these decisions. 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. 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. Moreover, we identify two parameters which play a prominent role in the SDU sizing decision: p, which captures the demand for SDU beds, and ν, which captures the supply gains by moving nurses to the SDU. The insights from our work provide rigorous justification for the variation in SDU use seen in practice as well as highlight which factors should be considered when making such sizing decisions for critical care.

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 identify the main drivers which dictate how these units should be sized.

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 a delicate tradeoff between overall capacity gains (SDU) for all critical patient severities versus maintaining more capacity for the most severely ill patients (ICU).

In theory, having a single large unit where the level of care of each bed can be dynamically flexed up or down would be more desirable than fixing the nurse allocation, as discussed in a recent opinion/survey article (Vincent and Rubenfeld 2015). However, the authors admit there is no evidence that such a solution is better than separate units and implementing this large unit may be practically infeasible at many institutions. First, all the beds would need to be equipped and legally certified to provide critical care. If some of these beds are only rarely, if ever, used in a critical care capacity, this would incur unnecessary overhead costs. Second, not having a dedicated unit for semi-critical patients will likely result in critical patients receiving priority in bed assignment over semi-critical patients in the large common unit. This would potentially lead to higher than desirable levels of off-placement of semi-critical patients. Third, nursing staff have been very reluctant to adopt such a solution as they prefer to have a level of predictability during their shift, which changing the level of care provided would not allow. Due to the high stress required to provide critical care, ICU nurses have the highest turnover. Implementing flexible staffing would increase dissatisfaction which could result in higher turnover and more medical errors (Strachota et al. 2003). 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 (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.

While physical space, beds, or specialized equipment could be the constraining resource, in many cases, nurses staffing is the bottleneck. For example, in California, despite availability of more physical beds, only 75% of adult ICU beds are staffed (State of California Office of Statewide Health Planning & Development 2010-2011). Thus, our primary focus will be on how to allocate nurses between the ICU and SDU. Many hospitals use critical-care nurses to staff the SDU (e.g. Eachempati et al. (2004), Harding (2009)) in order to ensure that the nurses are capable of dealing with any complications which could arise in the unit. However, if a hospital (e.g. Aloe et al. (2009)) elected to use medical-surgical nurses in their SDU, allocating more beds to the SDU would have an additional benefit (over capacity gains) of lower staffing costs. Because of strict nurse-to-patient ratios, the number of nurses fully dictates the number of beds and we will use the two terms (nurse versus bed allocation) interchangeably.

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 our analytic model on the case where the SDU is a true ‘Step Down Unit’ and patients are admitted only after being discharged from the ICU. We will then use simulation to examine how our insights translate to more complex patient flow patterns.

One could consider utilizing a simulation study (e.g. Mathews and Long (2015)) to exhaustively search over different possible ICU and SDU bed allocations. While this can provide useful prescriptive insight for a specific hospital setting, simulation studies can obscure the type of insights made possible via analysis of an analytic model, which is what we focus on in this work. In identifying important parameters which drive the bed allocation and balking threshold decisions, we also provide an initial framwork for hospital administrators to think about collecting data when making sizing decisions for their own institution. Another approach one could consider is to utilize Dynamic Programming (e.g. Best et al. (2015)). Given the complexity of bed allocation problem, Best et al. (2015) rely on heuristic solutions–this likely would also be required in our setting given the additional patient flow dynamics we incorporate into our model. As such, we turn to a queueing model approach and rely on fluid and diffusion approximations to gain insights.

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.

The hospital's 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 find that even under the optimal SDU sizing decisions, the number of beds allocated to the SDU is likely to be highly varied 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.

- We identify main drivers which influence the joint sizing decision of ICUs and SDUs. In particular, we find two parameters which arise in our first and second order analysis of the optimal nurse allocation between ICUs and SDUs. The first factor relates to the demand of SDUs, as captured by the proportion of critical patients who become semi-critical, p. The second factor related to the supply of capacity, as captured by the ratio between effective ICU capacity and effective SDU capacity, ν. These two factors arise additively, suggesting they both have equal importance in influencing the ICU and SDU sizing decision.

We find that optimizing the balking threshold is a second (or higher) order factor. As such, the tradeoff between balking and waiting is a second order effect, while the tradeoff between capacity for critical patients (ICU) and overall capacity (SDU) is a first order factor.

- Via simulation analysis, we examine whether our insights translate to a more complex model of patient flow. We find that the solutions obtained from our first and second order approximations result in good outcomes compared to an exhaustive search, despite our analytic model not incorporating all patient dynamics. This suggests that the main drivers of the unit sizing decision are robust to model specifications.

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 length-of-stay (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 (2014)). Most of the prior work focuses on a single unit and have not considered the impact of a step-down unit. 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 approach to consider how SDU bed availability impacts patient outcomes. 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. 2012, Zhang and Ayhan 2013, Kirkizlar et al. 2012). 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

Patient flows through the ICU and SDU can be very complex, so we start by focusing on a streamlined model in order to allow for tractability and to highlight the main factors which influence the optimal sizing of these units. In Section 5, we use simulation to examine whether our insights derived from our analytic model extend to more general patient flow dynamics.

Similar to Mathews and Long (2015), we consider two possible health states for each patient: **Critical** or **Semi-critical**. 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 in the critical state, he will become a semi-critical patient with probability p; with probability $1-p$, he leaves the ICU/SDU system. Practically, this can correspond to a number of different situations, such as the patient being transferred to 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 the recovery pattern for all patients of a single type is homogenous and 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 the case where nursing costs are the bottleneck, so we must determine how to allocate 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. In such instances, the costs for each nurse is invariable to the unit she is assigned. However, if nurses without critical care credentials are used in the SDU, the costs for SDU nurses would be lower than that for ICU nurses. 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. Thus, the nurse allocation and bed capacity decisions are interchangeable. We assume that 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. In Appendix EC-2, we consider the case where physical beds are the bottleneck and find that many of our insights also carry over.

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.

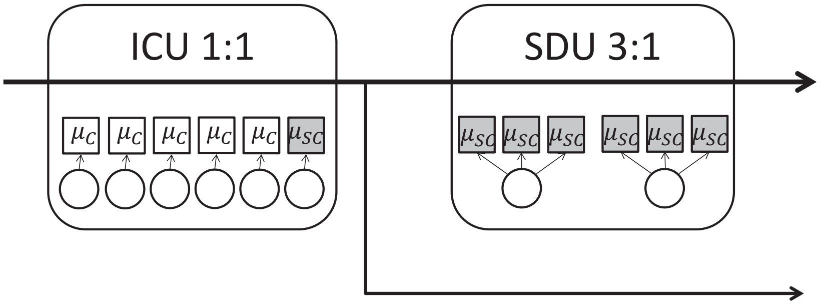


Figure 1 **Nurses are depicted as circles, patients are depicted at squares. Critical patients are served in the ICU. A critical patient may become a semi-critical patient upon finishing service in the ICU. semi-critical patients are depicted in gray and are served in the SDU or ICU. One semi-critical 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. 2016), 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. The '?' in the figure represents the assignment decision for the semi-critical patient.

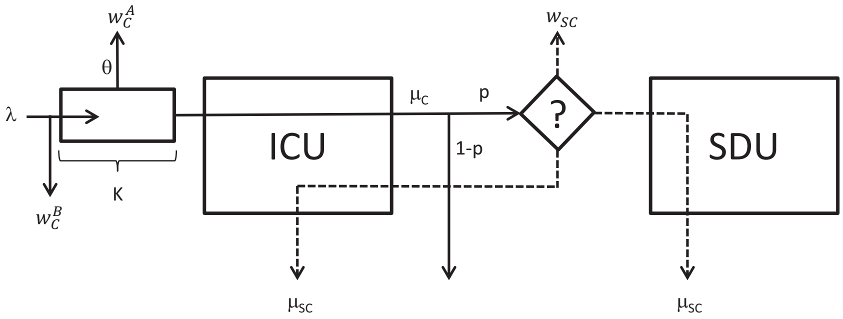


Figure 2 **ICU-SDU queueing model: The '?' represents the assignment decision of a semi-critical patient. Solid lines depict critical patient flows while dotted lines depict semi-critical patient flows.**

Our objective is to minimize the long run average balking, holding, abandonment, and bumping costs. These costs capture the impact of lack of access to care. 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 $\xi = 1$; if $Q(t) < K$, the patient enters the queue and $\xi = 0$. $\psi(Q(t), Z_C(t), Z_{SC}(t)) : \mathbb{Z}_+^3 \to \{0,1\}$ 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 balking threshold, K, as well as to specify the number of ICU and SDU beds in order to minimize the following cost function:

Optimization Problem 1

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

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. The bumping costs depend on the decision epochs when a semi-critical can be bumped: 1) when a critical patient becomes semi-critical, which occurs at rate $p\mu_C[B_I \wedge Z_C(t)]$ and 2) when a new critical patient arrives.*

In this work, we examine a stylized model of patient flows through 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. We will see this is a main driver in effectively managing ICUs and SDUs. In analyzing the patient flows described in this section, we can gain many insights into the role of the SDU and, in Section 5, we find that they extend to a more general model of patient flows.

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.

2.1. Cost parameters

It is reasonable to assume the optimal policy will depend on the different cost parameters: w_C^B, w_C^Q, and w_{SC}. Our formulation allows for *any* quality metric–it could capture clinical costs such as the net decrease in quality-adjusted life years (QALYs) or financial costs, such as loss in revenue due to not treating a patient in the ICU and/or the differences in reimbursement rates depending on where patients are treated. We now discuss a number of clinically relevant costs, which hospitals are likely to consider when making decisions surrounding ICUs and SDUs.

Mortality Risk: A natural cost metric is mortality. Specifically, there is some risk of death associated with each patient, even if the patient follows the ‘desired’ care pathway. However, if a Critical patient is unable to get ICU care and must wait, possibly so long he eventually abandons the queue, or is sent to another hospital, then it is reasonable to consider how this may impact the patient’s nominal mortality risk. Similarly, bumping a Semi-critical patient out of the SDU may increase the likelihood of death. In these cases, w_C^Q, w_C^B and w_{SC} could capture the *increase in mortality risk* due to waiting/abandonment, balking or bumping, respectively. Then, solving the optimization problem in (2) would correspond to selecting the

ICU and SDU sizes which would minimize the mortality rate of Critical and Semi-critical patients. In some practical settings, this cost metric may be too crude to be of value as access to care is typically granted for patients whose mortality risk would be significantly increased. Thus, we also consider other clinical measures of interest.

Readmission Risk: Another measure the medical community has focused on is patient readmissions, and more specifically, the probability of readmission. This cost metric has clear clinical implications as readmitted patients tend to be worse off (Durbin and Kopel 1993). It also has operational implications as readmitted patients will utilize ICU and SDU beds, which could have been used for new patients.

For each of these clinical measures, the cost parameters w_C^Q, w_C^B and w_{SC} would correspond to the increase in mortality risk or readmission risk due to waiting/abandonment, balking, or bumping. Then, solving the optimization problem in (2) would correspond to selecting the ICU and SDU sizes which would minimize the number of corresponding adverse patient outcomes. While a hospital administrator may wish to focus on one clinical outcome, one could also consider a weighted sum and/or other potential cost measures.

3. Balancing capacity needs with capacity gains

We begin our analysis via a fluid modeling approach to examine the optimal allocation of nurses and balking threshold given the balking, queue length, and bumping cost parameters. We find that the optimal allocation of nurses can be characterized by a well-defined threshold which captures the balance between capacity needs and capacity gains. 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 of order that is less than N. This way, we can focus on the main drivers of the balking threshold and nurse allocation. We begin by defining our fluid scaling. For notational compactness, we omit the indexing of λ by N. Let $\bar{\lambda} := \lambda/N, b_i := B_i/N, k := K/N$ for $i = I, S$ and note that by (1),

$$\frac{b_I}{r_I} + \frac{b_S}{r_S} \leq 1. \tag{3}$$

We say a function $f(x) := o(x)$ if $f(x)/x \to 0$ as $x \to \infty$ and $f(x) := \mathcal{O}(x)$ if $f(x)/x \leq c > 0$ as $x \to \infty$. The following proposition provides conditions such that the fluid costs are non-zero.

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

2. *Otherwise, if* $\frac{\lambda}{r_S\mu_{SC}}\left(p + \frac{r_S\mu_{SC}}{r_I\mu_C}\right) > N$, *then for any feasible bed allocation and balking threshold the total cost rate in Eqn.* (2) *is at least* $\mathcal{O}(N)$.

It is easy to see that in scenario 1, setting $K = \infty$, $B_I = \lambda/\mu_C$ and $B_s = \lambda p/\mu_{SC}$ will result in zero fluid costs. As such, in order to focus on the more interesting cases of non-zero costs, we consider the case where the following assumption is satisfied:

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

$$\frac{\lambda}{r_S\mu_{SC}}\left(p+\frac{r_S\mu_{SC}}{r_I\mu_C}\right)>N. \tag{4}$$

The following proposition helps simplify the fluid analysis substantially.

Proposition 2 *Under Assumption 2 and under any optimal bed allocation, we have that neither unit is underloaded. That is, we have that if (B_I,B_S) is optimal, then $B_I\leq\lambda/\mu_C+o(N)$ and $B_S\leq pB_I\mu_C/\mu_{SC}+o(N)$.*

The proof of the proposition follows simply by observing that if either of the units is underloaded, one could strictly improve the cost by transferring nurses to the other unit.

Corollary 1 *Under Assumption 2, the number of ICU beds occupied by semi-critical patients under the optimal allocation is $o(N)$.*

The implications of this corollary is that the interaction between the two units is minimal at the fluid scale in the sense that the patient types are effectively treated as separate units, and the system reduces to two queues in tandem with zero buffer in front of the second queue.

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 fluid-level 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. (2015) we have that, in this case $k^*=\infty$.
- **Balking-Dominated Case ($w_C^Q/\theta>w_C^B$):** We let $\bar{q}_{\max}\triangleq(\bar{\lambda}-\mu_C b_I)/\theta\geq 0$ denote the maximum queue length on the fluid scale if balking were not allowed. The non-negativity of $\bar{q}_{\max}$ follows from Proposition 2. Due to the overloaded assumption and the priority given to critical patients, for any fixed $k\leq\bar{q}_{\max}$, the queue length will be equal to k. If $k>\bar{q}_{\max}$, then the queue length is equal to $\bar{q}_{\max}$. Then the corresponding queue length cost incurred is $w_C^Q\min\{k,\bar{q}_{\max}\}$ and the balking cost is $(\bar{\lambda}-b_I\mu_C-\theta\min\{k,\bar{q}_{\max}\})^+w_C^B$. 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 (recall Corollary 1). We determine threshold k^*, which minimizes the cost function $\min_{0\leq k\leq\bar{q}_{\max}}\left\{(w_C^Q-\theta w_C^B)k+w_C^B(\bar{\lambda}-b_I\mu_C)\right\}$. 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 3 *In the fluid model, under the overloaded regime, the optimal balking threshold is given as:*

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

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, by Corollary 1, $b_I^* \le \bar{\lambda}/\mu_C$. 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 regimes, 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 (3) is satisfied as an equality under the optimal allocation, we can specify our fluid objective in terms of b_I.

Optimization Problem 2 (Fluid Cost) *Our goal is thus to determine,* $0 \le b_I \le \left(r_I \wedge \frac{\bar{\lambda}}{\mu_C}\right)$ *and* $0 \le b_S \le r_S$*, the allocation of nurses to ICU and SDU beds, respectively, so as to minimize the cost function:*

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

We can solve the preceding optimization problem to determine how to allocate nurses between the ICU and SDU. Note that the objective in Equation (5) is piecewise linear in b_I. As such, the minimization is obtained at the boundary points, which essentially proves Proposition 4 (below), so the proof is omitted.

We find that the optimal policy is highly dependent on the relationship between w_C and w_{SC}. When the cost for lack of ICU access (w_C) is very large, 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 ν), 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. More formally, we have:

Proposition 4 *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:*

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

where

$$\nu = \frac{r_S\mu_{SC}}{r_I\mu_C} \textit{ and } \kappa = p + \nu$$

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. 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.

In interpreting the threshold, κ, which specifies the nurse allocation regime, we notice that it is comprised of p and ν. We make particular note of this quantity as it continues to arise as a main driver of the nurse allocation decision. p is a measure of the demand to the SDU as it indicates the proportion of patients who become semi-critical and can be treated in the SDU. In contrast, ν captures the supply side of the SDU as it indicates the effective capacity gains by moving a nurse from the ICU to the SDU. Hence, we see that the optimal nurse allocation is a matter of carefully balancing the supply gains and the demand needs of the semi-critical patients with the relative cost of lack of timely access to the ICU for critical patients. The fact that these parameters are additive also suggests that they play an equally important role.

We observe that the fluid regime is ‘bang-bang’ so 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 $\frac{w_C}{w_{SC}} = \kappa$.

In further interpreting the results of Proposition 4, 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 cost) 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, $w_C = \min\{w_C^Q/\theta, w_C^B\}$, w_{SC}, and $\kappa = p + \nu$.

In practice, we see that some hospitals have SDUs while others do not. Our analysis suggests that, under the optimal sizing decision of ICU and SDU, one should expect to see variation in the use of SDUs. While we cannot assess whether each hospital is sizing their SDU(s) in a reasonable manner, our analysis suggests that some of the variation seen in practice may be justified. Indeed, the threshold κ defined in Proposition 4 is the main driver dictating whether having an SDU is optimal or not. This threshold depends on the capacity needs of critical and semi-critical patients (as captured by $r_I \mu_C$ and $(r_S \mu_{SC}, p)$, respectively), which will vary based on patient mix and regulation, thereby resulting in different thresholds for different hospitals. We will see that these factors will again have a prominent role in the sizing decision as we refine our analysis.

4. Second order drivers of the nurse allocation and balking decision

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, which identified κ as a key parameter influencing the management of the ICU and SDU. 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 impactful. Moreover, the parameter $\kappa = p + \nu$ again proves to be a critical component in determining how such reallocations should be determined.

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. This approach involves centering the system by N times its fluid limit and then scaling by $1/\sqrt{N}$, so that only fluctuations which are on the order of $\sqrt{N}$ are considered. More details about diffusion analysis can be found in Halfin and Whitt (1981), Ward (2012) and Chapter 10.4 of Whitt (2002a).

4.1. Diffusion Analysis in the ID regime

In this section we assume that

$$\frac{w_C}{w_{SC}} > \kappa = p + \nu \text{ and } Nr_I \geq \frac{\lambda}{\mu_C} + o(N). \tag{6}$$

As such, by Proposition 4, the fluid solution determines that it is optimal to operate in the ID regime and allocate enough nurses to the ICU so that all critical demand is met. That is, the number of nurses allocated to the ICU satisfies $B_I^* = \lambda/\mu_C + o(N)$, and the ICU is considered to be critically loaded with respect to the critical patients (Mandelbaum and Zeltyn 2009). Note that if $Nr_I < \frac{\lambda}{\mu_C} + o(N)$, then the ICU would be overloaded and reallocating any nurses to the SDU would only increase costs.

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 and operates in the QED (Quality and Efficiency Driven) regime with respect to the critical patients (Halfin and Whitt 1981, Garnett et al. 2002). Because, under the QED regime, there will be times when some ICU beds are not occupied by critical patients, the flow of the semi-critical patients is more intricate in this setting than in the fluid scale.

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. Theorem 1 in the Appendix precisely characterizes the patient dynamics at the diffusion level. Specifically, the dynamics of our system–according to Theorem 1–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 ($M/M/B_I + M$) model with finite or infinite buffer (Garnett et al. 2002, Kocaga and Ward 2010).

2. Both units are always full when considering fluctuations which are of order $\sqrt{N}$ or larger. If there are fewer than B_I critical patients in the system, then semi-critical patients fill the remaining ICU beds. We refer to this result as a 'State-space collapse', because the two-dimensional queueing system collapses to one dimension in this regime.

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: an ICU that is always full may appear to be the system bottleneck when, in fact, the reason why it is full could be due to spillover from 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. 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.

To evaluate the steady-state cost, over which we will optimize, we leverage results from Kocaga and Ward (2010) (Theorem 2 in the appendix) who generalize results from Garnett et al. (2002) and Browne and Whitt (1995) to include a balking threshold. Consistent with that paper, we consider a balking threshold K^N which is of order $\mathcal{O}(\sqrt{N})$. Specifically, we assume that $K^N = \hat{k}\sqrt{N}$. For a fixed balking threshold and nursing allocation, Theorem 2 provides diffusion approximations for the balking rate, $\hat{L}$, and queue length, $E[\hat{Q}]$, which can be used to determine the balking and queue costs, respectively.

We now derive diffusion approximations for the bumping rates in order to evaluate the bumping costs–the last component of our cost function. To do this, we leverage the state-space collapse results of Theorem 1. The starting point is that the bumping rate is approximately equal to the semi-critical arrival rate minus its total service rate. The arrival rate may be expressed as: $E[Z_C]\mu_C p$, where Z_C is the number of critical patients in ICU beds. By Theorem 1, the number of semi-critical patients in an SDU or ICU bed is equal to the total number of SDU beds plus the 'idle' ICU beds, which are not currently occupied by critical patients. Thus, 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 can determine, under the ID regime and the nurse allocation (7), the corresponding 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 $1/\sqrt{\lambda}$) which we wish to optimize over.

Optimization Problem 3 (Diffusion Cost in ID regime) *We optimize over β and $\hat{k}$, which determine the nursing allocation and balking threshold, respectively.*

$$\min_{\beta,\hat{k}} C(\beta,\hat{k}) := \min_{\beta,\hat{k}} w_C^B\hat{L} + w_C^Q E[\hat{Q}] + w_{SC}\left[\beta\sqrt{\mu_C}\,(p+\nu) - (\mu_{SC}+\mu_C p)E[\hat{I}]\right], \tag{8}$$

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

As in our fluid analysis of Section 3, we see in the expression above that the ratio of capacity gains, ν, and the semi-critical demand, measured by p, play a prominent role in the optimal balking threshold and nurse allocation decision. Again, they are additive, suggesting their comparable importance. As semi-critical demand, p, or SDU capacity gains, ν increase, β will decrease, meaning more nurses will be allocated to the SDU.

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

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

Note that in the queue-dominated case ($w_C^Q/\theta \le w_C^B$), one can verify that it is never optimal to let a patient balk and that $K^* = \infty$ as is stated by Proposition 1 of Kocaga et al. (2015).

Recall that by assumption, the system operates in the ID regime. However, when computing β^*, it is possible that its value will be so small that, in fact, the solution proposed is effectively in the CD regime. By the fluid analysis we know that this is first-order suboptimal. Thus, we set a lower bound on B_I^* and an upper bound on B_S^* in order to guarantee that the solution is still in the ID regime as dictated by the fluid solution. In doing so, the allocation of nurses is given by:

$$B_I^* := \max\left\{\frac{\lambda}{\mu_C} + \beta^*\sqrt{\frac{\lambda}{\mu_C}},\ r_I\frac{\nu}{\kappa}N\right\},$$

and

$$B_S^* = \min\left\{\frac{r_S}{r_I}\left(Nr_I - \frac{\lambda}{\mu_C} - \beta^*\sqrt{\frac{\lambda}{\mu_C}}\right),\ r_S\left(1-\frac{\nu}{\kappa}\right)N\right\}.$$

In Section 5, we see cases where the values of both the optimal $\hat{k}$ and β are non-trivial.

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. In the fluid analysis, ensuring enough capacity for the critical patients was all that mattered.

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

$$\frac{w_C}{w_{SC}} \le \kappa = p + \nu.$$

In this case, according to Proposition 4, we have that

$$B_I^* = b_I^* N + o(N), \quad b_I^* = r_I\frac{\nu}{\kappa}, \text{ and } B_S^* = b_S^* N + o(N), \quad b_S^* = r_S\left(1-\frac{\nu}{\kappa}\right).$$

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

$$B_I = b_I^* N + o(N) = \gamma R_I + \delta\sqrt{R_I} + o(\sqrt{R_I}), \quad R_I := \frac{\lambda}{\mu_C}, \tag{9}$$

where $\gamma = \frac{N r_S \mu_{SC}}{\lambda(p+\nu)}$ is less than 1 due to Assumption 2. Also, let

$$B_S = b_S^* N + o(N) = R_S + \beta\sqrt{R_S} + o(\sqrt{R_S}), \quad R_S := \frac{B_I \mu_C p}{\mu_{SC}}, \tag{10}$$

The relation $\frac{B_I}{r_I} + \frac{B_S}{r_S} = N + o(\sqrt{N})$ gives us

$$\delta := \delta(\beta) = -\beta\sqrt{\frac{N r_S p}{\lambda \mu_C}} \frac{\mu_{SC}}{(p+\nu)^{3/2}},$$

where β is only restricted by the non-negativity constraints on B_I and B_S. We aim to find a value for β that minimizes the expected balking plus queue plus bumping cost.

By definition, R_I is the offered load of the ICU. We argue that R_S is the offered load of the SDU. To see this, note that, since $\gamma < 1$, the ICU is 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 $\frac{B_I \mu_C p}{\mu_{SC}}$. As expected, the SDU is critically loaded, and operates in the QED regime.

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 ∞. It turns out that in the balking-dominated case $K^* = o(\sqrt{N})$ (see Corollary 3 in the appendix). Unlike in the ID regime, we do not see a second order impact of optimizing the balking threshold in the CD regime.

An interesting conclusion from these results 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

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

We leverage results from Jagerman (1974) to determine the probability of bumping:

$$Pr\{Bump\} := \limsup_{T\to\infty} \frac{1}{T}\int_0^T \psi(Q(t), Z_C(t), Z_{SC}(t)) = \frac{1}{\sqrt{B_S}} h(-\beta) + o(1/\sqrt{\lambda}).$$

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 the relevant diffusion cost function.

Optimization Problem 4 (Diffusion Cost in CD regime) *In this regime, the optimal balking threshold (up to order $\sqrt{N}$) is either 0 or ∞. Thus, we are left to optimize over β, which determines the nursing allocation.*

$$\min_\beta C(\beta) = \min_\beta \mu_{SC}\sqrt{\frac{r_S p}{p+\nu}}\left(w_C\frac{\beta}{p+\nu} + w_{SC}h(-\beta)\right). \tag{11}$$

Again, we see the parameter $p+\nu$ plays a vital part in determining the optimal nursing allocation.

Let $\beta^* := \arg\min_\beta C(\beta)$, and let $\delta^* := \delta(\beta^*)$. Similar to the ID regime, we set an upper bound on B_I^* and a lower bound on B_S^* to ensure the allocation is still in the CD regime as given by the fluid solution. Then our proposed solution in the CD regime is:

$$B_I^* = \min\left\{\gamma R_I + \delta^*\sqrt{R_I}, r_I N, \frac{\lambda}{\mu_C}\right\}, \quad R_I := \frac{\lambda}{\mu_C}, \tag{12}$$

and

$$B_S^* = \max\left\{R_S^* + \beta^*\sqrt{R_S^*}, \frac{r_S}{r_I}\left(Nr_I - \frac{\lambda}{\mu_C}\right)\right\}, \quad R_S^* := \frac{B_I^*\mu_C p}{\mu_{SC}}. \tag{13}$$

5. Simulation: Robustness of Main Drivers

We have utilized fluid and diffusion analysis to identify the main drivers which dictate how nurses should be allocated to ICU and SDU beds and when patients should be blocked from entering the system. 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. Moreover, we find that the balance between semi-critical demand, p, SDU capacity gains, ν, and the relative cost of lack of access to care for critical patients drive the nurse allocation decision. Our analysis was based upon a parsimonious model of patient flows through the ICU and SDU. While this model captures many salient features of the patient dynamics, we wish to examine whether our insights translate to a more complex system which includes additional features which arise in practice, such as SDU arrivals from units other than the ICU, patient returns to the ICU or SDU after transferring elsewhere, and off-placement of critical patients in the SDU. To do so, we use simulation to examine the quality of our approximations as deriving closed form expressions for this more general model seems unlikely.

5.1. Simulation Model

We begin by describing our simulation model. We consider a system with N_I and N_S nurses in the ICU and SDU, respectively. This dictates that the number of beds in each unit is $B_I = r_I N_I$ and $B_S = r_S N_S$. In order to focus on the ICU and SDU nurse / bed allocation decision, we assume there is ample capacity in the general ward. As in our original model, we consider two types of patients: **Critical** and **Semi-Critical**.

Critical patients arrive to the system according to a Poisson process with (possibly time-varying) rate λ_C. If a critical patient arrives and there is an available bed in the ICU, he is immediately admitted to the ICU and his 'service time', i.e. the time spent in the critical state, is log-normally distributed with mean $1/\mu_C$ and standard deviation σ_C. We use a log-normal distribution as it has been found to accurately capture LOS in the hospital (Armony et al. 2015) and ICU (Litvak et al. 2008). If there are no available beds in the ICU a number of things can happen: 1) if there is a semi-critical patient in the ICU, she will be bumped and the critical patient will take her ICU bed. 2) If all ICU beds are occupied by critical patients, but there are available SDU beds, the critical patient is admitted to the SDU but is 'served' at a rate $\hat{\mu} = \mu/x$, where $x > 1$. Thus, if the patient would have been critical for T time in the ICU, the same patient would require $x \times T$ time if treated in the SDU. We refer to such a critical patient as an 'off-placed' critical patient. The dis-utility to a critical patient who is off-placed in the SDU is not equivalent to waiting in the queue (outside of the ICU and SDU). Thus, instead of incurring cost w_C^Q, we assume these patients incur a cost at rate $y \times w_C^Q$ where $y \in [0, 1]$. 3) If there are no available ICU and SDU beds, the patient will enter the queue (and potentially abandon later) as long as the total number of critical patients in the system is less than the balking threshold K. If there are more than K critical patients in the system, the patient will balk.

Once a patient is no longer in the critical state, one of four events can occur 1) he will immediately become semi-critical with probability p, 2) he will return to the system as critical after an exponentially distributed delay with mean δ with probability $p^R_{C,C}$, 3) he will return to the system as semi-critical after an exponentially distributed delay with mean δ with probability $p^R_{C,SC}$, or 4) he will leave the ICU/SDU system.

If an ICU bed becomes available, priority for that bed is given as follows: 1) if the critical patient occupying that bed becomes semi-critical, she keeps the bed. 2) Otherwise, if there are any off-placed critical patients in the SDU, the patient with the longer remaining time in the critical state is transferred into the ICU. 3) A critical patient who is in the queue will be admitted to the ICU bed (potentially bumping a semi-critical patient) with priority given to patients who are in the return queue. 4) Finally, if there are any *semi-critical* patients who are off-placed outside of the ICU and SDU, the one with the longest remaining time in the semi-critical state is transferred into the ICU.

Semi-critical patients can arrive to the system via three pathways: 1) as an external arrival which is given by a Poisson process with (possibly time-varying) arrival rate λ_{SC}^{EXT}, 2) transitioning to semi-critical immediately after being critical, or 3) returning to the system after some time. If an SDU or ICU bed is available, the semi-critical patient will be immediately admitted to the bed. The time spent in the semi-critical state is log-normally distributed with mean $1/\mu_{SC}$ and standard deviation σ_{SC}, irrespective of the unit in which the patient resides. If there are no SDU or ICU beds available, the semi-critical patient will be off-placed in the general medical-surgical ward. If a bed becomes available, priority is given to the patient with the longest remaining time in the semi-critical state. If a semi-critical patient completes service in the ward (while off-placed), this is counted as part of the bumping rate.

Once a patient is no longer in the semi-critical state, one of three events can occur 1) she will return to the system as critical after an exponentially distributed delay with mean δ with probability $p^R_{SC,C}$, 2) she will return to the system as semi-critical after an exponential delay with mean δ with probability $p^R_{SC,SC}$, or 3) she will leave the ICU/SDU system.

If a patient returns to the system as critical (semi-critical) and there are no ICU (ICU/SDU) beds available, the patient will wait in a virtual queue. Practically, this could correspond to the patient being treated in the Emergency Department or general ward.

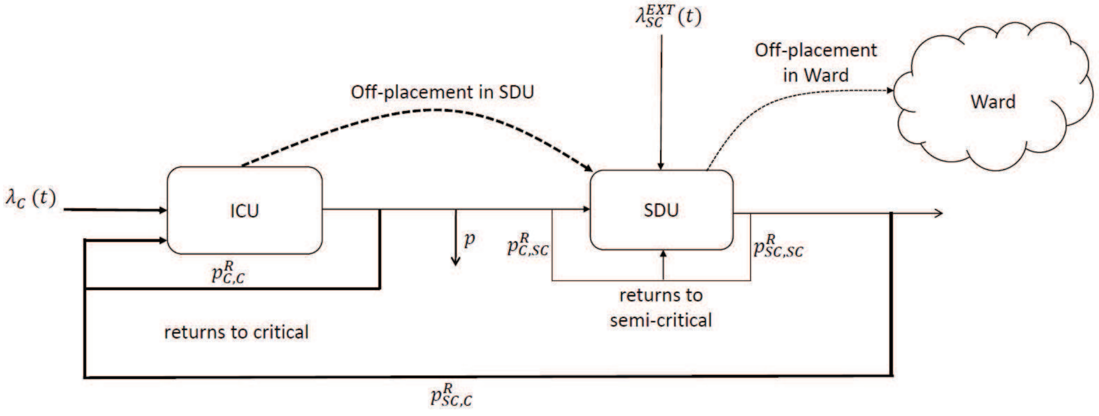


Figure 3

Figure 3 depicts our simulation model. We consider the allocation of N nurses between the ICU and SDU to care for patients who enter the system as critical. In our simulations, we consider five different policies:

1. **No SDU:** All N nurses are allocated to the ICU.

2. **Half-half:** Half of the N nurses are allocated to the ICU and the remaining $N/2$ are allocated to the SDU.

3. **Fluid:** We use our fluid solution from Section 3, which ignores readmissions, off-placements, and external arrivals.

4. **Diffusion:** We use our diffusion solution from Section 4, which ignores readmissions, off-placements, and external arrivals.

5. **Simulation:** We use simulation to run an exhaustive search over all possible allocations and present the best performing option.

In our original model of Section 2, we did not incorporate external arrivals of semi-critical patients. To account for these additional patients, we include an additional $B_{SDU}^{EXT} = \lambda_{SC}^{EXT}/\mu_{SC}$ beds and, correspondingly, $N_{SDU}^{EXT} = B_{SDU}^{EXT}/r_S$ nurses (in addition to the N nurses) to treat the external semi-critical arrivals.

We also optimize over the balking threshold. For the No SDU, Half-Half and fluid policies, we use the balking threshold dictated by the fluid solution in Proposition 3. For the Diffusion solution, we use the balking threshold specified by the solution of equation (8) or (11), depending on whether the system is in the ID or CD regime, respectively. For the Simulation policy, we search over all possible balking thresholds $K \in [0, 50]$ in the balking dominated regime, and set $K = \infty$ in the queue dominated regime.

5.2. Calibrating our Simulation

To start, we must first calibrate the parameters of our model. To do this, we utilize the existing medical literature. We leverage the results of Cady et al. (1995) who look at the impact of adding an SDU to the cardiothoracic ICU at the University of Missouri Hospitals to calibrate many of our parameters. We have that $1/\mu_C = 2.5$ days , $1/\mu_{SC} = 1.2$ days, $p = .65$ of the patients become semi-critical after being critical, and the nurse to patient ratio in the SDU is $1 : r_S = 1 : 3$. The ICU nurse-to-patient ratio is not given in this article, so we assume it to be one-to-one, $r_I = 1$. We set the probability of return to critical and semi-critical $p_{C,C}^R = p_{C,SC}^R = p_{SC,C}^R = p_{SC,SC}^R = .07$, which is similar to the rates given in Chan et al. (2016). Based on estimates from personal communication with medical professionals, we set $\theta = 1$ so that patients can tolerate waits of 1 day on average and $x = 1.5$ so that treating critical patients in the SDU takes 50% as long as in the ICU. Because the queue cost, w_C^Q, incorporates holding and abandonment costs, which are likely more detrimental than being off-placed, we assume that $y < 1$. Specifically, we assume that the cost of off-placement of critical patients is 30% less than the queue cost, i.e. $y = .3$. We consider the average time to return to service as 1 day, so that $\delta = 1$, based on estimates from conversations with clinicians. In Section 5.5, we consider other values of δ as robustness checks. We set the arrival rate of new critical and external semi-critical patients to be $\lambda_C = \lambda_{SC}^{EXT} = 8$ patients per day, with $B_{SDU}^{EXT} = 10$, and we consider how to allocate $N = 20$ nurses amongst the ICU and SDU. This corresponds to an ICU which is critically loaded if all nurses are allocated to the ICU, i.e. $\lambda_C = \mu_C r_I N$. We use a warmup of 1,000 days and average our results over 1,000,000 days.

Because the use of SDUs varies, we also use data from the surgical ICU/SDU at New York-Presbyterian Hospital (Eachempati et al. 2004). For this set of parameters, we have that $1/\mu_C = 4.8$ days , $1/\mu_{SC} = 2.3$ days, $p = .8$, $r_I = 2$, and $r_S = 4$. As before,we set the arrival rates to $\lambda_C = \mu_C r_I N = 8.33$ and $\lambda_{SC}^{EXT} = 8$. Table 1 summarizes our simulation parameters.

Lastly, we note that we also ran simulations with a time-varying arrival rate with a daily period to mimic the daily-cycles documented in the literature (e.g. Armony et al. (2015)). Similar to Chan et al. (2014b), we

	Model Primitives						Additional Parameters				
Source	$1/\mu_C$	$1/\mu_{SC}$	p	r_I	r_S	θ	λ	p^R	x	y	δ
Cady et al. (1995)	2.5 days	1.2 days	0.65	1†	2-3	1	8	.07	1.5	.3	1
Eachempati et al. (2004)	4.8 days	2.3 days	0.8	2	4		8.33				

Table 1 Summary of patient flow parameters for simulation model. The parameters $1/\mu_C, 1/\mu_{SC}, p, r_I$ and r_S are from the corresponding article on SDUs. †The ICU nurse-to-patient ratio is not given in this article, so we assume it to be one-to-one. The model primitives are the parameters which overlap between the analytic model in Section 2 and the simulation model of Section 5. The additional parameters are those which have been added to be included in the simulation model. Note that $\lambda = \lambda_C = \lambda_{SC}^{EXT}$ and $p^R = p_{i,j}^R$ for $i, j \in \{C, SC\}$.

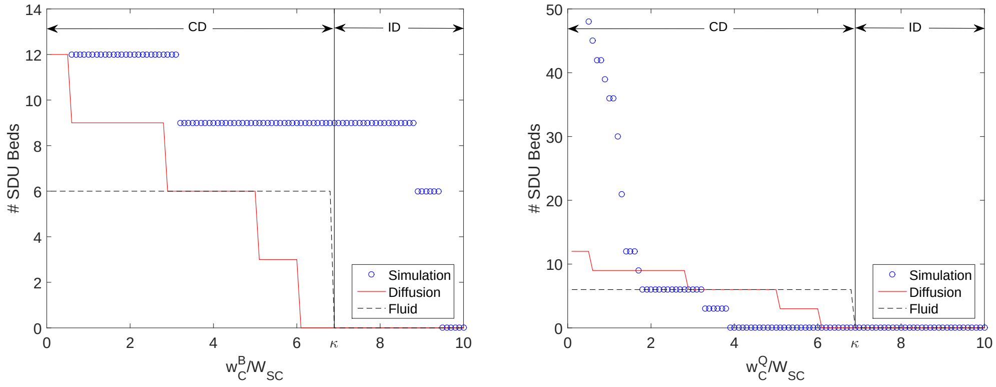
find that ignoring the time-variability, by simply using a fixed arrival rate equal to the daily average, results in practically the same performance and policy. This is not surprising due to the fact that the time-scale of variability is on the order of hours whereas the average 'service times' are in days.

5.3. Simulation Results

We start by examining the average costs incurred under various cost settings. 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 $w_C/w_{SC} = \kappa := p + \nu$. 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 κ, we assigned the number of ICU beds to be the average between the ID and CD diffusion solutions. That is, let $B_I^*(\text{ID}, \kappa)$ be the ID solution (minimizes Eqn. (8)) and let $B_I^*(\text{CD}, \kappa)$ be the CD solution (minimizes Eqn. (11)) when $w_C/w_{SC} = \kappa$. Then, our diffusion solution is $B_I^* = \frac{1}{2}[B_I^*(\text{ID}, \kappa) + B_I^*(\text{CD}, \kappa)]$, which also serves as a lower (upper) bound for the number of ICU beds in the ID (CD) regime.

Figure 4 compares the number of SDU beds from our analysis to the exhaustive search when there are 20 nurses to split amongst the ICU and SDU in the balking and queue dominated cases. The number of ICU beds can be easily determined via the following relationship: $B_I = r_I \times (20 - B_S/r_S)$. Note that because 1 nurse can treat up to 3 patients in the SDU, we see discrete jumps in multiples of 3 in the number of beds. As we can see in these figures, the solution determined by minimizing the cost in (8) and (11) 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.

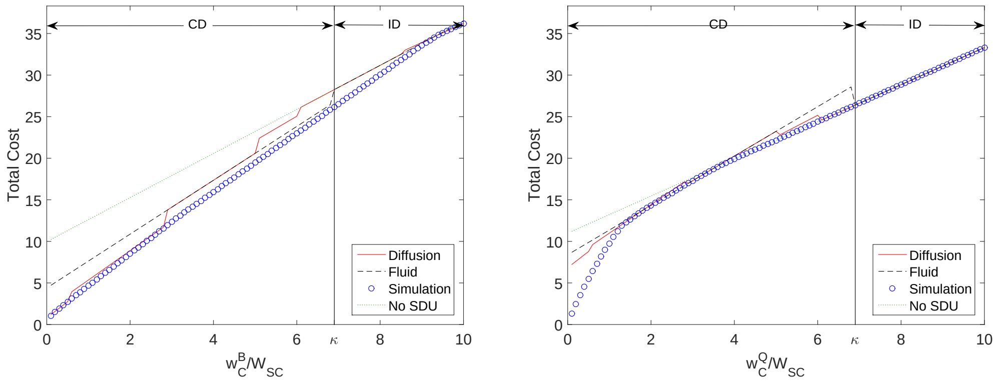
Though we see discrepancies in the number of beds in the ICU and SDU under the diffusion approximations, we find that the actual average cost incurred performs quite well. Figure 5 compares the simulated costs under the diffusion and fluid solutions to the minimum cost achieved via exhaustive search. We do not plot the cost incurred by the half-half allocation as its performance is much worse than all policies. In some cases, the performance of the diffusion and fluid policies is practically indistinguishable from that



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

Figure 4 **Optimal allocation of nurses to SDU beds via fluid and diffusion analysis compared to an exhaustive search.** $N = 20$ **nurses.** $w_{SC} = 1$**. Hospital parameters given by Cady et al. (1995).**

of the exhaustive search. We do see that the quality of the proposed solutions appear to degrade in the Queue-dominated case (when w_C^Q/w_{SC} is very small). We explore why this might be the case in Section 5.4.



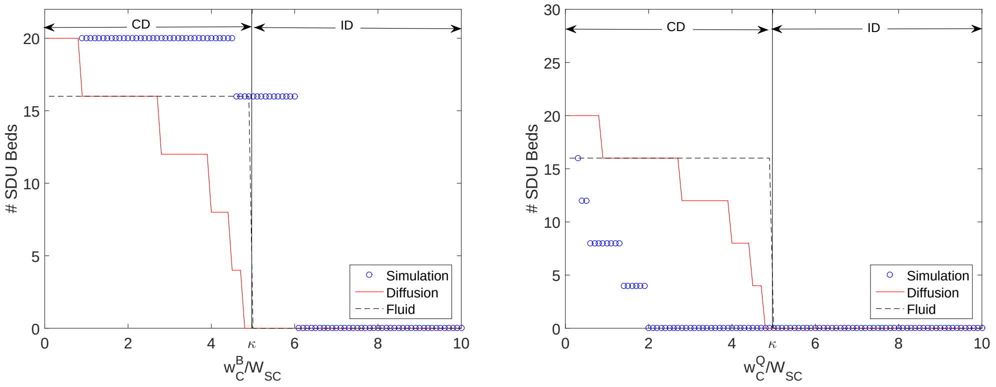
(a) Balking-Dominated Case: $w_C^Q/\theta = 15$ (b) Queue-Dominated Case: $w_C^B/\theta = 15$

Figure 5 **Cost incurred by various strategies to determine the balking threshold and nurse allocation.** $N = 20$ **nurses.** $w_{SC} = 1$**. Hospital parameters given by Cady et al. (1995).**

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.

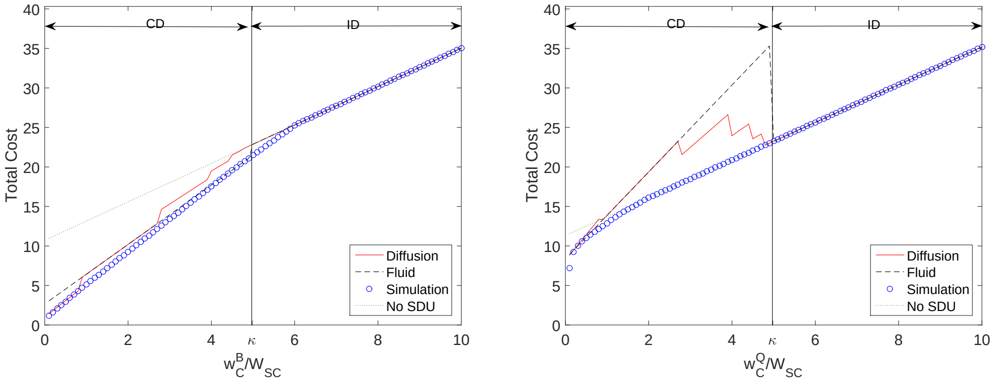
In the balking-dominated case, the balking threshold from the diffusion analysis is quite close to that from the simulation. In fact the discrepancies between the solutions is at most 1. For the sake of space, the corresponding figure is not included, but is available upon request. In the queue-dominated case, all of the balking thresholds coincide with $K = \infty$.

Figures 6 and 7 are the analogs of Figures 4 and 5 with the hospital parameters given by Eachempati et al. (2004). Because the nurse-to-patient ratios in Eachempati et al. (2004) require fewer nurses per patient than in Cady et al. (1995), the sizes of the units are twice as large for the Eachempati et al. (2004) parameters. In this case, the balking threshold is equal to 0 for all solutions in the balking-dominated case. We find that the qualitative results are similar, though the differences in performance of the fluid and diffusion solutions compared to the exhaustive search are more pronounced. Going forward, we will only provide results for the hospital parameters given by Cady et al. (1995).



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

Figure 6 **Optimal allocation of nurses to SDU beds via fluid and diffusion analysis compared to an exhaustive search. $N = 20$ nurses. $w_{SC} = 1$. Hospital parameters given by Eachempati et al. (2004).**

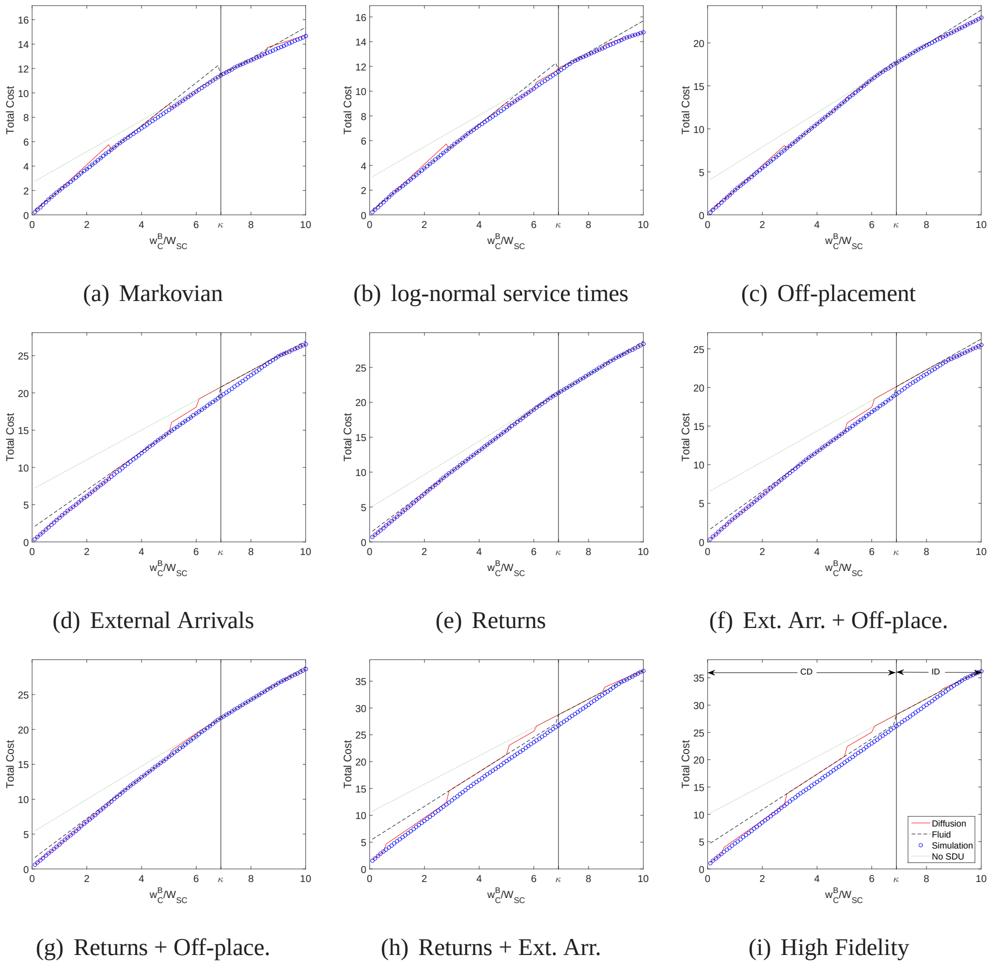


(a) Balking-Dominated Case: $w_C^Q/\theta = 15$ (b) Queue-Dominated Case: $w_C^B/\theta = 15$

Figure 7 **Cost incurred by various strategies to determine the balking threshold and nurse allocation. $N = 20$ nurses. $w_{SC} = 1$. Hospital parameters given by Eachempati et al. (2004).**

5.4. Transition from Markovian Model to High Fidelity Model

In Figures 5 and 7, we saw that – especially in the queue dominated case – the quality of the fluid and diffusion solutions derived from our analytic model can deviate from the optimal solution, as determined by exhaustive simulation, in the high fidelity model. We now aim to understand better what factors contribute to this degradation in performance. In particular, it is possible there are two sources of errors. First, the fluid and diffusion solutions themselves are approximations for finite systems. Second, our high fidelity simulation model incorporates a number of features which are not present in our analytic model and they may be causing the errors.



(a) Markovian

(b) log-normal service times

(c) Off-placement

(d) External Arrivals

(e) Returns

(f) Ext. Arr. + Off-place.

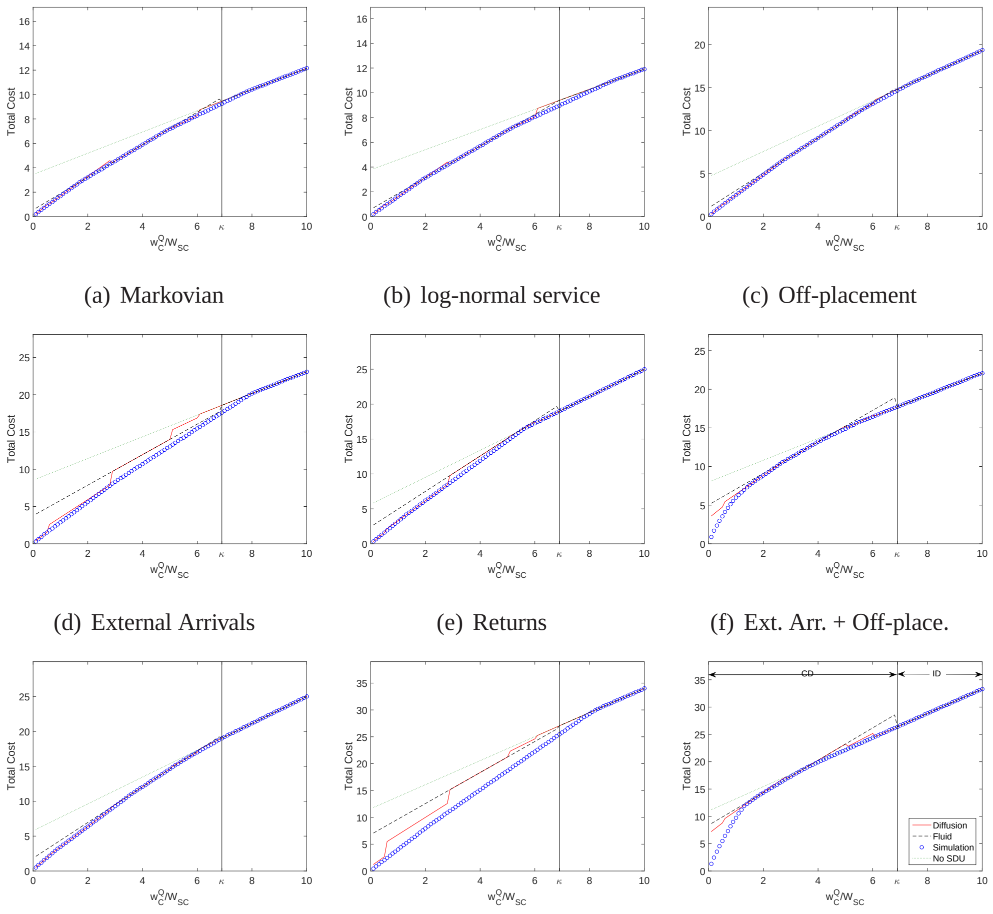
(g) Returns + Off-place.

(h) Returns + Ext. Arr.

(i) High Fidelity

Figure 8 Balking-Dominated Case: Cost incurred by various strategies to determine the balking threshold and nurse allocation. $N = 20$ **nurses.** $w_{SC} = 1$**. Hospital parameters given by Cady et al. (1995).**

Figure 8 and 9 present the cost incurred by the various strategies to determine the balking threshold and nurse allocation where the hospital parameters given by Cady et al. (1995). Each subfigure corresponds to a different simulation model. Specifically, we begin with the Markovian model which corresponds to our analytic model presented in Section 2. We incrementally add features to the model – log-normal service times, off-placement of patients, external arrivals of semi-critical patients, and returns to service – until all features are incorporated into our high fidelity simulation model described in Section 5.1.



(a) Markovian (b) log-normal service (c) Off-placement

(d) External Arrivals (e) Returns (f) Ext. Arr. + Off-place.

(g) Returns + Off-place. (h) Returns + Ext. Arr. (i) High Fidelity

Figure 9 Queue-Dominated Case: Cost incurred by various strategies to determine the balking threshold and nurse allocation. $N = 20$ **nurses.** $w_{SC} = 1$**. Hospital parameters given by Cady et al. (1995).**

We can see that in both the Balking and Queue Dominated cases, the solutions from our analysis of the Markovian model perform very well when comparing the performance under the simulated Markovian model. This suggests the deviations in performance are not due to the fact that our model has a finite number of nurses while the fluid and diffusion solutions are solutions in asymptotic regimes with $N \to \infty$.

We notice that the presence of off-placement and external arrivals of semi-critical patients appears to have a substantial impact on the quality of our proposed solutions. This is most evidence in the queue-dominated case, as seen in Figure 9. We believe this is likely because our model considers off-placed patients as part of the queue, even though they are receiving some ‘treatment’ in the off-placed unit, altering the queueing dynamics. In the balking-dominated case, there will be very few off-placed patients, so the impact on the quality of the solutions is diminished. The external arrivals seem to slightly deteriorate the performance of our solutions in both the queue and balking dominated cases – particularly near the threshold κ between the ID and CD regimes. It is in this regime that second order effects really drive the SDU sizing decision. As such, the external arrivals of semi-critical patients, which are of the first order, can have a marked impact on our proposed solutions.

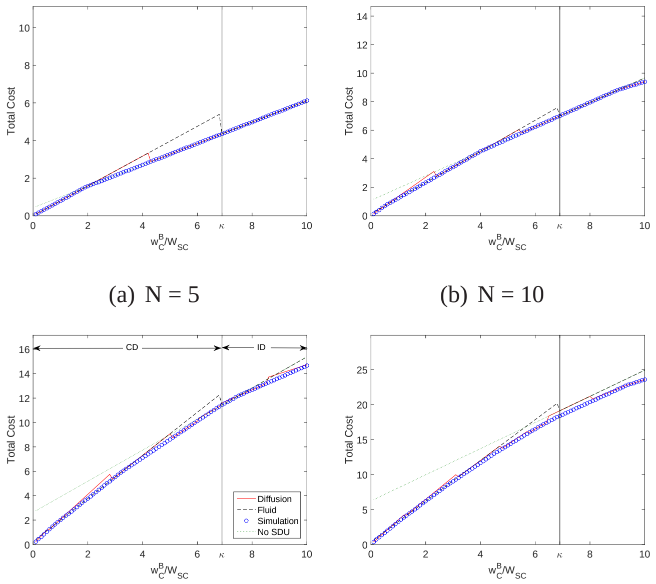
5.5. Robustness Checks

We also conducted a number of robustness checks of our high fidelity simulation to examine how sensitive our results are to changes in the parameters. First, we considered the case where off-placement of patients increased the probability of return to service by 10% and then 20%. Next, we varied the average time to return to service from $\delta = 1$ to $\delta = 2$ and $\delta = 3$. Finally, we considered arrival rates so that the system was not critically loaded. Recall that we set $\lambda = \mu_C r_I N$, so that the system would be critically loaded if all nurses were allocated to the ICU. We then considered three scenarios where $\lambda = .8 \times \mu_C r_I N$, $.85 \times \mu_C r_I N$, and $.9 \times \mu_C r_I N$. In all of these experiments, the qualitative performance of our proposed policies remains consistent (although there are very slight quantitative variations). Specifically, both the fluid and diffusion solutions perform quite well in the Balking-dominated case, while their performance degrades in the Queue-dominated case, especially by the threshold between the ID and CD regimes, κ.

We also examined the impact of the off-placement cost, yw_C^Q, by varying $y \in [0,1]$. The results in the Queue dominated case can be found in Figure EC.1 of EC-3. The performance of the diffusion/fluid solutions are reasonably accurate for moderate values of y (e.g. 10%-60%). We find that when $y \sim 1$, there are substantial deviations between the fluid and diffusion solutions with the optimal solution. As seen in Figure EC.1, the off-placement results in these deviations and increasing the costs associated with off-placement seems to magnify this effect. Interestingly, when $y = 0$, the optimal solution puts all nurses in the SDU (having no ICU). All patients are off-placed at no cost, resulting in very poor performance of the fluid and diffusion solutions. That said, it seems unreasonable to believe that off-placement would come at zero cost.

Next we examine the robustness of our insights to systems of different sizes. In particular, we simulate systems with $n = 5, 10, 20$, and 40 nurses to allocate between the ICU and SDU. Note that for these simulations we scale the arrival rate so that $\lambda = r_I \mu_C N$. Figures 10 and 11 show the performance of our proposed policies under Markovian dynamics (i.e. our analytic model presented in Section 2) in the Balking and Queue dominated cases, respectively. We can see that, as expected, the performance of the fluid and diffusion policies improve as the system size increases.

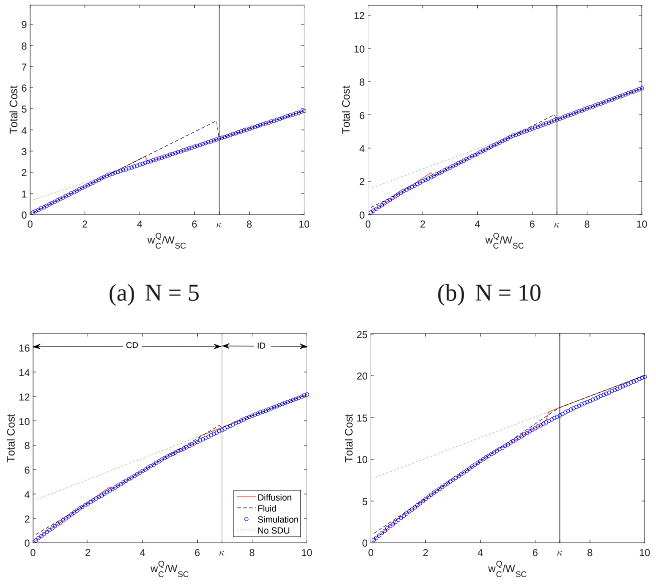
Figures 12 and 13 depict the high fidelity simulation model for different number of nurses. Here, we see that the quality of the solutions does not seem to improve with system size. In particular, it seems that the impact of the external arrivals and off-placement, which degrades the performance of our policies (see Section 5.4), outweighs the impact of system size. Thus, we find that while the performance of the fluid and diffusion policies improve under the Eachempati et al. (2004) parameters than under the Cady et al. (1995)



(a) N = 5 (b) N = 10

(c) N = 20 (d) N = 40

Figure 10 Markovian model: Cady Costs - Balk Dominated

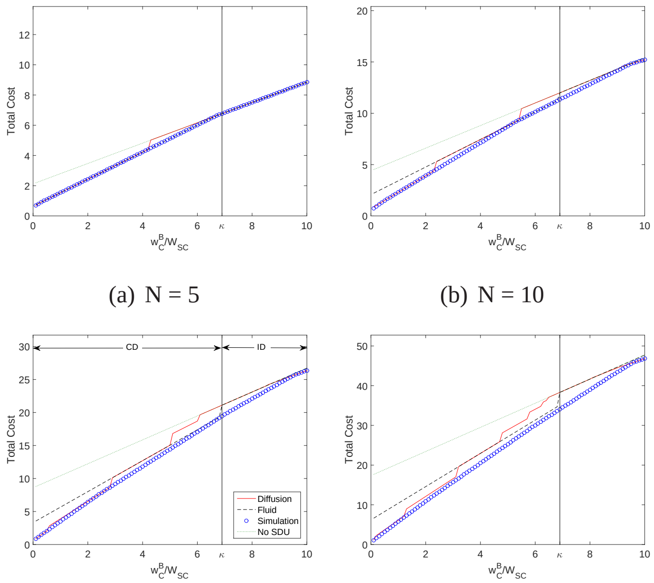


(a) N = 5 (b) N = 10

(c) N = 20 (d) N = 40

Figure 11 Markovian Model: Cady Costs - Queue Dominated

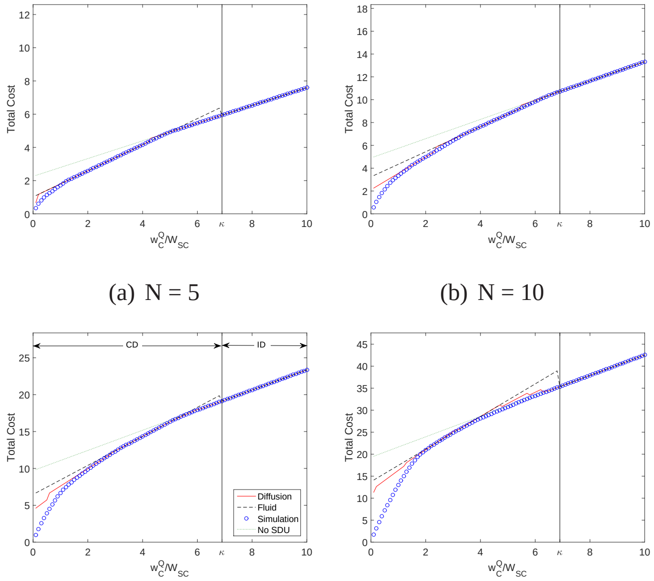
parameters in the Markovian simulation model due to the higher nurse to patient ratio which results in larger systems, this is not true for the high fidelity simulation model. For the sake of space, we do not include figures for the Eachempati et al. (2004) simulations, but they are available from the authors upon request.



(a) N = 5 (b) N = 10

(c) N = 20 (d) N = 40

Figure 12 **High Fidelity Model: Cady Costs - Balk Dominated**



(a) N = 5 (b) N = 10

(c) N = 20 (d) N = 40

Figure 13 **High Fidelity Model: Cady Costs - Queue Dominated**

6. Conclusions and Discussion

In this work, we consider the optimal allocation of nurses for the inpatient units used to treat the hospital's 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 and what factors drive the optimal use of SDUs. Within the medical community, there is an ongoing debate with regards to how hospitals should manage and size SDUs. Our work is an important first step towards helping to resolve this debate.

We propose a queueing model which allows us to examine how to optimally tradeoff capacity needs with capacity gains 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. Our results suggest that the first-order optimal solution is ‘bang-bang’ in that, 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 when second order terms are considered. Through our analysis, we identify the main drivers in managing the ICU/SDU sizing decision and balking threshold. We isolate two main parameters which impact the manner in which the nurse allocation decision should be made. One captures the demand for SDU beds as measured by the fraction of critical patients who become semi-critical; the other captures the supply gain from SDU beds as measured by the ratio of the effective capacity of a nurse in the SDU versus the ICU. We see that these two parameters play a critical role in the nurse allocation decision at both the fluid and diffusion level. Additionally, we find that optimizing the balking threshold beyond the dichotomy of allowing or not allowing balking only has a second order impact on reducing costs, if that. Using simulation, we find that our analysis in these asymptotic regimes can be quite accurate, even as we relax 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. In some cases, 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 a complementary explanation. 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. As such, we find that even if hospitals were sizing their SDUs in an optimal manner, we would still expect to see high variation in the use of SDUs. Moreover, our analysis allows us to isolate supply and demand characteristics of the ICU and the SDU which can be used to identify how hospitals should think about managing these units.

This work suggests several potential directions for future research. In our analytic model, we focused on patient flows from the ICU into the SDU and did not incorporate other patient flows through the ICU. These modeling choices were necessary for tractability and we used extensive simulations to develop and understanding of how our insights might translate to other settings. An interesting direction would be to consider other patient flows through the SDU. In our simulations, we see that our proposed policies work reasonably well even with alternative flows, so we conjecture that similar insights will carry over. That said, in the queue-dominated case off-placement and external arrivals contributed to deviations from our proposed solutions, so these are two features that one might want to consider as a reasonable next step. 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. Another interesting direction one could consider is the impact of state-dependent dynamics. A number of recent works (e.g. Chan et al. (2014b, 2016)) have found that patient flow parameters (e.g. μ_C, μ_{SC}, p) can be dependent on congestion, which begs the question as to how these dynamics may impact the management of the ICU and SDU. 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 is only used as a true Step-Down Unit, so that patients are only 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 isolate the main parameters p and ν driving the management of the SDU in order 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 Michael Huang his research assistance with our simulation study. The research by C. W. Chan was partially supported by the National Science Foundation [CMMI-1233547].

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. 2012. Flexible servers in tandem lines with setups. *Queueing Systems* **70**(2) 165–186.

Armony, M., S. Israelit, A. Mandelbaum, Y.N. Marmor, Y. Tseytlin, G.B. Yom-tov. 2015. Patient flow in hospitals: A data-based queueing-science perspective. *Stochastic Systems* **5**(1) 146–194.

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.

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. *Manufacturing & Service Operations Management* **17**(2) 157–176.

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. 2016. The Impact of Delays on Service Times in the Intensive Care Unit. *Management Science, to appear* .

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.

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.

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* **38**(4) 761–824.

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

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.

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 Noncardiac 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. 2012. Flexible servers in understaffed tandem lines. *Production and Operations Management* **21**(4) 761–777.

Kocaga, Y. L., M. Armony, A. R. Ward. 2015. Staffing call centers with uncertain arrival rates and co-sourcing. *Production and Operations Management* **24**(7) 1101–1117.

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.

Litvak, N., M. Van Rijsbergen, R. J. Boucherie, M. van Houdenhoven. 2008. Managing the overflow of intensive care patients. *European journal of operational research* **185**(3) 998–1010.

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* .

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.

State of California Office of Statewide Health Planning & Development. 2010-2011. Annual Financial Data. URL `http://www.oshpd.ca.gov/HID/Products/Hospitals/AnnFinanData/CmplteDataSet/index.asp,ac`

Stolyar, Alexander L. 2014. Tightness of stationary distributions of a flexible-server system in the halfin-whitt asymptotic regime. *arXiv preprint arXiv:1403.4896* .

Strachota, Ellen, Pamela Normandin, Nancy OBrien, Mary Clary, Belva Krukow. 2003. Reasons registered nurses leave or change employment status. *Journal of Nursing Administration* **33**(2) 111–117.

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.

Vincent, J.-L., G. D. Rubenfeld. 2015. Does intermediate care improve patient outcomes or reduce costs? *Critical Care* **19**(1) 89–94.

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

Ward, A. R. 2012. Asymptotic analysis of queueing systems with reneging: A survey of results for fifo, single class models. *Surveys in Operations Research and Management Science* **17**(1) 1–14.

Whitt, W. 2002a. *Stochastic-process limits: an introduction to stochastic-process limits and their application to queues*. Springer Science & Business Media.

Whitt, W. 2002b. *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. 2014. Erlang-r: A time-varying queue with reentrant customers, in support of health-care staffing. *Manufacturing & Service Operations Management* **16**(2) 283–299.

Zhang, Bo, H. Ayhan. 2013. Optimal admission control for tandem queues with loss. *IEEE Transactions on Automatic Control* **58**(1) 163–167.

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. Technical Results and Proofs

Proof of Proposition 1:

1. Suppose that $\limsup_{N\to\infty} \lambda \left(\frac{1}{r_I \mu_C} + \frac{p}{r_S \mu_{SC}} \right) \leq N$. Note, this implies that the offered load in the ICU can be met: $\limsup_{N\to\infty} \frac{\lambda}{r_I \mu_C} \leq N$. 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 $\frac{\lambda}{B_I \mu_C} \leq 1$, 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 $\limsup_{N\to\infty} \rho_C \leq 1$, 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 $\lim_{N\to\infty} \rho_C \leq 1$. □

EC-1.1. Diffusion Analysis in the ID regime

We examine the two-dimensional process with state $(Q + Z_C, Z_{SC})$, where 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. 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*

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

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 (2002b)).

PROOF OF THEOREM 1: Suppose that (4) holds in the limit. That is, assume that

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

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 $\tau_N = \inf\{t \ge 0;\ \hat{U}^N(t) < -\epsilon\}$ and $\tau_N' = \sup\{t \le \tau_N;\ \hat{U}^N(t) \ge -\epsilon/2\}$. 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 patient arrivals into the ICU from the queue in $(s,t]$; $D_C^N(s,t]$ is the number of critical patients who have completed their stay in the ICU and did not switch to a semi-critical state during $(s,t]$; and, $D_{SC}^N(s,t)$ 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

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

$$D_{SC}^N = S_3\left(\mu_{SC}\int_s^t Z_{SC}^N(r)\cdot dr\right) \leq S_3\left(\mu_{SC}\int_s^t \left(B_S^N + B_I^N - Z_C^N(r)\right)\cdot dr\right)$$
$$= (t-s)\cdot\left(\frac{\mu_{SC} r_S}{r_I}\left(N r_I - \frac{\lambda}{\mu_C}\right) + o(\lambda)\right).$$

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:

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

□

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 and in expectation (e.g. Gurvich and Whitt (2009a), Stolyar (2014)). A lengthy and rather technical mathematical argument is required to formally establish this result. This is outside the scope of this work; we simply postulate here that the same state-space collapse holds in steady-state and in expectation.

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'. We will leverage this result to evaluate the steady-state cost. To do so, we will rely on a result that follows directly from the results in Kocaga and Ward (2010). Note that these results generalize those of Garnett et al. (2002) and Browne and Whitt (1995).

Theorem 2 ***(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:

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

EC-1.2. Diffusion Analysis in the CD Regime

We start by examining the optimal balking threshold. First notice that by Theorem 4.3 of Mandelbaum and Zeltyn (2009), we have that in this regime, when no balking occurs $\lim_{\lambda\to\infty} EQ^N/\lambda = (1-\gamma)/\theta$. 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 5 ***(Balking threshold in the CD regime)*** *In the CD regime and under the nurse allocation of (9) and (10) if a threshold policy is used with threshold K^N that satisfies*

$$\limsup_{N\to\infty} \frac{K^N}{(1-\gamma)\lambda/\theta} = 1-\eta, \quad 0<\eta\le 1, \tag{EC-3}$$

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

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

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.

Proof of Proposition 5: Suppose that (EC-1) holds. Additionally, assume that the system operates in the CD regime and that (9) and (10) 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 $\epsilon>0$,

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

The proof follows along the lines of Reiman (1984). Fix $\epsilon>0$ and let $\tau_N = \inf\{t\ge 0;\ \hat{U}^N(t) < -\epsilon\}$ and $\tau'_N = \sup\{t\le\tau_N;\ \hat{U}^N(t)\ge -\epsilon/2\}$. 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

$$\begin{aligned} A^N(s,t) &= S_1\left(\int_s^t \lambda 1_{\{Q^N(r)<K^N\}}\cdot dr\right) = (t-s)\cdot(\lambda + o(\lambda)), \\ D_C^N(s,t) &= S_2\left(\mu_C \int_s^t Z_C^N(r)\cdot dr\right) \le S_2\left(\mu_C B_I(t-s)\right) = (t-s)\cdot(\gamma\lambda + o(\lambda)), \end{aligned} \tag{EC-4}$$

$$\begin{aligned}\Phi^N(s,t) &= S_3\left(\theta\int_s^t Q^N(r)\cdot dr\right) \leq S_3\left(\theta K^N(t-s)\right)\\ &\leq S_3\left((1-\gamma)(1-\eta/2)(t-s)+o(\lambda)\right) = (t-s)\cdot\left((1-\gamma)(1-\eta/2)+o(\lambda)\right).\end{aligned}$$

Finally, we have:

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

□

Corollary 2 *Under the conditions of Proposition 5, we have that the number of ICU beds that are occupied by critical patients is equal to $B_I - o(\sqrt{N})$.*

Corollary 3 *Under the conditions of Proposition 5, the optimal threshold in the balking-dominated case satisfies $K^{*N} = o(\sqrt{N})$.*

PROOF OF COROLLARY 3: By the fluid analysis we have that $K^{*N} = o(N)$. Therefore, K^{*N} satisfies (EC-3) with $\eta = 1$. By Corollary 2, 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 5, 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. When beds are the constrained resources

Consider a hospital with B critical care beds. The goal is to determine how to split these beds between the ICU (B_I) and SDU (B_S). Due to the strict nurse-to-patient ratios which must be maintained, determining the number of beds at each level of care will dictate how many nurses (N) are required to staff the units. We make this decision in order to minimize staffing costs in addition to the cost of lack of access to care, which are given by balking, abandonment, holding and bumping costs in our original model of Section 2. For simplicity, we consider the case with no balking. Thus the problem we aim to solve is

$$\begin{aligned}\min_{B_I,B_S,N\geq 0}\quad & w_N N + w_C R_{Ab} + w_{SC} R_{Bump} \qquad \text{(EC-5)}\\ \text{Subject to}\quad & B_I/r_I + B_S/r_S \leq N\end{aligned}$$

$$B_I + B_S \leq B,$$

where w_N is the staffing cost rate per nurse, $w_C = w_C^Q/\theta$ is the critical cost, w_{SC} is the bumping cost, and R_{Ab} and R_{Bump} are the abandonment and bumping rates, respectively.

By observing that the maximum number of nurses needed given a fixed number of beds B is $N = B/r_I$ (obtained when all beds are designated as ICU beds, $B_I = B$), the overload Assumption 2 translates into

Assumption 3 *The system operates in overload. That is,*

$$\frac{\lambda}{r_S\mu_{SC}}\left(p + \frac{r_S\mu_{SC}}{r_I\mu_C}\right) > \frac{B}{r_I}. \tag{EC-6}$$

Assuming a setting with a large number of beds B, we now consider fluid analysis to obtain insights into the first order drivers of the bed allocation decision. Redefine $\bar{\lambda} = \lambda/B, \quad b_I = B_I/B, \quad b_S = B_S/B,$ and $n = N/B$. then the corresponding fluid problem that only considers terms of order B is

$$\begin{aligned} \min_{b_I,b_S,n\geq 0} \quad & w_N n + w_C\left(\bar{\lambda} - b_I\mu_C\right)^+ + w_{SC}\left(p\left(b_I\mu_C \wedge \bar{\lambda}\right) - b_S\mu_{SC}\right)^+ \\ \text{Subject to} \quad & b_I/r_I + b_S/r_S \leq n \\ & b_I + b_S \leq 1. \end{aligned} \tag{EC-7}$$

To solve the problem (EC-7), we first fix $0 \leq b_I \leq 1 \wedge \bar{\lambda}/\mu_C$ and consider the optimal values of b_S and n as a function of b_I. We will then optimize over b_I. Note that it is sufficient to consider values of $b_I \leq \bar{\lambda}/\mu_C$ due to Proposition 2. The next lemma helps to identify the optimal value of b_S given a value of b_I, by using the tradeoff between staffing and bumping costs, as well as the value of n in optimality.

Lemma 1 *Given $0 \leq b_I \leq 1 \wedge \bar{\lambda}/\mu_C$, then*

$$n = \frac{b_I}{r_I} + \frac{b_S}{r_S}, \tag{EC-8}$$

and

$$b_S = \begin{cases} \min\left\{1 - b_I, p\frac{b_I\mu_C}{\mu_{SC}}\mu_{SC}\right\}, & \textit{if } \frac{w_N}{w_{SC}} < r_S\mu_{SC}; \\ 0, & \textit{if } \frac{w_N}{w_{SC}} \geq r_S\mu_{SC}; \end{cases}$$

PROOF: Given $0 \leq b_I \leq 1$ and $0 \leq b_S \leq 1 - b_I$, we have $n \geq b_I/r_I + b_S/r_S$ from feasibility. Now if $n > n_0 := b_I/r_I + b_S/r_S$, then the solution (b_I, b_S, n) is feasible and incurs a higher staffing cost than the feasible solution (b_I, b_S, n_0) without improving the abandonment or bumping cost. This proves (EC-8).

Next fix $0 \leq b_I \leq 1 \wedge \bar{\lambda}/\mu_C$. For feasibility, we must have that $0 \leq b_S \leq 1 - b_I$, so by (EC-8), we have that $n = b_I/r_I + b_S/r_S$. We start by considering the case where $1 - b_I < p\frac{b_I\mu_C}{\mu_{SC}}$. For any $b_S \in (0, 1 - b_I)$, let us examine the marginal value of increasing (decreasing) n. When n increases by dn, the marginal staffing cost increases by $w_N \cdot dn$. At the same time, this increase in n allows us to correspondingly increase b_S

by $r_S \cdot dn$. Because $b_S < p\frac{b_I\mu_C}{\mu_{SC}}$, the bumping cost is non-zero and this additional capacity decreases the bumping cost by $w_{SC}r_S\mu_{SC} \cdot dn$. All together, the marginal effect on the cost is $(w_N - w_{SC}r_S\mu_{SC}) \cdot dn$. Thus, depending on whether this marginal cost is positive or negative b_S should be minimal (0) or maximal ($1 - b_I$), respectively.

Now we consider the case where $1 - b_I \geq p\frac{b_I\mu_C}{\mu_{SC}}$. In this case, if $b_S = p\frac{b_I\mu_C}{\mu_{SC}}$ the bumping cost is equal to zero. Therefore, increasing b_S beyond this point will only increase the overall cost. Thus, depending on whether the marginal cost of increasing n (calculated above) is positive or negative b_S should be minimal (0) or maximal ($p\frac{b_I\mu_C}{\mu_{SC}}$), respectively. □

We can now reformulate the objective function (EC-7) in terms of b_I and optimize over this variable. Specifically, we wish to solve:

$$\min_{0 \leq b_I \leq 1 \wedge \bar{\lambda}/\mu_C} w_N(b_I/r_I + b_S/r_S) + w_C(\bar{\lambda} - b_I\mu_C) + w_{SC}(pb_I\mu_C - b_S\mu_{SC}), \tag{EC-9}$$

where b_S is determined according to Lemma 1. We have two cases to consider depending on the relationship between $\frac{w_N}{w_{SC}}$ and $r_S\mu_{SC}$.

Proposition 6 (No SDU) *If $\frac{w_N}{w_{SC}} \geq r_S\mu_{SC}$ then the optimal solution of (EC-9) is to have no SDU ($b_S^* = 0$), an ICU with b_I^* beds, with*

$$b_I^* = \begin{cases} 1 \wedge \bar{\lambda}/\mu_C & \textit{if } \frac{w_N}{w_C - pw_{SC}} \leq r_I\mu_C, \quad \textbf{\textit{ID regime}} \\ 0 & \textit{otherwise}, \quad \textbf{\textit{No critical care}} \end{cases} \tag{EC-10}$$

and $n^* = b_I^*/r_I$.

The intuition behind this result is that w_N is the marginal cost increase of increasing n, while $r_Iw_C\mu_C$ and $r_Iw_{SC}p\mu_C$ are the marginal decrease in abandonment cost and marginal increase in bumping cost, respectively. Thus, depending on whether the marginal cost is positive or negative, b_I^* will obtain its minimal or maximal possible value.

PROOF: Because $\frac{w_N}{w_{SC}} \geq r_S\mu_{SC}$, $b_S^* = 0$ by Lemma 1. Minimizing (EC-9) then becomes:

$$\min_{0 \leq b_I \leq 1 \wedge \bar{\lambda}/\mu_C} w_N b_I/r_I + w_C(\bar{\lambda} - b_I\mu_C) + w_{SC}pb_I\mu_C.$$

This is a linear function of b_I which obtains its minimum at 0 or $1 \wedge \bar{\lambda}/\mu_C$ when the coefficient of b_I is positive or negative, respectively. □

We now examine the case where $\frac{w_N}{w_{SC}} < r_S\mu_{SC}$.

Proposition 7 *If $\frac{w_N}{r_S\mu_{SC}} < w_{SC}$ then*

$$b_I^* = \begin{cases} 1 \wedge \bar{\lambda}/\mu_C & \textit{if } w_C\mu_C \geq w_N(1/r_I - 1/r_S) + w_{SC}(p\mu_C + \mu_{SC}), \quad \textbf{\textit{ID regime}} \\ 0 & \textit{if } \frac{w_C}{w_N} < \frac{1}{r_S\mu_{SC}}\left(p + \frac{r_S\mu_{SC}}{r_I\mu_C}\right), \quad \textbf{\textit{No critical care}} \\ \frac{\mu_{SC}}{p\mu_C + \mu_{SC}} & \textit{otherwise}, \quad \textbf{\textit{CD regime}} \end{cases} \tag{EC-11}$$

$b_S^* = (1 - b_I^*) \wedge pb_I^*\mu_C/\mu_{SC}$ *and $n^* = b_I^*/r_I + b_S^*/r_S^*$.*

PROOF: Because $\frac{w_N}{w_{SC}} < r_S\mu_{SC}$, by Lemma 1, $b_S = (1-b_I) \wedge pb_I\mu_C/\mu_{SC}$. We have two cases to consider depending on the relationship between $1-b_I$ and $pb_I\mu_C/\mu_{SC}$.

1. $1-b_I \le pb_I\mu_C/\mu_{SC}$: Minimizing (EC-9) becomes:

$$\min_{\frac{\mu_{SC}}{p\mu_C+\mu_{SC}} \le b_I \le 1\wedge \frac{\bar{\lambda}}{\mu_C}} \{w_N(b_I/r_I + (1-b_I)/r_S) + w_C(\bar{\lambda} - b_I\mu_C) + w_{SC}(pb_I\mu_C - (1-b_I)\mu_{SC})\}$$

which is a linear function of b_I. Thus, on the interval $\left[\frac{\mu_{SC}}{p\mu_C+\mu_{SC}}, 1\wedge\frac{\bar{\lambda}}{\mu_C}\right]$ the optimal value of b_I is

$$b_I^* = \begin{cases} 1\wedge\frac{\bar{\lambda}}{\mu_C} & \text{if } \; w_C\mu_C \ge w_N(1/r_I - 1/r_S) + w_{SC}(p\mu_C + \mu_{SC}), \\ \frac{\mu_{SC}}{p\mu_C+\mu_{SC}} & \text{otherwise,} \end{cases} \tag{EC-12}$$

where we note that $\frac{\mu_{SC}}{p\mu_C+\mu_{SC}} \le \frac{\hat{\lambda}}{\mu_C}$ due to Assumption 3.

2. $1-b_I > pb_I\mu_C/\mu_{SC}$: There are no bumping costs in this case. Minimizing (EC-9) becomes:

$$\min_{0\le b_I \le \frac{\mu_{SC}}{p\mu_C+\mu_{SC}}} \{w_N(b_I/r_I + pb_I\mu_C/\mu_{SC}r_S) + w_C(\bar{\lambda} - b_I\mu_C)\}.$$

Similar to case #1, on the interval $\left[0, \frac{\mu_{SC}}{p\mu_C+\mu_{SC}}\right]$ the optimal value of b_I is

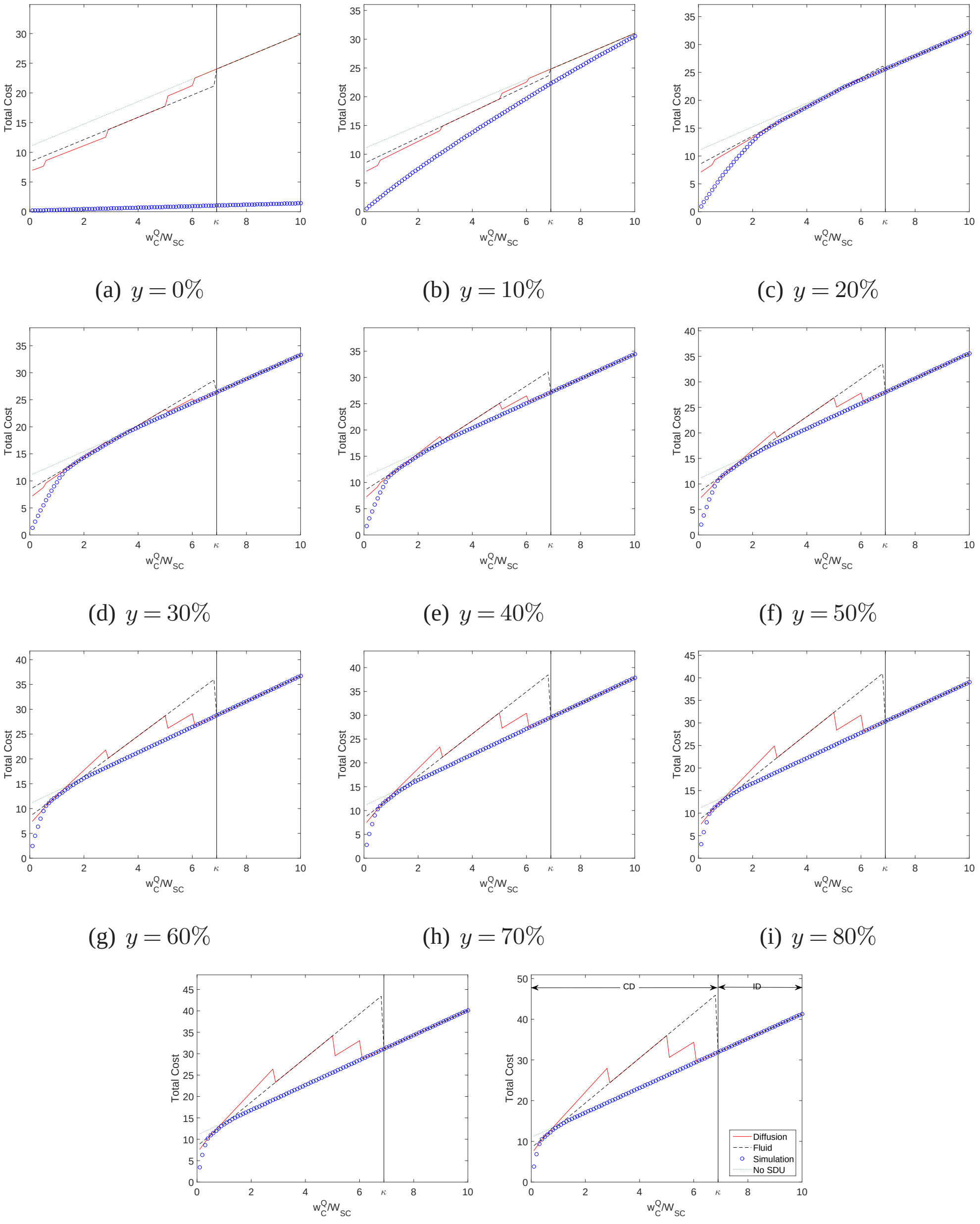
$$b_I^* = \begin{cases} \frac{\mu_{SC}}{p\mu_C+\mu_{SC}} & \text{if } \; \frac{w_C}{w_N} < \frac{1}{r_S\mu_{SC}}\left(p + \frac{r_S\mu_{SC}}{r_I\mu_C}\right), \\ 0 & \text{otherwise.} \end{cases} \tag{EC-13}$$

To complete the proof, we need to see which is the optimal value of b_I if both conditions $w_C\mu_C \ge w_N(1/r_I - 1/r_S) + w_{SC}(p\mu_C + \mu_{SC})$ and $\frac{w_C}{w_N} < \frac{1}{r_S\mu_{SC}}\left(p + \frac{r_S\mu_{SC}}{r_I\mu_C}\right)$ are satisfied. In this case, the two candidate values for b_I^* are either $1\wedge\frac{\bar{\lambda}}{\mu_C}$ or 0. But note that under the assumptions of the proposition, the two inequalities hold if and only if $w_C\mu_C = w_N(1/r_I - 1/r_S) + w_{SC}(p\mu_C + \mu_{SC}) = w_N(1/r_I + p\mu_C/(\mu_{SC}r_S))$ and $w_N/w_{SC} = r_S\mu_{SC}$, in which case both values of b_I^* are optimal. □

We make a few observations about the bed allocation solutions in this alternative formulation. First, we note that, again, the optimal SDU size may be substantial or zero. Similar to before, whether or not the hospital should have an SDU depends on the effective capacity of a single nurse in the SDU ($r_s\mu_{SC}$); however, because nurses are no longer the bottleneck, the corresponding cost ratio is now between the cost of that additional nurse, w_N, and w_{SC} instead of the cost of misplacing a critical patient, w_C.

In allocating beds to the ICU and SDU, we see there is still an ID-like regime where all (or as much as possible) of the critical patient demand is met, and a CD-like regime where a substantial number of beds is allocated to the SDU and all of the Semi-critical demand is met. The inclusion of nursing costs in the optimization framework introduces a third regime where the staffing costs are so high that it is optimal to have no ICU and no SDU (Note that the parameter $\kappa = p + \nu$ again plays a key role in this decision). While some (typically small, rural) hospitals do not have an ICU, it is generally due to strategic reasons, rather than because staffing costs are extraordinarily high.

EC-3. Supplemental Figure



(a) $y = 0\%$ (b) $y = 10\%$ (c) $y = 20\%$

(d) $y = 30\%$ (e) $y = 40\%$ (f) $y = 50\%$

(g) $y = 60\%$ (h) $y = 70\%$ (i) $y = 80\%$

(j) $y = 90\%$ (k) $y = 100\%$

Figure EC.1 **Queue-Dominated Case: Cost incurred by various strategies. Sensitivity Analysis for varying values of off-placement cost, yw_C^Q. Hospital parameters as given by Cady et al. (1995).**